CLEF 2023 SimpleText Track

What Happens if General Users Search Scientific Texts?

Liana Ermakova[1(✉)], Eric SanJuan[2], Stéphane Huet[2], Olivier Augereau[3], Hosein Azarbonyad[4], and Jaap Kamps[5]

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

Abstract. The general public tends to avoid reliable sources such as scientific literature due to their complex language and lacking background knowledge. Instead, they rely on shallow and derived sources on the web and in social media – often published for commercial or political incentives, rather than the informational value. Can text simplification help to remove some of these access barriers? This paper presents the CLEF 2023 SimpleText track tackling technical and evaluation challenges of scientific information access for a general audience. We provide appropriate reusable data and benchmarks for scientific text simplification, and promote novel research to reduce barriers in understanding complex texts. Our overall use-case is to create a simplified summary of multiple scientific documents based on a popular science query which provides a user with an accessible overview on this specific topic. The track has the following three concrete tasks. Task 1 (*What is in, or out?*): selecting passages to include in a simplified summary. Task 2 (*What is unclear?*): difficult concept identification and explanation. Task 3 (*Rewrite this!*): text simplification - rewriting scientific text. The three tasks together form a pipeline of a scientific text simplification system.

Keywords: Scientific text simplification · (Multi-document) summarization · Terminology extraction · Keyword extraction · Contextualization · Background knowledge · Scientific information distortion · Information retrieval

1 Introduction

Scientific texts such as research publications are difficult to understand for the general public, or even for scientists outside the exact specialism. The CLEF 2023 SimpleText track is unique in its focus on text simplification for scientific texts, and its general public use-case naturally combining information retrieval and natural language processing aspects.

J. Kamps et al. (Eds.): ECIR 2023, LNCS 13982, pp. 536–545, 2023.
https://doi.org/10.1007/978-3-031-28241-6_62

Text complexity or reading levels and text simplification in general have been studied for long in linguistics, education science, and natural language processing. Simplified texts are more accessible for non-native speakers [28], young readers, people with reading disabilities [5,13,22] or needed for reading assistance (e.g. congenitally deaf people) [15] or lower level of education. Improving text comprehensibility remains a challenge as it is difficult to define a desirable output of simplification [14]. Traditional readability scores are limited to word or sentence length, while vocabulary overlap based metrics do not consider information distortion. Recently, text simplification is gaining interest. The workshop on Scholarly Document Processing[1] is targeting an NLP audience [4]. They hosted tasks on scientific document summarization including a Lay Summary task. At EMNLP 2022, the TSAR (Text Simplification, Accessibility, and Readability)[2] hosted a lexical simplification task, and TermEval 2020 ran a shared task on automatic term extraction [23]. In contrast to that, SimpleText is not limited to lexical and grammatical simplification.

SimpleText aims to improve information access to scientific knowledge for a general audience, by providing appropriate reusable data and benchmarks for text simplification, promoting novel research and tools to reduce barriers in understanding complex texts. In contrast to the previous work, we focus on (1) information selection which is suitable for a general public; (2) searching for difficult concepts, including words, abbreviations, etc. that need to be explained and can not be discard; and (3) evaluation of information distortion which might occur during the simplification process.

The track's setup is based on the following pipeline: (1) select the information to be included in a simplified summary; (2) decide whether the selected information is sufficient and comprehensible or provide some background knowledge if not; (3) improve the readability of the text [7]. This results in the following three tasks [11]:

- ***Task 1: What is in, or out?*** *Selecting passages to include in a simplified summary.*
- ***Task 2: What is unclear?*** *Difficult concept identification and explanation.*
- ***Task 3: Rewrite this!*** *Text simplification - rewriting scientific text.*

We also welcome any submission that uses our data in other ways as a fourth open task.

In the rest of this paper, we will first reflect on the CLEF 2022 edition of the track in Sect. 2, and then provide a detailed description of each task in Sect. 3.

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

2 Results and Lessons Learnt from SimpleText'22

In the first year of running SimpleText as a track at CLEF 2022, its counted a total of 62 registered teams [11]. A total of 40 teams downloaded data from the server. A total of 9 distinct teams submitted 24 runs, of which 10 runs were updated. For Task 1 (selecting passages/abstracts to include) [26], 6 runs were submitted. For Task 2 (identifying difficult terms) [9], we received 4 runs. For Task 3 (rewriting text) [10], a total of 14 submissions was made. We have seen several post-submission experiments, and with all 2022 data available, the track is expected to gain in participation in 2023.

For Task 1, we saw a clear difference in the reading level of journalistic and scientific articles. The 2022 topical relevance qrels provide a unique resource that can be reused and enriched with additional judgments. In 2023, we will extend relevance judgment with supplementary labels on text complexity and credibility of the source publication, based on large-scale automatic and small-scale manual judgments further enriching the 2022 qrels. As the recall base is small, we have to expand the test collection by increasing pooling depth and adding new subtopics and queries for the same set of popular science articles.

In the 2022 edition, the Task 2 was limited to difficult term spotting. However, several runs for Task 3 inserted some context or definition for difficult terms in additional to language simplification [11,25]. This shows a demand for a corpus with explanations of difficult terms integrated in a text. Thus, we will update Task 2 to provide further context for difficult terms. The evaluation stage allowed to increase the annotated data for term difficulty spotting. We will reuse these data as a first stage of the annotation for the corpus in 2023 and provide additional evaluation data.

Multiple SimpleText participants applied T5-based text simplification models which previously demonstrated strong performance [27]. However, we observed that direct application of large pre-trained models often keeps sentences unchanged as in case of the runs of PortLinguE with 36% of unchanged sentences [11,17]. We also observed that large pre-trained models tend to insert unnecessary and even false information in the simplification due to their generative nature. This is a general problem of generative models attracting massive attention in AI, and studying further safeguard against over-generation and gratuitous insertions feels necessary. Thus, we will continue to provide human evaluation results with regard to the errors produced during simplification, which distinguishes SimpleText from existing benchmarks using only automatic text simplification evaluation metrics. Our general observation is that state of the art text simplification systems perform well, but far below human simplifications in terms of the length and the complexity of the resulting simplifications.

Our shared tasks are interconnected. The corpus is based on abstracts in response to a popular science request. While some complex terms do not need to be explained as they will be further removed at the language simplification step, others must be kept even if they are too complex in order to avoid severe information distortion. In this case additional context or explanations could be inserted, integrating Tasks 2 and 3. And the other way around, information

about the text complexity and the amount of revisions can inform the ranking stage of Task 1. For example, we can promote abstracts with more favorable reading levels in the ranking, and ensure our user is guided to relevant and already accessible abstracts first [18,19].

3 SimpleText 2023 Tasks

We will keep the three tasks for the 2023 edition. 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 2022 data.

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

Given a popular science article targeted to a general audience, this task aims at retrieving passages, that 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 the 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)[3] [29]. This collection was extracted from DBLP, ACM, MAG (Microsoft Academic Graph), and other sources. It contains: 4,894,083 bibliographic references published before 2020, 4,232,520 abstracts in English, 3,058,315 authors with their affiliations, and 45,565,790 ACM citations. We provide an ElasticSearch index to allow participants to retrieve passages or abstracts using BM25 [24]. Through a simple API, queries can be done on the textual content of abstracts together with authorship. Thus, the shared dataset provides: document abstract content for LDA (Latent Dirichlet Allocation) or Word Embedding (WE); document authors for coauthoring analysis; citation relationship between documents for co-citation analysis; citations by author for author impact factor analysis.

On the other hand, press articles, targeted to a general audience, are drawn from two sources: *The Guardian*, a major international newspaper for a general audience with a tech section, and *Tech Xplore*,[4] a web site taking part in the Science X Network to provide a comprehensive coverage of engineering and technology advances. Each of these popular science article 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

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

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. In the last year's edition, 40 articles, 20 of each source, were made available [11]. We plan to expand it with 10 other topics used as a test set. The 2022 qrels cover many topics (31) and queries (67) but with a limited pooling depth. In 2023, we will increase the pooling depth with at least 50 judged documents per query.

Evaluation. Topical relevance was only evaluated last year with a 0-5 score on the relevance degree towards the content of the original article [11]. Whereas this large scale can measure how close the retrieved abstract to the topic, the title or the textual content is, other facets, yet important in the context of text simplification, were missing. In 2023, we will continue evaluating on topical relevance, but also on text complexity (using readability measures and comparison to manually attributed scores), and source authoritativeness (using academic impact measures). The provided test collection will be simplified to three scores on a 0-2 scale:

- **Topic relevance**: Not relevant (0), relevant (1), highly relevant (2);
- **Text complexity**: Easy (0), difficult (1), very difficult (2);
- **Source credibility**: Low (0), medium (1), high credibility (2).

While these criteria can provide different levels of comparison between systems, we will still compute a unique ranking score using NDCG (as well as other measures) based on the fusion of the various criteria [21].

3.2 Task 2: Difficult Concept Identification and Explanation for a General Audience

The goal of this task is to decide which concepts in scientific abstracts require explanation and contextualization in order to help a reader to understand the scientific text. Complex Word Identification (CWI) and Lexical Simplification (LS) are the most popular approaches to assess and reduce the complexity [6,16, 31]. In the context of a query, some key concepts need to be contextualized with a definition, example and/or use-case that are easier to understand for a reader. There is ongoing research on this by generating definitions with a controllable complexity [1].

In 2023, we ask participants to identify such concepts and to provide useful and understandable explanations for them. Thus, the task has two steps:

1. to retrieve up to 5 difficult terms in a given passage from a scientific abstract;
2. to provide an explanation of these difficult terms (e.g. definition, abbreviation deciphering, example etc.).

Data. The corpus of Task 2 is based on the sentences in high-ranked abstracts to the requests of Task 1. For the first step of the task, i.e. retrieving difficult terms, we will use the train data collected in 2022 [11]. As for the test data, we

will provide additional passages coming from the DBLP abstracts as in Task 1. For the second step of the task we will provide additional training data for definition generation, extracted from a much larger corpus of full text articles. This training data contains pairs of $<sentence, concept>$ and a label per pair is provided. The binary label indicates whether the sentence provides a good definition for the concept or not. Samples in this dataset are extracted from books and articles published in ScienceDirect[5]. This dataset contains 43,368 samples distributed across 8 different domains and all pairs in this dataset are annotated by subject matter experts. There are a total of 9,870 positive samples (meaning that the sentence provides a good definition for the corresponding concept) and 33,498 negative samples. The average length of sentences in this dataset is 24.5 words. In addition to this dataset, participants are encouraged to use existing datasets extracted from other resources such as the WCL dataset [20] to train the definition generation model. Participants are also encourage to use gazetteers, wikification resources as well as resources for abbreviation deciphering.

Evaluation. As in 2022, we will evaluate complex concept spotting in terms of their complexity and the detected concept spans [11]. For the explanations of difficult terms, the evaluation set will contain 1,000 concepts and their definitions extracted by subject matter experts. We will automatically evaluate provided explanations by comparing them to references (e.g. ROUGE, cosine similarity etc.). We will provide manual evaluation 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. definition, abbreviation deciphering, examples, use cases etc.

3.3 Task 3: Text Simplification - Rewriting 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, supplemented with additional training data from the health domain. Our training data is a truly parallel corpus of directly simplified sentences (648 sentences for now) coming from scientific abstracts from the DBLP Citation Network Dataset for *Computer Science* and Google Scholar and PubMed articles on *Health and Medicine* [7,8,10,11]. 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 (English native speaker) working together [8,10,11].

All the existing large corpora used post-hoc aligned sentences [2,3,30,32,33]. The SimpleText corpus [11] contains directly simplified sentences, and is not much smaller than existing high-quality corpora like NEWSELA [30] (2,259 sentences).

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

Our track is the first to focus on scientific text simplification rather than news articles. In 2023, we will expand the training and evaluation data.

Evaluation. In 2023, we will emphasize large-scale automatic evaluation measures (SARI, ROUGE, compression, readability) that provide a reusable test collection.

These will be supplemented with small-scale detailed human evaluation of other aspects, essential for deeper analysis. As in 2022, 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) [11]. Rather than focus only on this evaluation which is similar to easy-to-read guidelines suggested in previous research [34], we prefer to assess the results according to information distortion which can be bring during simplification process. We distinguish the following types of information distortion with corresponding severity level: Style (1); Insertion of unnecessary details with regard to a query (1); Redundancy (without lexical overlap) (2); Insertion of false or unsupported information (3); Omission of essential details with regard to a query (4); Overgeneralization (5); Oversimplification(5); Topic shift (5); Contra sense / contradiction (6); Ambiguity (6); Nonsense (7).

4 Conclusions

This paper described the setup of the CLEF 2023 SimpleText track, which contains three interconnected tasks on scientific text simplification. Within the SimpleText track, we have already released extensive corpora and manually labeled data:

- a large corpus of over 4 million scientific abstracts that can be used for popular science;
- scientific terms from sentences coming from scientific abstracts with manually attributed difficulty scores;
- a parallel corpus of manually simplified sentences from scientific literature;
- 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.

Acknowledgment. This track would not have been possible without the great support of numerous individuals. We want thank in particular Silvia Araujo, Patrice Bellot, Julien Boccou, Pierre De Loor, Radia Hannachi, Helen McCombie, Diana Nurbakova, Irina Ovchinnikov, and Léa Talec; the students of the Université de Bretagne Occidentale; and all the 2022 track participants for their great help in discussing and shaping the track, and in creating all the evaluation data and training data for 2023. 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*).

References

1. August, T., Reinecke, K., Smith, N.A.: Generating scientific definitions with controllable complexity. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 8298–8317 (2022)
2. Bott, S., Saggion, H.: An unsupervised alignment algorithm for text simplification corpus construction. In: Proceedings of the Workshop on Monolingual Text-To-Text Generation, pp. 20–26 (2011)
3. Cardon, R., Grabar, N.: French biomedical text simplification: when small and precise helps. In: Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain, pp. 710–716. International Committee on Computational Linguistics (2020). https://www.aclweb.org/anthology/2020.coling-main.62
4. Chandrasekaran, M.K., et al.: Overview of the first workshop on scholarly document processing (SDP). In: Proceedings of the First Workshop on Scholarly Document Processing, pp. 1–6. Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.sdp-1.1. https://aclanthology.org/2020.sdp-1.1/
5. Chen, P., Rochford, J., Kennedy, D.N., Djamasbi, S., Fay, P., Scott, W.: Automatic text simplification for people with intellectual disabilities. In: Artificial Intelligence Science and Technology, pp. 725–731. World Scientific (2016). https://www.worldscientific.com/doi/abs/10.1142/9789813206823_0091
6. Cruz, F., Coustaty, M., Augereau, O., Kise, K., Journet, N.: An interactive recommendation system for 2nd language vocabulary learning-vocabulometer 2.0. In: 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW), vol. 3, pp. 28–32. IEEE (2019)
7. Ermakova, L., et al.: Overview of SimpleText 2021 - CLEF workshop on text simplification for scientific information access. In: Candan, K.S., et al. (eds.) CLEF 2021. LNCS, vol. 12880, pp. 432–449. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-85251-1_27
8. Ermakova, L., et al.: Automatic simplification of scientific texts: SimpleText lab at CLEF-2022. In: Hagen, M., et al. (eds.) ECIR 2022. LNCS, vol. 13186, pp. 364–373. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-99739-7_46
9. Ermakova, L., Ovchinnikova, I., Kamps, J., Nurbakova, D., Araújo, S., Hannachi, R.: Overview of the CLEF 2022 SimpleText task 2: complexity spotting in scientific abstracts. In: Faggioli et al. [12]
10. Ermakova, L., Ovchinnikova, I., Kamps, J., Nurbakova, D., Araújo, S., Hannachi, R.: Overview of the CLEF 2022 SimpleText task 3: query biased simplification of scientific texts. In: Faggioli et al. [12]
11. Ermakova, L., et al.: Overview of the CLEF 2022 SimpleText lab: automatic simplification of scientific texts. In: Barrón-Cedeño, A., et al. (eds.) CLEF 2022. LNCS, vol. 13390, pp. 470–494. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-13643-6_28
12. Faggioli, G., Ferro, N., Hanbury, A., Potthast, M. (eds.): Proceedings of the Working Notes of CLEF 2022: Conference and Labs of the Evaluation Forum. CEUR Workshop Proceedings (2022)
13. Gala, N., Tack, A., Javourey-Drevet, L., François, T., Ziegler, J.C.: Alector: a parallel corpus of simplified French texts with alignments of misreadings by poor and dyslexic readers. In: Language Resources and Evaluation for Language Technologies (LREC) (2020)

14. Grabar, N., Saggion, H.: 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, pp. 453–463 (2022)
15. Inui, K., Fujita, A., Takahashi, T., Iida, R., Iwakura, T.: Text simplification for reading assistance: a project note. In: Proceedings of the Second International Workshop on Paraphrasing - Volume 16, PARAPHRASE 2003, pp. 9–16. ACL, USA (2003). https://doi.org/10.3115/1118984.1118986
16. Kochmar, E., Gooding, S., Shardlow, M.: Detecting multiword expression type helps lexical complexity assessment. In: LREC 2020: Proceedings of the 12th Conference on Language Resources and Evaluation (2020)
17. Monteiro, J., Aguiar, M., Araújo, S.: Using a pre-trained SimpleT5 model for text simplification in a limited corpus. In: Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, 5–8 September 2022, Bologna, Italy. CEUR Workshop Proceedings, CEUR-WS.org (2022)
18. Mostert, F., Sampatsing, A., Spronk, M., Kamps, J.: University of Amsterdam at the CLEF 2022 SimpleText track. In: Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, 5–8 September 2022, Bologna, Italy. CEUR Workshop Proceedings, CEUR-WS.org (2022)
19. Nakatani, M., Jatowt, A., Tanaka, K.: Easiest-first search: towards comprehension-based web search. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management, pp. 2057–2060 (2009)
20. Navigli, R., Velardi, P.: Learning word-class lattices for definition and hypernym extraction. In: ACL, pp. 1318–1327 (2010)
21. Ravana, S.D., Moffat, A.: Score aggregation techniques in retrieval experimentation. In: Proceedings of the Twentieth Australasian Conference on Australasian Database, vol. 92, pp. 57–66 (2009)
22. Rello, L., Baeza-Yates, R., Bott, S., Saggion, H.: Simplify or help? Text simplification strategies for people with dyslexia. In: Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility, pp. 1–10 (2013)
23. Rigouts Terryn, A., Hoste, V., Drouin, P., Lefever, E.: 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), pp. 85–94. European Language Resources Association (ELRA) (2020)
24. Robertson, S., Zaragoza, H., et al.: The probabilistic relevance framework: BM25 and beyond. Found. Trends® Inf. Retrieval **3**(4), 333–389 (2009)
25. Rubio, A., Martínez, P.: HULAT-UC3M at SimpleText@CLEF-2022: scientific text simplification using BART. In: Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, 5–8 September 2022. CEUR Workshop Proceedings, CEUR-WS.org (2022)
26. SanJuan, E., Huet, S., Kamps, J., Ermakova, L.: Overview of the CLEF 2022 SimpleText task 1: passage selection for a simplified summary. In: Faggioli et al. [12]
27. Sheang, K.C., Saggion, H.: Controllable sentence simplification with a unified text-to-text transfer transformer. In: Proceedings of the 14th International Conference on Natural Language Generation, pp. 341–352 (2021)
28. Siddharthan, A.: An architecture for a text simplification system (2002). https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.1.9968&rank=1

29. Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., Su, Z.: ArnetMiner: extraction and mining of academic social networks. In: KDD 2008, pp. 990–998 (2008)
30. Xu, W., Callison-Burch, C., Napoles, C.: Problems in current text simplification research: new data can help. Trans. ACL 3, 283–297 (2015). https://www.mitpressjournals.org/doi/abs/10.1162/tacl_a_00139
31. Yimam, S.M., et al.: 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) (2018)
32. Zhang, X., Lapata, M.: Sentence simplification with deep reinforcement learning. In: EMNLP 2017: Conference on Empirical Methods in Natural Language Processing, pp. 584–594. Association for Computational Linguistics (2017)
33. Zhu, Z., Bernhard, D., Gurevych, I.: A monolingual tree-based translation model for sentence simplification. In: Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), Beijing, China, pp. 1353–1361. Coling 2010 Organizing Committee (2010). https://www.aclweb.org/anthology/C10-1152
34. Štajner, S., Sheang, K.C., Saggion, H.: Sentence Simplification Capabilities of Transfer-Based Models (2022)