Overview of the CLEF 2023 SimpleText Task 1: Passage Selection for a Simplified Summary

Éric SanJuan[1], Stéphane Huet[1], Jaap Kamps[2] and Liana Ermakova[3]

[1]Avignon Université, LIA, France
[2]University of Amsterdam, Amsterdam, The Netherlands
[3]Université de Bretagne Occidentale, HCTI, France

Abstract

This paper presents an overview of the CLEF 2023 SimpleText Task 1: Content Selection, asking systems to retrieve scientific abstracts in response to a query prompted by a popular science article. Overall, the SimpleText track provides an evaluation platform for the automatic simplification of scientific texts. We discuss the details of the task set-up. First, the SimpleText Corpus with over 4 million academic papers and abstracts. Second, the Topics based on 40 popular science articles in the news and the 114 Queries prompted by them. Third, the Formats of requests and results, the Evaluation labels and Evaluation measures used. Fourth, the Results of the runs submitted by our participants.

Keywords

information retrieval, scientific documents, text simplification, scientific information retrieval, non-expert queries, press outlets, query-document relationships (Q-rels), popularized science

1. Introduction

This paper presents an overview of the first task in the SimpleText track on automatic simplification of scientific texts following up on the CLEF 2021 SimpleText Workshop [1] and CLEF 2022 SimpleText Track [2]. The main goal of the SimpleText track is to provide data and benchmarks to advance research in this area. This paper focuses on *Task 1: What is in (or out)? Selecting passages to include in a simplified summary*. This task is part of a pipeline with the two other SimpleText tasks, namely *Task 2: What is unclear? Difficult concept identification and explanation* and *Task 3: Rewrite this! Given a query, simplify passages from scientific abstracts*. For a comprehensive understanding of the other tasks, the overview papers of Task 2 [3] and Task 3 [4], as well as the Track overview paper [5], provide detailed information and insights.

Scientific literacy is a vital skill for individuals. It serves as a key component of critical thinking, enabling individuals to make objective decisions and assess the validity and significance of research findings. Scientific literacy helps to differentiate between reliable evidence and unsubstantiated claims and navigate the complex landscape of scientific advancements. Despite increasing digitization and open access to scientific literature, several barriers remain that impede non-experts from accessing unbiased scientific information from these texts. One

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece
eric.sanjuan@univ-avignon.fr (É. SanJuan); liana.ermakova@univ-brest.fr (L. Ermakova)
https://simpletext-project.com/ (L. Ermakova)
0000-0002-7598-7474 (L. Ermakova)

primary challenge is the difficulty in comprehending scientific literature, which arises from its reliance on specialized knowledge and the utilization of complex terminology. As a result, non-experts may face obstacles when attempting to understand and interpret scientific papers. Retrieval of relevant yet credible and understandable scientific documents is still a challenge as search engines virtually ignore documents' difficulty.

The rest of the paper is organized as follows. Section 2 provides details on the datasets utilized and the evaluation metrics employed in the study. Section 3 offers an overview of the retrieval approaches adopted by the participants, specifically focusing on the scientific text. In Section 4, the official submissions' results are presented and discussed. Finally, Section 5 summarizes the findings and outlines potential directions for future research.

2. SimpleText Task 1 Test Collection

This section provides an overview of the resulting test collection, detailing the corpus, the topics and queries, the input and output format, as well as the used evaluation measures.

2.1. Corpus: DBLP abstracts

The corpus utilized in this task is the Citation Network Dataset: DBLP+Citation, ACM Citation network (12th version released in 2020[1])[6, 7] containing 4,894,081 papers (2020-04-09) with abstract content and field of subjects, made available from Microsoft academic services[8]. An ElasticSearch index is provided to participants with access through an API. A JSON dump of the index is also released. Besides, document bibliographic information and abstract content in the form of an inverted positional index, can be retrieved from OpenAlex[2][9] using their document ids as work references (W).

2.2. Topics: Press articles

Topics are a selection of press articles from the tech section of The Guardian newspaper (topics G01 to G20) and the Tech Xplore website (topics T01 to T20). URLs to original articles and textual content of each topic are provided to participants. All abstracts extracted from the document collection by participants are expected to be relevant to subjects addressed in the press articles.

2.2.1. Queries as facets

Between one and four keywords queries are provided with each topic. It has been manually checked that each query allows retrieving relevant passages that could be inserted as citations in the press article.

[1] https://www.aminer.org
[2] https://docs.openalex.org/

Table 1
Statistics on qrels.

Qrels	Topics	#Queries	#Assessed abstracts			#Avg Ass.
			0	1	2	
2022 test	G1-G20, T2,4,5,10-12,15-16,T18-20	72	192	187	107	6.8
2023 train	G1-G15	29	728	338	237	44.9
2023 test	G16-G20, T1-T5	34	2260	357	1218	112.8

2.2.2. Qrels

Quality relevance of abstracts w.r.t. topics is given in Simpletext_2023_task1_train.qrels, distributed to participants. This file extends the qrels released last year with a significant increase of the depth of judgments of abstracts per query (Table 1, lines 1 and 2). Thus, the average number of assessed abstracts by query has raised from 6.8 to 44.9. Relevance annotations are provided on a 0-2 scale (the higher the more relevant) for 29 queries associated with the first 15 articles from the Guardian.[3]

From runs submitted this year (see Section 3), a new qrels file was built using pooling on 10 topics different from the train file: the last 5 topics from the Guardian and the first 5 from Tech Xplore (Table 1, line 3). For pooling, only top 10 abstracts of each submitted run were considered for manual assessment. While the train file was released to participants to allow them to have supervised approaches, the test file is only used for the track evaluation.

2.3. Expected results

2.3.1. Ad-hoc passage retrieval

Participants had to retrieve, for each topic and each query, all passages from DBLP abstracts, related to the query and relevant to be inserted as a citation in the paper associated with the topic. Some passages could require simplification. We encouraged participants to take into account passage complexity as well as its credibility/influentialness.

2.3.2. Open passage retrieval (optional)

Participants were also encouraged to extract supplementary relevant queries from the titles or content articles and to provide results based on these supplementary queries.

2.3.3. Output format

Results had to be provided in a TREC style JSON or TSV format with the following fields:

run_id Run ID starting with : team_id_task_id_method_used, e.g. *UBO_task_1_TFIDF*

manual Whether the run is manual {0,1}

[3]Judgments made last year on a 0-5 scale were transformed with the following conversion rules: $0/1 \rightarrow 0; 2/3 \rightarrow 1; 4/5 \rightarrow 2$.

Table 2
CLEF 2023 SimpleText Task 1 on content selection: example of output

Run	M/A	Topic	Query	Doc	Rel	Comb	Passage
ST1_task1_1	0	G01	G01.1	1564531496	0.97	0.85	A CDA is a mobile user device, similar to a Personal Digital Assistant (PDA). It supports the citizen when dealing with public authorities and proves his rights - if desired, even without revealing his identity.
ST1_task1_1	0	G01	G01.1	3000234933	0.9	0.9	People are becoming increasingly comfortable using Digital Assistants (DAs) to interact with services or connected objects
ST1_task1_1	0	G01	G01.2	1448624402	0.6	0.3	As extensive experimental research has shown individuals suffer from diverse biases in decision-making.

topic_id Topic ID

query_id Query ID used to retrieve the document (if one of the queries provided for the topic was used; 0 otherwise)

doc_id ID of the retrieved document (to be extracted from the JSON output)

rel_score Relevance score of the passage (in the [0-1] scale)

comb_score General score that may combine relevance and other aspects: readability, citation measures…(in the [0-1] scale)

passage Text of the selected passage

For each query, the maximum number of distinct DBLP references (doc_id field) was 100 and the total length of passages could not exceed 1000 tokens. The idea of taking into account complexity is to have passages easier to understand for non-experts, while credibility score aims at guiding them on the expertise of authors and the value of publication w.r.t. the article topic. For example, complexity scores can be evaluated using readability score and credibility scores using bibliometrics.

Here is an output format example:

An example of the output is shown in Table 2. For each topic, the maximum number of distinct DBLP references (_id JSON field) was 100 and the total length of passages was not to exceed 1,000 tokens.

2.4. Evaluation Metrics

Passage relevance has been assessed based on:

- lexical and semantic overlap of extracted passages with topic article content
- manual relevance assessment of a pool of passages (relevance scores provided by participants will be used to measure ranking quality)
- manual assessment by non-expert users of credibility and complexity

3. Participants' approaches

Five teams submitted 39 runs in total.

Elsevier (represented as $Elsevier*$ in Table 7) [10] made 10 submissions to Task 1. Their submissions focused on evaluating the performance of neural rankers, utilizing both zero-shot approaches and unsupervised fine-tuning techniques on scientific documents.

The University of Amsterdam ($UAms_*$) [11] entered 10 submissions for Task 1. Initially, they contributed three baseline rankers aimed at enhancing the pool of judgments. These baseline rankers included an ElasticSearch run utilizing keyword queries (non-phrase), as well as a cross-encoder reranking approach applied to the top 100 and top 1,000 results obtained from ElasticSearch. They made four additional submissions that focused on evaluating the credibility of the retrieved results. These submissions took into consideration factors such as the recency and number of citations for each paper to assess their credibility. Finally, they submitted three runs specifically aimed at addressing the readability of the retrieved results.

The University of Maine (AIIR Lab, $maine\_*$) [12] submitted 5 runs for Task 1. Their submissions involved experimenting with cross-encoder and bi-encoder models, comparing their performance to lexical models.

The University of Milano Bicocca ($unimib\_DoSSIER\_*$) [13] submitted 2 runs for Task 1. Their submissions encompassed domain-specific approaches for scientific documents, including probabilistic lexical ranking, hierarchical document classification, and pseudo-relevance feedback (PRF).

4. Results

4.1. Retrieval Effectiveness

Table 3 shows the results of the CLEF 2023 Simpletext Task 1, based on the 34 test queries. The main measure of the task is NDCG@10, and the table is sorted on this measure for convenience. Let us note that some participants used the possibility of having two different scores in their run. Since ranking made according to the relevance or combination scores may vary in this case, we add in the result table ^{rel} and ^{comb} for runs with two different scores.

A number of observations stand out. First and foremost, we see in general that the top of the Table is dominated by neural rankers; in particular, cross-encoders trained on MSMarco applied in a zero-shot way (or variants thereof), perform well for ranking scientific abstracts on NDCG@10 and other early precision measures. Traditional lexical retrieval models perform

reasonably but at some distance from the top-scoring runs, with the neural runs typically re-ranking such a lexical baseline run.

Second, looking at more recall-oriented measures, such as MAP and bpref, the picture is more mixed. This is indicating some approaches privilege precision over recall, whereas other approaches seem to promote all recall levels.

Third, some submissions aimed to balance the topical relevance with the readability or credibility of the results. We observe that these runs still achieve competitive retrieval effectiveness, despite removing or down-ranking highly relevant abstracts that have for example a high text complexity or are dated with low numbers of citations.

The document collection contains two different sets of topics. On the one hand, the Guardian topics (G) are built from articles related to societal issues: privacy, ubiquity, misinformation, etc. and are usually associated with general queries that must be disambiguated in the context of the articles. On the other hand, the Tech Xplore topics (T) are linked to an original scientific paper and deal with more technical facets: neural networks, indoor positioning system, RISC architecture, etc. Further analysis was performed on the behavior of systems according to these two sets and shown in Tables 4 and 5. A comparison of these results exhibits that more relevant abstracts were found for the T topics, with higher scores overall. However, ranking against NDCG@10 leads to a different ranking of systems. While ElsevierSimpleText_run8 still outperforms other systems on T topics, maine_CrossEncoder1 becomes the 1st system on G topics.

Going back to the training qrels released to participants, we observe as expected that supervised models learned on these data have the highest scores (Table 6). Runs submitted by the University of Maine and to a lesser extent by the University of Amsterdam outperform others. Runs by Elsevier also resorted to neural rerankers, but other training data were used. Only G topics were included in the train qrels, which may explain why systems behave differently between G or T topics used in the test qrels.

4.2. Analysis of Readability

Table 7 shows several statistics over to the top 10 results retrieved for the entire topic set for Task 1:

- citation analysis (impact factor based on ACM records and average number of references per document),
- textual analysis (document length and FKGL scores).

Let us note that when two different scores were provided for a run, only the combined one was considered in this evaluation.

We make a number of observations.

First, it appears that the most effective ranking models tend to retrieve abstracts that are not only longer, but also exhibit greater length variability. These retrieved abstracts often have higher impact factors and extensive bibliographies. There also seems to be a discernible difference between the lengths of abstracts retrieved by lexical-based systems compared to those retrieved by neural-based systems.

Table 3
Evaluation of SimpleText Task 1 (Test qrels: G16-G20+T01-T05).

Run	MRR	Precision		NDCG		Bpref	MAP
		10	20	10	20		
ElsevierSimpleText_run8	0.8082	0.5618	0.3515	0.5881	0.4422	0.2371	0.1633
ElsevierSimpleText_run7	0.7136	0.5618	0.4103	0.5704	0.4627	0.2626	0.1915
maine_CrossEncoder1^{rel}	0.8106	0.5382	0.4456	0.5675	0.4908	0.3317	0.2810
maine_CrossEncoderFinetuned1^{rel}	0.7691	0.5559	0.4441	0.5542	0.4840	0.3433	0.2572
maine_CrossEncoder1^{comb}	0.7309	0.5265	0.4500	0.5455	0.4841	0.3337	0.2754
maine_CrossEncoderFinetuned1^{comb}	0.7338	0.4971	0.4000	0.4859	0.4295	0.3443	0.2385
ElsevierSimpleText_run5	0.6600	0.4765	0.3838	0.4826	0.4186	0.2542	0.1828
maine_CrossEncoderFinetuned2^{rel}	0.6588	0.4971	0.4088	0.4821	0.4254	0.3185	0.2242
ElsevierSimpleText_run2	0.7010	0.4676	0.4059	0.4791	0.4282	0.2528	0.1942
UAms_CE100^{rel}	0.7050	0.4912	0.4044	0.4782	0.4236	0.2616	0.2011
ElsevierSimpleText_run6	0.6402	0.4676	0.3853	0.4723	0.4185	0.2557	0.1809
ElsevierSimpleText_run4	0.6774	0.4529	0.3794	0.4721	0.4116	0.2485	0.1898
ElsevierSimpleText_run9	0.5933	0.4735	0.3176	0.4655	0.3595	0.1758	0.1238
ElsevierSimpleText_run1	0.6821	0.4588	0.3824	0.4626	0.4071	0.2573	0.1823
maine_CrossEncoderFinetuned2^{comb}	0.7082	0.4706	0.3926	0.4617	0.4089	0.3259	0.2253
UAms_CE1k_Filter	0.6403	0.4765	0.3559	0.4533	0.3743	0.2727	0.1936
ElsevierSimpleText_run3	0.6502	0.4471	0.3779	0.4460	0.3994	0.2558	0.1785
UAms_CE1k^{rel}	0.6329	0.4735	0.4044	0.4448	0.4049	0.2797	0.2051
maine_PI2TFIDF^{rel}	0.5791	0.4382	0.2853	0.4212	0.3313	0.2159	0.1410
UAms_ElF_Cred44	0.6888	0.4324	0.3338	0.4103	0.3499	0.2395	0.1719
UAms_CE100^{comb}	0.6779	0.3971	0.3456	0.4016	0.3642	0.2658	0.1792
maine_PI2TFIDF^{comb}	0.5626	0.4176	0.2809	0.4014	0.3218	0.2155	0.1364
UAms_Elastic	0.6424	0.4059	0.3456	0.3910	0.3541	0.2501	0.1895
UAms_ElF_Cred53	0.6429	0.4088	0.3382	0.3883	0.3468	0.2454	0.1833
UAms_ElF_Cred44Read	0.6625	0.3971	0.3147	0.3723	0.3282	0.2123	0.1403
UAms_CE1k_Combine^{comb}	0.5880	0.4147	0.3515	0.3706	0.3398	0.2700	0.1865
UAms_CE1k^{comb}	0.5880	0.4147	0.3515	0.3706	0.3398	0.2700	0.1865
UAms_ElF_Read25	0.6076	0.3735	0.3074	0.3539	0.3190	0.2194	0.1522
UAms_ElF_Cred53Read	0.6088	0.3676	0.3059	0.3469	0.3153	0.2133	0.1456
maine_tripletloss^{rel}	0.5425	0.3500	0.2162	0.3439	0.2557	0.1296	0.0690
maine_tripletloss^{comb}	0.5502	0.3382	0.2176	0.3353	0.2561	0.1335	0.0696
unimib_DoSSIER_2	0.5201	0.2853	0.2515	0.2980	0.2683	0.1898	0.1141
unimib_DoSSIER_4	0.5202	0.2853	0.2441	0.2972	0.2632	0.1873	0.1111

Second, in terms of readability levels, the overwhelming majority of systems retrieve abstracts with an FKGL of around 14 — corresponding to university-level texts. This is entirely as expected since the corpus is based on scientific text, known to be written for experts with higher text complexity than for example newspaper articles.

Third, two systems retrieve abstracts with an FKGL of 11-12 — corresponding to the exit level of compulsory education, and the reading level of the average newspaper reader targeted by the use case of the track. These runs still achieved very reasonable retrieval effectiveness

Table 4
Evaluation of SimpleText Task 1 (Test G-only qrels).

Run	MRR	Precision		NDCG		Bpref	MAP
		10	20	10	20		
ElsevierSimpleText_run1	0.5602	0.2941	0.2441	0.3027	0.2680	0.1616	0.1155
ElsevierSimpleText_run2	0.5882	0.2765	0.2559	0.2942	0.2755	0.1622	0.1170
ElsevierSimpleText_run3	0.5387	0.2882	0.2471	0.2945	0.2681	0.1625	0.1159
ElsevierSimpleText_run4	0.5460	0.3000	0.2294	0.3186	0.2694	0.1614	0.1207
ElsevierSimpleText_run5	0.5145	0.3412	0.2559	0.3369	0.2896	0.1722	0.1222
ElsevierSimpleText_run6	0.5157	0.3412	0.2676	0.3423	0.3027	0.1733	0.1261
ElsevierSimpleText_run7	0.4958	0.3765	0.2588	0.3699	0.2952	0.1730	0.1222
ElsevierSimpleText_run8	0.6850	0.3706	0.2235	0.3948	0.3002	0.1672	0.1132
ElsevierSimpleText_run9	0.4102	0.3235	0.2118	0.2990	0.2366	0.1219	0.0816
UAms_CE100^{rel}	0.6041	0.3353	0.2794	0.3185	0.2920	0.1863	0.1335
UAms_CE100^{comb}	0.6817	0.3647	0.2765	0.3724	0.3136	0.2030	0.1482
UAms_CE1k^{rel}	0.5104	0.3412	0.2824	0.3128	0.2851	0.1879	0.1280
UAms_CE1k_Combine^{comb}	0.5796	0.3294	0.2853	0.3179	0.2933	0.1951	0.1455
UAms_CE1k_Filter^{comb}	0.4976	0.3471	0.2471	0.3294	0.2688	0.1906	0.1311
UAms_ElF_Cred44	0.5899	0.3059	0.2265	0.2823	0.2482	0.1858	0.1276
UAms_ElF_Cred44Read	0.6299	0.2882	0.2088	0.2666	0.2388	0.1703	0.1106
UAms_ElF_Cred53	0.5373	0.2941	0.2235	0.2685	0.2453	0.1903	0.1376
UAms_ElF_Cred53Read	0.5801	0.2588	0.1971	0.2469	0.2269	0.1725	0.1148
UAms_ElF_Read25	0.5801	0.2647	0.2029	0.2555	0.2356	0.1799	0.1220
UAms_Elastic	0.5373	0.2941	0.2294	0.2724	0.2519	0.1954	0.1432
maine_CrossEncoder1^{rel}	0.6947	0.4353	0.3529	0.4483	0.3973	0.2983	0.2441
maine_CrossEncoder1^{comb}	0.5892	0.4176	0.3647	0.4185	0.3930	0.3020	0.2390
maine_CrossEncoderFinetuned1^{rel}	0.7471	0.4471	0.3529	0.4448	0.3872	0.3019	0.2218
maine_CrossEncoderFinetuned1^{comb}	0.7765	0.4235	0.3353	0.4263	0.3848	0.3031	0.2068
maine_CrossEncoderFinetuned2^{rel}	0.5738	0.3882	0.3324	0.3582	0.3274	0.2659	0.1722
maine_CrossEncoderFinetuned2^{comb}	0.6959	0.3941	0.3353	0.3868	0.3542	0.2749	0.1842
maine_PI2TFIDF^{rel}	0.5628	0.3882	0.2235	0.3318	0.2731	0.2235	0.1492
maine_PI2TFIDF^{comb}	0.5216	0.3647	0.2265	0.3283	0.2690	0.2254	0.1477
maine_tripletloss^{rel}	0.4475	0.3471	0.1882	0.3069	0.2213	0.1279	0.0731
maine_tripletloss^{comb}	0.4433	0.3294	0.1882	0.2960	0.2196	0.1352	0.0745
uninib_DoSSIER_2	0.4349	0.1588	0.1353	0.1693	0.1534	0.1397	0.0601
uninib_DoSSIER_4	0.4351	0.1529	0.1324	0.1657	0.1514	0.1379	0.0584

(NDCG@10 0.37-0.45 in Table 3) while only retrieving abstracts with the desirable readability level.

5. Conclusion

In this CLEF lab track, a range of language models has been systematically examine and compare by participants in an attempt to elucidate their strengths and weaknesses when employed for accessing scientific information, especially amid the burgeoning science communication

Table 5
Evaluation of SimpleText Task 1 (Test T-only qrels).

Run	MRR	Precision		NDCG		Bpref	MAP
		10	20	10	20		
ElsevierSimpleText_run1	0.8039	0.6235	0.5206	0.6224	0.5462	0.3530	0.2491
ElsevierSimpleText_run2	0.8137	0.6588	0.5559	0.6641	0.5809	0.3434	0.2713
ElsevierSimpleText_run3	0.7618	0.6059	0.5088	0.5975	0.5308	0.3492	0.2411
ElsevierSimpleText_run4	0.8088	0.6059	0.5294	0.6257	0.5538	0.3357	0.2588
ElsevierSimpleText_run5	0.8056	0.6118	0.5118	0.6283	0.5477	0.3363	0.2435
ElsevierSimpleText_run6	0.7647	0.5941	0.5029	0.6024	0.5342	0.3380	0.2358
ElsevierSimpleText_run7	0.9314	0.7471	0.5618	0.7709	0.6302	0.3522	0.2609
ElsevierSimpleText_run8	0.9314	0.7529	0.4794	0.7809	0.5838	0.3071	0.2134
ElsevierSimpleText_run9	0.7764	0.6235	0.4235	0.6319	0.4823	0.2297	0.1659
UAms_CE100^{rel}	0.6041	0.6471	0.5294	0.6379	0.5552	0.3369	0.2687
UAms_CE100^{comb}	0.6740	0.4294	0.4147	0.4309	0.4148	0.3287	0.2101
UAms_CE1k^{rel}	0.5104	0.6059	0.5265	0.5768	0.5247	0.3716	0.2822
UAms_CE1k_Combine^{comb}	0.5965	0.5000	0.4176	0.4232	0.3862	0.3448	0.2276
UAms_CE1k_Filter^{comb}	0.7829	0.6059	0.4647	0.5773	0.4798	0.3548	0.2561
UAms_ElF_Cred44	0.5899	0.5588	0.4412	0.5382	0.4516	0.2932	0.2161
UAms_ElF_Cred44Read	0.6299	0.5059	0.4206	0.4780	0.4176	0.2542	0.1700
UAms_ElF_Cred53	0.5373	0.5235	0.4529	0.5081	0.4483	0.3005	0.2289
UAms_ElF_Cred53Read	0.5801	0.4765	0.4147	0.4468	0.4037	0.2541	0.1764
UAms_ElF_Read25	0.5801	0.4824	0.4118	0.4523	0.4023	0.2589	0.1824
UAms_Elastic	0.5373	0.5176	0.4618	0.5098	0.4565	0.3050	0.2358
maine_CrossEncoder1^{rel}	0.9265	0.6412	0.5382	0.6867	0.5842	0.3651	0.3180
maine_CrossEncoder1^{comb}	0.8725	0.6353	0.5353	0.6726	0.5752	0.3654	0.3119
maine_CrossEncoderFinetuned1^{rel}	0.7912	0.6647	0.5559	0.6636	0.5809	0.3848	0.2926
maine_CrossEncoderFinetuned1^{comb}	0.6912	0.5706	0.4647	0.5455	0.4742	0.3854	0.2703
maine_CrossEncoderFinetuned2^{rel}	0.7437	0.6059	0.5000	0.6059	0.5234	0.3710	0.2761
maine_CrossEncoderFinetuned2^{comb}	0.7206	0.5471	0.4500	0.5366	0.4636	0.3769	0.2663
maine_PI2TFIDF^{rel}	0.5955	0.5118	0.3471	0.5106	0.3895	0.2084	0.1327
maine_PI2TFIDF^{comb}	0.6036	0.4706	0.3353	0.4745	0.3746	0.2057	0.1251
maine_tripletloss^{rel}	0.6374	0.3529	0.2441	0.3809	0.2900	0.1313	0.0649
maine_tripletloss^{comb}	0.6572	0.3471	0.2471	0.3746	0.2926	0.1317	0.0646
uninib_DoSSIER_2	0.6053	0.4118	0.3676	0.4267	0.3832	0.2399	0.1681
uninib_DoSSIER_4	0.6053	0.4176	0.3559	0.4288	0.3750	0.2367	0.1639

landscape.

Task lab results involve the construction of a unique test set that pairs layperson queries with ideal scientific documents. The queries have been derived from a large-scale student experiment, providing a realistic representation of the nature of layperson queries. The test set has been used to train various participating systems, and the results have been enhanced with expert manual annotation of an additional pool of results.

This comprehensive exploration provides valuable insights into the effectiveness of each model and presents a critical comparison of their performances. The findings from this study

Table 6
Evaluation of SimpleText Task 1 (Train qrels: G01-G15).

Run	MRR	Precision		NDCG		Bpref	MAP
		10	20	10	20		
ElsevierSimpleText_run1	0.5294	0.2207	0.1931	0.2199	0.2296	0.3313	0.1815
ElsevierSimpleText_run2	0.5234	0.2621	0.2655	0.2366	0.2832	0.3952	0.2773
ElsevierSimpleText_run3	0.4897	0.2207	0.1845	0.2126	0.2177	0.3137	0.1694
ElsevierSimpleText_run4	0.4447	0.2655	0.2534	0.2437	0.2776	0.3894	0.2781
ElsevierSimpleText_run5	0.4156	0.2103	0.1862	0.2097	0.2277	0.3305	0.1742
ElsevierSimpleText_run6	0.3989	0.2000	0.1707	0.2000	0.2098	0.3091	0.1642
ElsevierSimpleText_run7	0.3512	0.2241	0.1828	0.1871	0.1986	0.3725	0.1498
ElsevierSimpleText_run8	0.2691	0.1828	0.1500	0.1526	0.1673	0.3585	0.1281
ElsevierSimpleText_run9	0.3238	0.1448	0.1241	0.1310	0.1319	0.1434	0.0906
UAms_CE100^{rel}	0.5252	0.3034	0.2690	0.2947	0.3145	0.4012	0.3033
UAms_CE100^{comb}	0.4371	0.3069	0.2466	0.2489	0.2795	0.3998	0.2838
UAms_CE1k^{rel}	0.4608	0.2379	0.1948	0.2307	0.2421	0.3335	0.2001
UAms_CE1k_Combine^{comb}	0.3182	0.1966	0.1897	0.1633	0.2005	0.3211	0.1714
UAms_CE1k_Filter^{comb}	0.4952	0.2414	0.1879	0.2431	0.2423	0.3249	0.1934
UAms_ElF_Cred44	0.6375	0.3655	0.3086	0.3658	0.3696	0.3531	0.3405
UAms_ElF_Cred44Read	0.5746	0.3379	0.2845	0.3079	0.3145	0.2724	0.2294
UAms_ElF_Cred53	0.5682	0.3759	0.3414	0.3729	0.4029	0.4284	0.4050
UAms_ElF_Cred53Read	0.5312	0.3379	0.3103	0.3053	0.3323	0.3089	0.2706
UAms_ElF_Read25	0.5257	0.3310	0.3086	0.2970	0.3264	0.3082	0.2741
UAms_Elastic	0.5605	0.3655	0.3345	0.3627	0.3924	0.4226	0.4072
maine_CrossEncoder1^{rel}	0.7102	0.4448	0.4086	0.4604	0.5017	0.4629	0.5064
maine_CrossEncoder1^{comb}	0.7165	0.4414	0.4155	0.4597	0.5084	0.4619	0.5023
maine_CrossEncoderFinetuned1^{rel}	0.9418	0.7517	0.6086	0.6861	0.7272	0.8730	0.7821
maine_CrossEncoderFinetuned1^{comb}	0.7959	0.6483	0.5552	0.5759	0.6324	0.8778	0.6645
maine_CrossEncoderFinetuned2^{rel}	0.8230	0.5207	0.4328	0.4883	0.5186	0.7296	0.5109
maine_CrossEncoderFinetuned2^{comb}	0.8346	0.5069	0.4379	0.4888	0.5270	0.7230	0.5042
maine_PI2TFIDF^{rel}	0.2022	0.1379	0.1069	0.1146	0.1201	0.2125	0.0868
maine_PI2TFIDF^{comb}	0.2191	0.1310	0.1069	0.1173	0.1226	0.2166	0.0891
maine_tripletloss^{rel}	0.5966	0.3793	0.2948	0.3461	0.3659	0.6041	0.3332
maine_tripletloss^{comb}	0.5629	0.3690	0.2966	0.3300	0.3611	0.6041	0.3292
uninib_DoSSIER_2	0.4802	0.2310	0.2086	0.2492	0.2625	0.2568	0.2449
uninib_DoSSIER_4	0.4462	0.2241	0.2069	0.2451	0.2596	0.2514	0.2384

have the potential to inform future research directions and aid the development of more user-friendly AI tools, leading to more accurate and effective retrieval of scientific literature for laypeople.

The comprehensive analysis of the CLEF 2023 SimpleText track leads to the overall conclusion that state-of-the-art models have made significant progress. However, it is evident that there is still substantial room for improvement in the field. This indicates that further advancements and refinements are necessary to enhance the performance and capabilities of the models.

Table 7
Text Analysis of SimpleText Task 1 output.

Run	Impact	#Refs	Length		FKGL	
			Mean	Median	Mean	Median
ElsevierSimpleText_run1	1.88	0.95	965.02	921.00	13.80	13.80
ElsevierSimpleText_run2	2.24	1.36	1017.57	981.00	13.98	13.90
ElsevierSimpleText_run3	1.80	0.94	951.64	912.00	13.71	13.75
ElsevierSimpleText_run4	2.10	1.21	1011.10	994.00	13.95	13.90
ElsevierSimpleText_run5	1.78	0.71	993.14	972.50	13.76	13.80
ElsevierSimpleText_run6	1.59	0.65	995.65	975.50	13.75	13.90
ElsevierSimpleText_run7	2.37	0.94	1101.23	1075.50	13.87	13.80
ElsevierSimpleText_run8	0.60	0.50	1089.90	1045.00	14.09	14.00
ElsevierSimpleText_run9	0.71	0.54	1016.96	991.00	13.66	13.70
UAms_CE100	3.20	1.64	1028.78	975.00	14.59	14.50
UAms_CE1k	2.41	1.24	1071.67	985.50	14.70	14.60
UAms_CE1k_Combine	0.84	0.49	924.38	839.00	10.84	11.20
UAms_CE1k_Filter	1.09	0.62	988.00	913.50	12.40	12.70
UAms_ElF_Cred44	3.32	1.62	973.03	970.50	13.60	14.50
UAms_ElF_Cred44Read	1.85	1.34	799.29	851.00	13.18	14.20
UAms_ElF_Cred53	2.89	1.49	938.41	932.00	13.73	14.40
UAms_ElF_Cred53Read	1.70	1.28	774.76	823.00	13.29	14.30
UAms_ElF_Read25	1.60	1.25	767.70	819.00	13.09	14.20
UAms_Elastic	2.84	1.45	922.36	917.00	13.49	14.30
maine_CrossEncoder1	4.22	2.86	961.17	923.00	14.64	14.60
maine_CrossEncoderFinetuned1	4.41	3.37	1003.75	988.00	15.01	14.80
maine_CrossEncoderFinetuned2	3.49	3.04	988.86	951.50	14.95	14.80
maine_PI2TFIDF	3.35	2.58	893.29	894.00	14.03	14.00
maine_tripletloss	4.76	3.29	969.09	973.50	14.69	14.60
unimib_DoSSIER_2	1.44	1.33	1024.48	994.00	14.77	14.60
unimib_DoSSIER_4	1.44	1.33	238.63	212.00	15.11	15.00

Acknowledgments

This research was funded, in whole or in part, by the French National Research Agency (ANR) under the project ANR-22-CE23-0019-01. We would like to thank Radia Hannachi, Silvia Araújo, Pierre De Loor, Olga Popova, Diana Nurbakova, Quentin Dubreuil, Aurianne Damoy, Angelique Robert, Julien Gaudin, and all other colleagues and participants who helped run this track.

References

[1] L. Ermakova, P. Bellot, P. Braslavski, J. Kamps, J. Mothe, D. Nurbakova, I. Ovchinnikova, E. SanJuan, Overview of simpletext 2021 - CLEF workshop on text simplification for scientific information access, in: CLEF'21: Experimental IR Meets Multilinguality, Multimodality, and Interaction - 12th International Conference of the CLEF Association,

volume 12880 of *Lecture Notes in Computer Science*, Springer, 2021, pp. 432–449. URL: https://doi.org/10.1007/978-3-030-85251-1_27.

[2] L. Ermakova, E. SanJuan, J. Kamps, S. Huet, I. Ovchinnikova, D. Nurbakova, S. Araújo, R. Hannachi, É. Mathurin, P. Bellot, Overview of the CLEF 2022 simpletext lab: Automatic simplification of scientific texts, in: CLEF'22: Experimental IR Meets Multilinguality, Multimodality, and Interaction - 13th International Conference of the CLEF Association, volume 13390 of *Lecture Notes in Computer Science*, Springer, 2022, pp. 470–494. URL: https://doi.org/10.1007/978-3-031-13643-6_28.

[3] L. Ermakova, H. Azarbonyad, S. Bertin, O. Augereau, Overview of the CLEF 2023 SimpleText Task 2: Difficult Concept Identification and Explanation, in: [14], 2023.

[4] L. Ermakova, S. Bertin, H. McCombie, J. Kamps, Overview of the CLEF 2023 SimpleText Task 3: Scientific text simplification, in: [14], 2023.

[5] L. Ermakova, E. SanJuan, S. Huet, H. Azarbonyad, O. Augereau, J. Kamps, Overview of the CLEF 2023 SimpleText Lab: Automatic simplification of scientific texts, in: A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, A. Giachanou, D. Li, M. Aliannejadi, M. Vlachos, G. Faggioli, N. Ferro (Eds.), CLEF'23: Proceedings of the Fourteenth International Conference of the CLEF Association, Lecture Notes in Computer Science, Springer, 2023.

[6] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, Z. Su, Arnetminer: Extraction and mining of academic social networks, in: KDD'08, 2008, pp. 990–998.

[7] J. Tang, A. C. Fong, B. Wang, J. Zhang, A unified probabilistic framework for name disambiguation in digital library, IEEE Transactions on Knowledge and Data Engineering 24 (2012) 975–987.

[8] A. Sinha, Z. Shen, Y. Song, H. Ma, D. Eide, B.-j. P. Hsu, K. Wang, An overview of microsoft academic service (mas) and applications, in: Proceedings of the 24th international conference on world wide web, ACM, 2015, pp. 243–246.

[9] J. Priem, H. Piwowar, R. Orr, Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts, 2022. arXiv:2205.01833.

[10] A. Capari, H. Azarbonyad, G. Tsatsaronis, Z. Afzal, Elsevier at SimpleText: Passage Retrieval by Fine-tuning GPL on Scientific Documents, in: [14], 2023.

[11] R. Hutter, J. Sutmuller, M. Adib, D. Rau, J. Kamps, University of Amsterdam at the CLEF 2023 SimpleText Track, in: [14], 2023.

[12] B. Mansouri, S. Durgin, S. Franklin, S. Fletcher, R. Campos, AIIR and LIAAD Labs Systems for CLEF 2023 SimpleText, in: [14], 2023.

[13] O. E. Mendoza, G. Pasi, Domain Context-centered Retrieval for the Content Selection task in the Simplification of Scientific Literature, in: [14], 2023.

[14] M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), Working Notes of CLEF 2023: Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS.org, 2023.