Overview of the CLEF 2024 SimpleText Task 1: Retrieve Passages to Include in 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 2024 SimpleText Task 1 on *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 CLEF 2024 SimpleText Task 1 on *Content Selection*, asking systems to retrieve scientific abstracts in response to a query prompted by a popular science article. This task performs a key element of the overall approach of the CLEF 2024 SimpleText Track.

The track as a whole offers valuable data and benchmarks to facilitate discussions on the challenges associated with automatic text simplification. It presents an interconnected framework that encompasses various tasks, providing a comprehensive view of the complexities involved:

Task 1 on *Content Selection*: retrieve passages to include in a simplified summary.

Task 2 on *Complexity Spotting*: identify and explain diffficult concepts.

Task 3 on *Text Simplification*: simplify scientific text.

Task 4 on *SOTA?*: tracking the state-of-the-art in scholarly publications.

This paper presents an overview of the first task in the SimpleText track on automatic simplification of scientific texts following up on the CLEF 2024 SimpleText Workshop. For a comprehensive overview of the other tasks, the task overview papers on Task 2 [1], Task 3 [2], and Task 4 [3], as well as the track overview paper [4], provide detailed information and further insights.

A total of 45 teams registered for our SimpleText track at CLEF 2024. A total of 20 teams submitted 207 runs in total for the Track, of which 11 teams submitted a total of 42 runs for Task 1. The statistics for the Task 1 runs submitted are presented in Table 1. However, some runs had problems that we could not resolve. We do not detail them in the paper as well as the 0-scored runs.

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

CLEF 2024: Conference and Labs of the Evaluation Forum, September 9—12, 2024, Grenoble, France
eric.sanjuan@univ-avignon.fr (É. SanJuan); liana.ermakova@univ-brest.fr (L. Ermakova)
https://simpletext-project.com/ (L. Ermakova)
0000-0002-6614-0087 (J. Kamps); 0000-0002-7598-7474 (L. Ermakova)

Table 1
CLEF 2024 Simpletext Task 1 official run submission statistics

Task	AIIR Lab	AMATU	Arampatzis	Elsevier	L3S	LIA	PiTheory	Sharigans	SINAI	SONAR	AB/DPV	Dajana/Katya	Frane/Andrea	Petra/Regina	Ruby	Tomislav/Rowan	UAmsterdam	UBO	UniPD	UZH Pandas	Total
1	5		9	10		5		1			1		1	1		2	6	1			42

adopted by the participants, specifically focusing on the scientific text. In Section 4, the official submissions' results are presented and discussed. Section 5 conducts a comprehensive analysis of the results, examining various significant aspects. Finally, Section 6 summarizes the findings and outlines potential directions for future research.

2. Task 1: Retrieve Passages to Include in a Simplified Summary

This section details *Task 1: Content Selection* on retrieve passages to include in a simplified summary.

2.1. Description

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.

2.2. 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 fields of computer science and engineering. We reuse the collection of academic abstracts from the Citation Network Dataset (12th version released in 2020)[1] [5]. This collection was extracted from DBLP, ACM, MAG (Microsoft Academic Graph), and other sources. It includes 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 [6].

In 2024, we added between 2 and 5 new queries (with IDs of the form G*.C*) for each of the 20 articles from the Guardian. These topics were generated by ChatGPT 4, with a prompt asking to list the main subtopics related to computer science; they were manually inspected to check they are linked to the original article and are not redundant. They are longer, containing around ten words and focusing on a specific point related to the article. An example of a keyword query is "*system on chip*" (T06.1) and an example of a long query is "*How AI systems, especially virtual assistants, can perpetuate gender stereotypes?*" (G01.C1).

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

The C1 queries were generated based on the following prompt: *In the attached article from the Guardian, list the main sub topics related to computer science and for each topic find at least five related references to scientific publications before 2019 that would have been relevant to be cited in this article. Just provide the references, don't try to get the full text.* We then considered as query the first sub topic. We also considered to use ChatGPT results as a complete run, but few references were returned, many were not indexed in computer science and some did not even exist. That emphasizes the real difficulty of the task of retrieving references to be included in a popular science article.

2.3. Baselines

An ElasticSearch index is provided to participants with access through an API. A JSON dump of the index is also available for participants. This index can be accessed online through queries, e.g. https://clef.termwatch.eu/dblp1/_search?q=biases&size=1000 for the query "Biases."

We additionally provided two supplementary baselines leveraging bag-of-words models and sparse vector document representations. The first baseline (denoted by "meili" in the results tables) was generated using the Meilisearch system[2] and relies on a bucket sort approach. The second baseline (denoted by "boolean") was constructed using a simple boolean model powered by PostgreSQL GIN text indexing.

For each topic, organizers manually assessed each proposed keyword query retrieved by the baseline run powered by Elasticsearch, ensuring that it retrieved at least five relevant documents. As a consequence, this boolean system, which retrieves all abstracts containing all query keywords within the abstract, is expected to artificially achieve high recall levels at a depth of 5. However, this approach suffers from two limitations: it misses relevant abstracts that do not contain all keywords, and it retrieves irrelevant abstracts that happen to contain all query keywords.

In the case of the long C1 queries, we manually extracted the largest subset of terms that retrieved at least five relevant documents. For these queries, the boolean approach is essentially a manual run, which is indicated by an asterisk (*) in the results tables.

Despite their effectiveness, neural models are computationally expensive, requiring significant training data and processing power. Consequently, most participants rely on a hybrid document retrieval approach. This approach leverages a two-stage process:

1. Initial Retrieval: This phase employs a more traditional and less resource-intensive method, such as tf-idf vectorization. This initial retrieval identifies a set of potentially relevant documents.
2. Re-ranking: The documents retrieved in the first stage are then re-ranked using the more nuanced dense representations provided by neural models. This step refines the initial retrieval results based on the semantic understanding of the neural models.

In previous editions, participants relied on the provided ElasticSearch baseline for the initial retrieval phase. To enhance run diversity and address resource limitations, the organizers this year provided access to two vector databases containing pre-computed paragraph embeddings (for titles and abstracts). These vector databases enable to compare the efficiency of scientific document retrieval techniques using asymmetric sparse document retrieval (based on tf-idf) and symmetric dense passage retrieval (based on pre-computed embeddings).

Two embedding vectors were based on the paragraph cross-encoder MS MARCO Mini LM (all-MiniLM-L6-v2)[3]. These embeddings, along with a search API based on them, have been released to participants. Documents are ranked based on the dot product between the query and the abstract (vir_abstract) or the title (vir_title) using the pg_vector[4] PostgreSQL extension and an ivvflat dense vector index (k-means vector clustering with $\sqrt{|D|}$ centroids).

These dense vector and the boolean baselines can be accessed online through a CGI API[5] with three parameters:

[2] https://www.meilisearch.com/
[3] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2
[4] https://github.com/pgvector/pgvector
[5] https://clef.termwatch.eu/stvir_test

corpus : title, abstract or bool

phrase : text passage as query

length : number of results to be retrieved

In the case of non boolean query, this API generates the vector embedding of the query on the fly before retrieving results using SQL syntax. For example, for the query "Exploring the use of AI to improve success rates and speed in the pharmaceutical research field", the top 100 documents whose abstracts are most similar to the query (based on dot product) can be retrieved in JSON format using the following syntax:

```
https://clef.termwatch.eu/stvir\_test?
    corpus=abstract\\
    \&phrase=Exploring the use of AI to improve success rates and speed
        in the pharmaceutical research field\\
    \&length=100
```

In addition to the dot product similarity measure, we also experimented with cosine distance. However, this alternative approach yielded comparable results.

The Boolean and dense vector baselines are provided as a PostgreSQL database containing four tables:

1. Complete Documents (JSON): full documents in JSON format, enabling access to all content.
2. Textual Content (Boolean Search): title and abstracts of documents, facilitating efficient boolean search operations.
3. Title Embeddings: pre-computed dense vector representations (embeddings) of the document titles.
4. Truncated Abstract Embeddings: pre-computed dense vector representations (embeddings) of the first 110 tokens of each document's abstract.

2.4. Formats

Ad-hoc passage retrieval Participants should retrieve, for each topic and each query, DBLP abstracts related to the query and relevant to be inserted as a citation in the paper associated with the topic. We encourage participants to take into account passage complexity as well as its credibility/influentialness.

Open passage retrieval (optional) Participants are encouraged to extract supplementary relevant queries from the titles or content articles and to provide results based on these supplementary queries.

Output format Results should be provided in a TREC style JSON format with the following fields:

1. *run_id*: Run ID starting with *<team_id>_<task_id>_<method_used>*, e.g. *UBO_Task1_TFIDF*
2. *manual*: Whether the run is manual {0,1}
3. *topic_id*: Topic ID
4. *query_id*: Query ID used to retrieve the document (if one of the queries provided for the topic was used; 0 otherwise)
5. *doc_id*: ID of the retrieved document (to be extracted from the JSON output)
6. *rel_score*: Relevance score of the passage (in the [0-1] scale)
7. *comb_score*: General score that may combine relevance and other aspects: readability, citation measures…(in the [0-1] scale)
8. passage: Text of the selected passage

For each query, the maximum number of distinct DBLP references (*doc_id* field) must be 100 and the total length of passages should not exceed 1,000 tokens. The idea of taking into account complexity is to have passages easier to understand for non-experts, while the 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 scores and credibility scores using bibliometrics.

Here is an output format example:


```
[{
    "run_id":"UBO_Task1_TFIDF",
    "manual":0,
    "topic_id":"G01",
    "query_id":"G01.1",
    "doc_id":1564531496,
    "rel_score":0.97,
    "comb_score":0.85,
    "passage":"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."
},
{
    "run_id":"UBO_Task1_TFIDF",
    "manual":0,
    "topic_id":"G01",
    "query_id":"G01.1",
    "doc_id":3000234933,
    "rel_score":0.9,
    "comb_score":0.9,
    "passage":"People are becoming increasingly comfortable using Digital Assistants (DAs)
    ↪  to interact with services or connected objects"
},
{
    "run_id":"UBO_Task1_TFIDF",
    "manual":0,
    "topic_id":"G01",
    "query_id":"G01.2",
    "doc_id":1448624402,
    "rel_score":0.6,
    "comb_score":0.3,
    "passage":"As extensive experimental research has shown individuals suffer from diverse
    ↪  biases in decision-making."
}]
```


2.5. Evaluation

To assess topical relevance, we assigned a 0-2 score to each retrieved document based on its content alignment with the original article. To expand the training data for relevance judgments (qrels), we pooled all documents retrieved at depth 10 from all submitted systems. This approach significantly increased the size of the qrels by 9,990 documents, with a particular focus on newly introduced long queries for the Guardian corpus and T06-T11 queries that previously lacked relevance assessments.

Table 2 summarizes the test collection constructed for the CLEF 2024 SimpleText Task 1, in relation to the earlier years (note that earlier topics have been reused in the "train" data).

While generating the long C1 queries using state-of-the-art LLMs, we were surprised by the inability of these models, specifically ChatGPT4, to find relevant references in the computer science domain suitable for inclusion in large audience tech articles. This raises questions about the inherent difficulty of the task and the potential necessity of combining multiple retrieval systems to improve recall. This need was addressed by both participants and organizers this year.

Many participants employed multiple LLMs, not for initial retrieval, but as rerankers within their systems. Additionally, several participants utilized different implementations of BM25 compared to the one provided by the organizers for the retrieval stage. These novel end-to-end retrieval approaches,

Table 2
CLEF 2024 SimpleText Task 1 Test Collection Statistics.

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	G01–G15	29	728	338	237	44.9
2023 test	G16–G20, T01-T05	34	2260	357	1218	112.8
2024 train	G01–G20, T01-T05	64	3,675	768	1,655	95.5
2024 test	G1.C1–G10.C1, T06–T11	30	2,775	1,500	579	128.5
2024 test extended	G1-G10, T01-T20	96	6,463	2,491	1,036	104.1

coupled with the 4 new baselines provided, resulted in an unexpectedly high number of unassessed documents among the top ten retrieved documents per run. This phenomenon included queries from previous editions. For instance, among queries G01-G10, there were $3,843$ new documents not returned in the top ten of previous editions. Notably, 954 of these documents appeared relevant to at least one existing topic, and 576 were relevant to one of the newly introduced long C1 queries. This confirms the task's inherent difficulty but also demonstrates the potential to achieve high recall levels at depth 10.

In addition to topical relevance, we took into account other key aspects of the track, such as the text complexity and the credibility of the retrieved results. These evaluations were performed using automatic metrics.

3. Scientific Passage Retrieval Approaches

In this section, we discuss a range of scientific text retrieval approaches that have been applied by the participants of the track. A total of 11 teams submitted 42 runs in total.

AB/DPV Varadi and Bartulović [7] submitted 1 run for Task 1. They used our ElasticSearch API and took into account an FKGL readability score for their combined score.

Sharingans Ali et al. [8] also submitted 1 run. They experimented with the ColBERT neural ranker and used GPT 3.5 to select the most informative and concise passages for inclusion in the summary.

Tomislav/Rowan Mann and Mikulandric [9] submitted a total of 2 runs. They took the top 100 results retrieved by ElasticSearch. Then, they used cosine similarity on TF-IDF vectors as the relevance score and FKGL score as the combined score.

Petra/Regina Elagina and Vučić [10] submitted 1 run, for the first 3 queries only, with the same approach as the previous system.

AIIRLab Largey et al. [11] submitted a total of 5 runs and proposed several models. First, since input queries are short keyword terms, they used query expansion with LLaMA 3 and reranked the top 5,000 results retrieved by TF-IDF with a bi-encoder or a cross-encoder. Second, they applied LLaMa3 as a pairwise re-ranker. Third, they leveraged ElasticSearch with fine-tuned cross-encoders.

UBO Vendeville et al. [12] submitted a total of 1 run. They used PyTerrier[6] to retrieve documents from TF-IDF scores. Then, the MonoT5 reranker provided by PyTerrier was employed to reorder all extracted documents.

[6] https://pyterrier.readthedocs.io/

UAmsterdam Bakker et al. [13] submitted a total of 6 runs for Task 1. First, they focused on regular information retrieval effectiveness with 2 vanilla baseline runs on an Anserini index, using either BM25 or BM25+RM3, and 2 other runs generated with neural cross-encoder rerankings of these runs by an MS MARCO-trained ranker. Second, 2 further runs filter out the most complex abstracts per request, using the median FKGL readability measure.

Elsevier Capari et al. [14] submitted a total of 10 runs. Their approaches mainly centered on creating a ranking model. They started by assessing the performance of several models on a proprietary test collection of scientific papers. Then, the top-performing model was fine-tuned on a large set of unlabeled documents using the Generative Pseudo Labeling approach. They also experimented with generating new search queries.

LIA submitted a total of 5 runs as baselines for Task 1. All five have been included in the pool of results for qrel evaluation.

Ruby This team (No paper received) submitted a total of 1 run for Task 1. Their approach relies on ElasticSearch and a TF-IDF score.

Arampatzis This team (No paper received) submitted a total of 9 runs for Task 1. As these reports are very close, the Tables below only report their evaluation made on their first run.

4. Results

This section details the results of the task, for both the train and test data.

4.1. Released database

All data and results have been organized within a relational database, which will be released to all active participants. This release will facilitate:

- Computation of Diverse Scores.
- Addressing qrel Issues.
- Easy Generation of Supplementary Runs.

One particular benefit of the relational database is the ability to easily extend the qrels based on dense vector similarity and similarity thresholds. This capability is especially relevant given the observation that seemingly identical abstracts in the DBLP dataset appear with different relevance labels.

4.2. Train results

Table 3 shows the results of the CLEF 2023 Simpletext Task 1 using the training qrels provided to participants. These evaluation data encompass 64 queries over 25 topics (G01-G20 and T01-T05). Runs presented in the table are sorted by the main measure of the task, i.e. NDCG@10. Let us note that for this table, we only considered top results retrieved according to relevance scores. The scores of two competitive runs taking into account combined scores are also provided in the two last lines.

4.3. Test results

Table 4 still shows relevance evaluation, with a ranking by NDCG@10 on queries absent from the training qrels. This time, we included two separated lines for a single run, when top retrieved documents differ using combined score (*comb*) or relevance score (*rel*).

We also inspect how systems behave on two different subsets of queries used in the test. Tables 5 and 6 distinguish the scores on the long G*.C1 queries related to general computer science topics,

Table 3
Results for CLEF 2024 SimpleText Task 1 on the Train qrels: G01-G20 and T01-T05.

Run	MRR	Precision		NDCG		Bpref	MAP
		10	20	10	20		
AIIRLab_Task1_LLaMABiEncoder	0.7570	0.6467	0.4133	0.4955	0.4206	0.3463	0.2227
AIIRLab_Task1_LLaMAReranker2	0.7531	0.6200	0.4008	0.4708	0.4014	0.3364	0.2086
AIIRLab_Task1_LLaMAReranker	0.7459	0.6083	0.4000	0.4643	0.3983	0.3405	0.2070
AIIRLab_Task1_LLaMACrossEncoder	0.7849	0.5150	0.3675	0.4117	0.3732	0.3413	0.1985
LIA_vir_title	0.6680	0.4433	0.2758	0.3405	0.2766	0.2742	0.1191
AIIRLAB_Task1_CERRF	0.6329	0.4033	0.3083	0.3246	0.3030	0.2113	0.1375
Arampatzis_1.GPT2_search_results	0.5732	0.3933	0.1967	0.2972	0.2184	0.0876	0.0676
UAms_Task1_Anserini_rm3	0.5613	0.3817	0.2833	0.2805	0.2541	0.2842	0.1408
UAms_Task1_Anserini_bm25	0.5493	0.3733	0.3208	0.2803	0.2827	0.3003	0.1536
Elsevier@SimpleText_task_1_run8	0.6173	0.3633	0.2458	0.2800	0.2406	0.1673	0.0993
LIA_vir_abstract	0.6015	0.3867	0.2633	0.2795	0.2405	0.2738	0.1168
UAms_Task1_CE100_CAR	0.4800	0.3683	0.2958	0.2637	0.2591	0.2048	0.1207
UAms_Task1_CE100	0.4800	0.3683	0.2958	0.2637	0.2591	0.2048	0.1207
LIA_bool	0.5646	0.3517	0.2400	0.2552	0.2238	0.2134	0.1037
UAms_Task1_CE1K_CAR	0.4408	0.3483	0.2833	0.2419	0.2418	0.2030	0.1147
UAms_Task1_CE1K	0.4408	0.3483	0.2833	0.2419	0.2418	0.2778	0.1347
Ruby_Task_1	0.5231	0.3050	0.2425	0.2387	0.2281	0.1696	0.1018
Elsevier@SimpleText_task_1_run10	0.5072	0.2983	0.2000	0.2335	0.1983	0.1356	0.0815
Elsevier@SimpleText_task_1_run4	0.5360	0.3100	0.2292	0.2285	0.2081	0.1476	0.0874
LIA_elastic	0.4540	0.2817	0.2067	0.2213	0.1977	0.2275	0.1103
AB/DPV_SimpleText_task1_FKGL	0.4538	0.2817	0.2067	0.2213	0.1977	0.1623	0.0948
Elsevier@SimpleText_task_1_run6	0.4686	0.2900	0.2283	0.2145	0.2014	0.1699	0.0947
Tomislav/Rowan_SimpleText_T1_1	0.5023	0.2683	0.1933	0.2108	0.1910	0.0972	0.0650
Elsevier@SimpleText_task_1_run1	0.5263	0.2400	0.2267	0.2108	0.2150	0.1733	0.1011
Elsevier@SimpleText_task_1_run5	0.4464	0.2767	0.2200	0.2023	0.1919	0.1664	0.0913
LIA_meili	0.4372	0.2883	0.1792	0.1833	0.1570	0.2024	0.0691
Elsevier@SimpleText_task_1_run9	0.4149	0.2583	0.1775	0.1833	0.1622	0.1240	0.0645
Elsevier@SimpleText_task_1_run7	0.3731	0.2367	0.1717	0.1735	0.1577	0.1194	0.0588
UBO_Task1_TFIDFT5	0.4134	0.1933	0.1775	0.1621	0.1625	0.1647	0.0730
Elsevier@SimpleText_task_1_run3	0.3879	0.1700	0.1508	0.1498	0.1469	0.1246	0.0654
Elsevier@SimpleText_task_1_run2	0.3186	0.1500	0.1742	0.1279	0.1446	0.1416	0.0696
Sharingans_Task1_marco-GPT3	0.4167	0.0417	0.0208	0.0658	0.0466	0.0085	0.0085
Tomislav/Rowan_SimpleText_T1_2	0.0108	0.0100	0.0067	0.0057	0.0051	0.0030	0.0011
Petra/Regina_results_simpleText_task_1	0.0013	0.0000	0.0025	0.0000	0.0018	0.0016	0.0004
UAms_Task1_CE100_CAR†	0.6709	0.4687	0.3937	0.4530	0.3972	0.3144	0.1922
UAms_Task1_CE1K_CAR†	0.6403	0.4219	0.3672	0.4032	0.3646	0.3411	0.1904

† Evaluated on comb_score.

which can be a source of public debates (like privacy, quantum computing, bitcoins...) from those established on the short Tech Xplore queries, which are more specific and related with a scientific paper in peer-reviewed venues (indoor positioning system, RISC-V architecture for space computing, underwater WiFi developed using LEDs and lasers...). Rankings on these two subsets are very similar, which shows the consistency of relevance results across queries.

Table 4
Results for CLEF 2024 SimpleText Task 1 on the Test qrels (G01.C1-G10.C1 and T06-T11).

Run	MRR	Precision		NDCG		Bpref	MAP
		10	20	10	20		
AIIRLab_Task1_LLaMABiEncoder^{rel}	0.9444	0.8167	0.5517	0.6311	0.5240	0.3559	0.2304
AIIRLab_Task1_LLaMAReranker2^{rel}	0.9300	0.7933	0.5417	0.6092	0.5082	0.3495	0.2177
AIIRLab_Task1_LLaMAReranker2^{comb}	0.9361	0.7833	0.5450	0.6063	0.5107	0.3494	0.2192
AIIRLab_Task1_LLaMAReranker^{rel}	0.8944	0.7967	0.5583	0.5991	0.5070	0.3541	0.2200
AIIRLab_Task1_LLaMAReranker^{comb}	0.9111	0.7833	0.5600	0.5982	0.5112	0.3543	0.2217
LIA_vir_title	0.8454	0.6933	0.4383	0.5090	0.4010	0.3594	0.1534
AIIRLab_Task1_LLaMACrossEncoder^{rel}	0.7975	0.6933	0.5100	0.4879	0.4335	0.3404	0.1970
LIA_vir_abstract	0.7683	0.6000	0.4067	0.4269	0.3539	0.3857	0.1603
UAms_Task1_Anserini_rm3	0.7878	0.5700	0.4350	0.3945	0.3506	0.4010	0.1824
UAms_Task1_Anserini_bm25	0.7187	0.5500	0.4883	0.3774	0.3721	0.3994	0.1972
UAms_Task1_CE1K^{comb}	0.5950	0.5333	0.4583	0.3726	0.3659	0.4032	0.1939
UAms_Task1_CE1K_CAR^{rel}	0.5950	0.5333	0.4583	0.3726	0.3659	0.2701	0.1605
UAms_Task1_CE1K^{rel}	0.5950	0.5333	0.4583	0.3726	0.3659	0.4032	0.1939
UAms_Task1_CE100^{rel}	0.6618	0.5300	0.4567	0.3705	0.3579	0.2657	0.1579
UAms_Task1_CE100_CAR^{rel}	0.6618	0.5300	0.4567	0.3705	0.3579	0.2657	0.1579
UAms_Task1_CE100^{comb}	0.6618	0.5300	0.4567	0.3705	0.3579	0.2657	0.1579
UAms_Task1_CE1K_CAR^{comb}	0.6611	0.5133	0.3400	0.3654	0.2998	0.2676	0.1348
AIIRLAB_Task1_CERRF	0.7264	0.5033	0.4000	0.3584	0.3239	0.2204	0.1309
Arampatzis_1.GPT2_search^{rel}	0.6986	0.5100	0.2550	0.3522	0.2465	0.0742	0.0577
UBO_Task1_TFIDFT5	0.7132	0.4833	0.3817	0.3506	0.3215	0.2354	0.1274
Arampatzis_1.GPT2_search^{comb}	0.6814	0.5100	0.2550	0.3449	0.2423	0.0741	0.0563
LIA_bool*	0.7242	0.5233	0.3633	0.3409	0.2906	0.2661	0.1199
AIIRLab_Task1_LLaMACrossEncoder^{comb}	0.6609	0.4867	0.4100	0.3363	0.3281	0.3364	0.1579
AIIRLab_Task1_LLaMABiEncoder^{comb}	0.6078	0.4867	0.3783	0.3246	0.3024	0.3393	0.1474
UAms_Task1_CE100_CAR^{comb}	0.6420	0.4700	0.3433	0.3236	0.2790	0.2657	0.1321
Elsevier@SimpleText_task_1_run8	0.7123	0.4533	0.3367	0.3152	0.2755	0.1582	0.0906
Elsevier@SimpleText_task_1_run4	0.6162	0.4300	0.3217	0.3075	0.2692	0.1642	0.1005
Elsevier@SimpleText_task_1_run10	0.5117	0.4067	0.2767	0.2949	0.2401	0.1236	0.0729
LIA_elastic	0.6173	0.3733	0.2900	0.2818	0.2442	0.3016	0.1325
AB&DPV_SimpleText_task1_FKGL^{rel}	0.6173	0.3733	0.2900	0.2818	0.2442	0.1966	0.1078
Ruby_Task_1^{rel}	0.5470	0.4233	0.3533	0.2790	0.2688	0.1980	0.1110
LIA_meili	0.6386	0.4700	0.2867	0.2736	0.2242	0.2377	0.0833
Elsevier@SimpleText_task_1_run6	0.5333	0.3833	0.3117	0.2654	0.2445	0.1841	0.0973
Ruby_Task_1^{comb}	0.5910	0.3767	0.3000	0.2641	0.2407	0.1961	0.0980
Tomislav/Rowan&Rowan_SimpleText_T1_1^{rel}	0.5444	0.3733	0.2750	0.2477	0.2201	0.0963	0.0601
Elsevier@SimpleText_task_1_run5	0.4867	0.3533	0.2883	0.2415	0.2238	0.1834	0.0943
Elsevier@SimpleText_task_1_run1	0.5589	0.3000	0.3300	0.2262	0.2407	0.1978	0.1018
Tomislav/Rowan&Rowan_SimpleText_T1_1^{comb}	0.5309	0.3200	0.2333	0.2044	0.1790	0.0938	0.0509
Elsevier@SimpleText_task_1_run7	0.4026	0.3200	0.2250	0.2039	0.1733	0.1085	0.0565
Elsevier@SimpleText_task_1_run9	0.3868	0.3300	0.2283	0.1971	0.1710	0.1103	0.0590
Elsevier@SimpleText_task_1_run3	0.4733	0.2367	0.2033	0.1872	0.1712	0.1587	0.0714
Elsevier@SimpleText_task_1_run2	0.4193	0.2233	0.2433	0.1812	0.1878	0.1768	0.0820
AB&DPV_SimpleText_task1_FKGL^{comb}	0.4380	0.2333	0.2100	0.1476	0.1496	0.1909	0.0667
Sharingans_Task1_marco-GPT3	0.6667	0.0667	0.0333	0.1167	0.0807	0.0107	0.0107
Tomislav/Rowan&Rowan_SimpleText_T1_2	0.0217	0.0233	0.0150	0.0156	0.0124	0.0062	0.0025
Petra&Regina_simpleText_task_1	0.0026	0.0000	0.0050	0.0000	0.0035	0.0031	0.0007

Table 5
Results for CLEF 2024 SimpleText Task 1 on the Test qrels limited to G01.C1-G10.C1.

Run	MRR	Precision		NDCG		Bpref	MAP
		10	20	10	20		
AIIRLab_Task1_LLaMABiEncoder^{rel}	0.9500	0.7600	0.5125	0.5546	0.4777	0.3150	0.1919
AIIRLab_Task1_LLaMAReranker2^{rel}	0.9350	0.7450	0.5200	0.5366	0.4733	0.3115	0.1868
AIIRLab_Task1_LLaMAReranker2^{comb}	0.9375	0.7400	0.5225	0.5334	0.4740	0.3111	0.1887
AIIRLab_Task1_LLaMAReranker^{rel}	0.8667	0.7500	0.5300	0.5292	0.4657	0.3127	0.1845
AIIRLab_Task1_LLaMAReranker^{comb}	0.8917	0.7400	0.5300	0.5260	0.4678	0.3126	0.1854
AIIRLab_Task1_LLaMACrossEncoder^{rel}	0.7892	0.6650	0.4750	0.4399	0.3957	0.3032	0.1667
LIA_vir_title	0.8014	0.6100	0.3750	0.4043	0.3307	0.2793	0.0985
LIA_bool*	0.7613	0.5800	0.4175	0.3531	0.3194	0.3384	0.1452
AIIRLab_Task1_LLaMABiEncoder^{comb}	0.6500	0.5750	0.4225	0.3526	0.3268	0.2988	0.1433
AIIRLab_Task1_LLaMACrossEncoder^{comb}	0.6767	0.5650	0.4500	0.3503	0.3448	0.2968	0.1504
LIA_meili	0.7017	0.6100	0.3800	0.3477	0.2929	0.3175	0.1145
UAms_Task1_Anserini_rm3	0.7150	0.5250	0.4075	0.3248	0.3078	0.3486	0.1463
AIIRLAB_Task1_CERRF	0.6800	0.4950	0.3975	0.3159	0.3047	0.1943	0.1104
UAms_Task1_Anserini_bm25	0.6364	0.5050	0.4600	0.3060	0.3269	0.3549	0.1651
LIA_vir_abstract	0.6774	0.4900	0.3025	0.3053	0.2537	0.3020	0.0906
Arampatzis_1.GPT2_search^{comb}	0.6588	0.4900	0.2450	0.3050	0.2237	0.0651	0.0476
UAms_Task1_CE1K_CAR^{comb}	0.5583	0.5000	0.3300	0.3042	0.2630	0.2297	0.1091
Arampatzis_1.GPT2_search^{rel}	0.6333	0.4900	0.2450	0.2993	0.2193	0.0646	0.0453
Elsevier@SimpleText_task_1_run8	0.6780	0.4400	0.2950	0.2847	0.2424	0.1131	0.0614
UBO_Task1_TFIDFT5	0.6198	0.4500	0.3425	0.2774	0.2610	0.1911	0.0903
UAms_Task1_CE1K_CAR^{rel}	0.4592	0.4700	0.4200	0.2642	0.2931	0.2289	0.1221
UAms_Task1_CE1K	0.4592	0.4700	0.4200	0.2642	0.2931	0.3573	0.1539
UAms_Task1_CE100_CAR^{comb}	0.5880	0.4450	0.3225	0.2574	0.2383	0.2341	0.1074
Ruby_Task_1^{rel}	0.5550	0.4100	0.3600	0.2546	0.2587	0.1677	0.0966
UAms_Task1_CE100^{comb}	0.5260	0.4550	0.4000	0.2515	0.2750	0.2328	0.1217
UAms_Task1_CE100	0.5260	0.4550	0.4000	0.2515	0.2750	0.2328	0.1217
Tomislav/Rowan&Rowan_SimpleText_T1_1^{rel}	0.5550	0.4000	0.3200	0.2467	0.2380	0.1125	0.0675
Ruby_Task_1^{comb}	0.5510	0.3850	0.3150	0.2409	0.2310	0.1649	0.0840
Elsevier@SimpleText_task_1_run4	0.5297	0.3350	0.2400	0.2153	0.1933	0.0923	0.0452
LIA_elastic	0.5163	0.3000	0.2325	0.2010	0.1851	0.2540	0.0988
AB&DPV_SimpleText_task1_FKGL^{rel}	0.5163	0.3000	0.2325	0.2010	0.1851	0.1589	0.0762
Tomislav/Rowan&Rowan_SimpleText_T1_1^{comb}	0.4839	0.3300	0.2575	0.1977	0.1846	0.1088	0.0560
Elsevier@SimpleText_task_1_run10	0.3786	0.3150	0.2225	0.1906	0.1694	0.0891	0.0407
Elsevier@SimpleText_task_1_run6	0.4898	0.2950	0.2325	0.1790	0.1712	0.1343	0.0539
AB&DPV_SimpleText_task1_FKGL^{comb}	0.4048	0.2600	0.2200	0.1446	0.1481	0.1560	0.0608
Elsevier@SimpleText_task_1_run9	0.2848	0.2900	0.1925	0.1356	0.1226	0.0838	0.0366
Elsevier@SimpleText_task_1_run5	0.3961	0.2400	0.1950	0.1330	0.1359	0.1294	0.0482
Elsevier@SimpleText_task_1_run1	0.4550	0.1700	0.2800	0.1306	0.1884	0.1622	0.0722
Elsevier@SimpleText_task_1_run7	0.3148	0.2500	0.1750	0.1250	0.1164	0.0815	0.0311
Elsevier@SimpleText_task_1_run3	0.3725	0.1200	0.1325	0.1022	0.1098	0.1306	0.0435
Sharingans_Task1_marco-GPT3	0.5000	0.0500	0.0250	0.0816	0.0589	0.0070	0.0070
Elsevier@SimpleText_task_1_run2	0.2995	0.0800	0.1600	0.0679	0.1091	0.1209	0.0384

Table 6
Results for CLEF 2024 SimpleText Task 1 on the Test qrels limited to T06-T11.

Run	MRR	Precision		NDCG		Bpref	MAP
		10	20	10	20		
AIIRLab_Task1_LLaMABiEncoder^{rel}	0.9500	0.7600	0.5125	0.5546	0.4777	0.3150	0.1919
AIIRLab_Task1_LLaMAReranker2^{rel}	0.9350	0.7450	0.5200	0.5366	0.4733	0.3115	0.1868
AIIRLab_Task1_LLaMAReranker2^{comb}	0.9375	0.7400	0.5225	0.5334	0.4740	0.3111	0.1887
AIIRLab_Task1_LLaMAReranker^{rel}	0.8667	0.7500	0.5300	0.5292	0.4657	0.3127	0.1845
AIIRLab_Task1_LLaMAReranker^{comb}	0.8917	0.7400	0.5300	0.5260	0.4678	0.3126	0.1854
AIIRLab_Task1_LLaMACrossEncoder^{rel}	0.7892	0.6650	0.4750	0.4399	0.3957	0.3032	0.1667
LIA_vir_title	0.8014	0.6100	0.3750	0.4043	0.3307	0.2793	0.0985
LIA_bool	0.7613	0.5800	0.4175	0.3531	0.3194	0.3384	0.1452
AIIRLab_Task1_LLaMABiEncoder^{comb}	0.6500	0.5750	0.4225	0.3526	0.3268	0.2988	0.1433
AIIRLab_Task1_LLaMACrossEncoder^{comb}	0.6767	0.5650	0.4500	0.3503	0.3448	0.2968	0.1504
LIA_meili	0.7017	0.6100	0.3800	0.3477	0.2929	0.3175	0.1145
UAms_Task1_Anserini_rm3	0.7150	0.5250	0.4075	0.3248	0.3078	0.3486	0.1463
AIIRLAB_Task1_CERRF	0.6800	0.4950	0.3975	0.3159	0.3047	0.1943	0.1104
UAms_Task1_Anserini_bm25	0.6364	0.5050	0.4600	0.3060	0.3269	0.3549	0.1651
LIA_vir_abstract	0.6774	0.4900	0.3025	0.3053	0.2537	0.3020	0.0906
Arampatzis_1.GPT2_search^{comb}	0.6588	0.4900	0.2450	0.3050	0.2237	0.0651	0.0476
UAms_Task1_CE1K_CAR^{comb}	0.5583	0.5000	0.3300	0.3042	0.2630	0.2297	0.1091
Arampatzis_1.GPT2_search^{rel}	0.6333	0.4900	0.2450	0.2993	0.2193	0.0646	0.0453
Elsevier@SimpleText_task_1_run8	0.6780	0.4400	0.2950	0.2847	0.2424	0.1131	0.0614
UBO_Task1_TFIDFT5	0.6198	0.4500	0.3425	0.2774	0.2610	0.1911	0.0903
UAms_Task1_CE1K_CAR	0.4592	0.4700	0.4200	0.2642	0.2931	0.2289	0.1221
UAms_Task1_CE1K	0.4592	0.4700	0.4200	0.2642	0.2931	0.3573	0.1539
UAms_Task1_CE100_CAR^{comb}	0.5880	0.4450	0.3225	0.2574	0.2383	0.2341	0.1074
Ruby_Task_1^{rel}	0.5550	0.4100	0.3600	0.2546	0.2587	0.1677	0.0966
UAms_Task1_CE100_CAR^{rel}	0.5260	0.4550	0.4000	0.2515	0.2750	0.2328	0.1217
UAms_Task1_CE100	0.5260	0.4550	0.4000	0.2515	0.2750	0.2328	0.1217
Tomislav/Rowan&Rowan_SimpleText_T1_1^{rel}	0.5550	0.4000	0.3200	0.2467	0.2380	0.1125	0.0675
Ruby_Task_1^{comb}	0.5510	0.3850	0.3150	0.2409	0.2310	0.1649	0.0840
Elsevier@SimpleText_task_1_run4	0.5297	0.3350	0.2400	0.2153	0.1933	0.0923	0.0452
LIA_elastic	0.5163	0.3000	0.2325	0.2010	0.1851	0.2540	0.0988
AB&DPV_SimpleText_task1_FKGL^{rel}	0.5163	0.3000	0.2325	0.2010	0.1851	0.1589	0.0762
Tomislav/Rowan&Rowan_SimpleText_T1_1^{comb}	0.4839	0.3300	0.2575	0.1977	0.1846	0.1088	0.0560
Elsevier@SimpleText_task_1_run10	0.3786	0.3150	0.2225	0.1906	0.1694	0.0891	0.0407
Elsevier@SimpleText_task_1_run6	0.4898	0.2950	0.2325	0.1790	0.1712	0.1343	0.0539
AB&DPV_SimpleText_task1_FKGL^{comb}	0.4048	0.2600	0.2200	0.1446	0.1481	0.1560	0.0608
Elsevier@SimpleText_task_1_run9	0.2848	0.2900	0.1925	0.1356	0.1226	0.0838	0.0366
Elsevier@SimpleText_task_1_run5	0.3961	0.2400	0.1950	0.1330	0.1359	0.1294	0.0482
Elsevier@SimpleText_task_1_run1	0.4550	0.1700	0.2800	0.1306	0.1884	0.1622	0.0722
Elsevier@SimpleText_task_1_run7	0.3148	0.2500	0.1750	0.1250	0.1164	0.0815	0.0311
Elsevier@SimpleText_task_1_run3	0.3725	0.1200	0.1325	0.1022	0.1098	0.1306	0.0435
Sharingans_Task1_marco-GPT3	0.5000	0.0500	0.0250	0.0816	0.0589	0.0070	0.0070
Elsevier@SimpleText_task_1_run2	0.2995	0.0800	0.1600	0.0679	0.1091	0.1209	0.0384

Table 7
Evaluation of complexity and credibility for SimpleText Task 1 (over all 176 queries).

Run	Avg #Refs	Avg size of vocabulary	Ratio of long words	Ratio of complex words	FKGL avg	FKGL median
AB/DPV_SimpleText_task1_FKGL^{comb}	9.7	91.6	0.421	0.533	20.4	19.5
AB/DPV_SimpleText_task1_FKGL^{rel}	9.2	92.9	0.384	0.505	15.3	15.1
AIIRLAB_Task1_CERRF^{comb}	10.6	96.4	0.386	0.503	15.3	15.1
AIIRLab_Task1_LLaMABiEncoder^{comb}	9.5	98.1	0.411	0.515	20.7	20
AIIRLab_Task1_LLaMABiEncoder^{rel}	8.7	95.8	0.375	0.485	15.3	15.1
AIIRLab_Task1_LLaMACrossEncoder^{comb}	10.0	99.4	0.411	0.513	20.4	19.7
AIIRLab_Task1_LLaMACrossEncoder^{rel}	10.7	104.3	0.378	0.485	15.5	15.3
AIIRLab_Task1_LLaMAReranker^{comb}	8.8	96.1	0.377	0.487	15.7	15.4
AIIRLab_Task1_LLaMAReranker^{rel}	8.8	95.8	0.376	0.486	15.5	15.2
AIIRLab_Task1_LLaMAReranker2^{comb}	8.6	93.9	0.378	0.489	15.5	15.3
AIIRLab_Task1_LLaMAReranker2^{rel}	8.6	94	0.376	0.487	15.3	15.1
Arampatzis_1.GPT2_searchs	10.5	91.9	0.392	0.511	15.7	15.1
Elsevier@SimpleText_task_1_run1	10.0	92.6	0.385	0.498	15.4	15
Elsevier@SimpleText_task_1_run10	10.2	92.1	0.38	0.499	15.2	14.8
Elsevier@SimpleText_task_1_run2	11.2	99.1	0.38	0.495	15.2	15.1
Elsevier@SimpleText_task_1_run3	9.7	90.4	0.384	0.494	15.3	15
Elsevier@SimpleText_task_1_run4	10.7	99.1	0.375	0.495	15.1	14.9
Elsevier@SimpleText_task_1_run5	11.1	98	0.377	0.492	15	14.9
Elsevier@SimpleText_task_1_run6	11.2	96.7	0.378	0.492	15.2	15
Elsevier@SimpleText_task_1_run7	9.8	100.4	0.368	0.492	14.8	14.8
Elsevier@SimpleText_task_1_run8	10.3	94.4	0.387	0.504	15.5	15.3
Elsevier@SimpleText_task_1_run9	10.6	98.2	0.364	0.487	14.7	14.5
LIA_bool	13.0	139.3	0.388	0.513	20.9	17
LIA_elastic	9.2	92.9	0.384	0.505	15.3	15.1
LIA_meili	9.6	115.1	0.383	0.502	17.1	15.5
LIA_vir_abstract	7.2	69.8	0.378	0.484	14.6	14.3
LIA_vir_title	9.8	90.4	0.372	0.483	15	14.7
Petra/Reginas_simpleText_task1	5.5	86.1	0.386	0.509	15.4	15.3
Ruby_Task_1^{comb}	9.6	101.2	0.36	0.484	14	13.7
Ruby_Task_1^{rel}	9.7	92.9	0.389	0.503	15.9	15.2
Sharingans_Task1_marco-GPT3	9.8	59.8	0.373	0.436	15.5	15.5
Tomislav/Rowan_SimpleText_T1_1^{comb}	11.3	97.6	0.408	0.519	17.2	16.6
Tomislav/Rowan_SimpleText_T1_1^{rel}	9.9	93.2	0.391	0.505	15.9	15.4
Tomislav/Rowan_SimpleText_T1_2^{comb}	11.8	97.1	0.397	0.501	16.4	16.1
Tomislav/Rowan_SimpleText_T1_2^{rel}	10.9	91.1	0.392	0.496	15.8	15.7
Uams_Task1_Anserini_bm25	11.8	111.4	0.385	0.506	16.2	15.3
Uams_Task1_Anserini_rm3	11.9	112.9	0.387	0.508	16.8	16
UAms_Task1_CE100_CAR^{comb}	10.6	102.5	0.363	0.485	13.5	13.5
Uams_Task1_CE100	11.1	103.1	0.388	0.501	15.6	15.3
UAms_Task1_CE1K_CAR^{comb}	10.2	98.5	0.363	0.483	13.8	13.5
Uams_Task1_CE1K	10.8	101.4	0.387	0.499	15.9	15.4
UBO_Task1_TFIDFT5	10.3	99.2	0.386	0.498	15.4	15.2

5. Analysis

This section provides further analysis of the submitted runs, and the task as whole.

We complement the evaluation above by taking into consideration other aspects essential for Task 1. Table 7 highlights credibility and text complexity. We used simple automatic metrics to provide an overview of the importance and the complexity of the article. First, the average number of bibliographic references among the top 10 results of each query is provided. Second, we provide several metrics provided by the Python library readability[7]: the average size of vocabulary per abstract, the average ratio of words considered as long (i.e., with at least 7 characters), the average ratio of words considered

[7]https://pypi.org/project/readability/.

as complex (i.e., absent from the Dale-Chall word list of 3,000 words recognized by 80 % of fifth graders) and the averaged and median FKGL readability metrics.

A large majority of runs have a similar FKGL of 15, corresponding to university level texts, which can be expected since the document deals with advanced scientific topics. However, AIIRLab runs obtained with bi- or cross-encoders and ordered according to comb scores exhibit a significant higher FKGL readability scores. This difference is related to longer sentences retrieved with this score that with relevance score (average length of 31 words vs 23 words).

Only one run (Sharingans_Task1_marco-GPT3) provided a rephrased extract from the retrieved abstracts, while other runs gave the abstracts in full. This feature translates in the Table in a lower size of vocabulary in their passages.

6. Discussion and Conclusions

This concludes the results for the CLEF 2024 SimpleText Task 1: Content Selection on retrieve passages to include in a simplified summary. Our main findings are the following: First, the Tables on relevance are dominated by neural rankers, in particular, cross-encoders and LLaMA 3 used as a pairwise re-ranker. Second, a majority of participants relied on ElasticSearch search results. If neural models used in processing steps leveraged these results, other IR systems turned out to be competitive. For instance, LIA_vir_title operating with embedding sentences or UAms_Task1_Anserini_rm3, using an Anserini index have high relevance evaluations. Third, as expected, ranking over systems differs according to the considered criterion. Runs filtered against readability measures tend to have shorter sentences with a more or less drop in relevance. Remarkably, LLaMA 3 used as a reranker seems to not only help to select more relevant documents but also with more concise sentences.

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 and evaluation. Please visit the SimpleText website for more details on the track.[8]

Liana Ermakova is funded by the French National Research Agency (ANR) Automatic Simplification of Scientific Texts project (ANR-22-CE23-0019-01),[9] *and the MaDICS research group.*[10]

References

[1] G. M. Di Nunzio, F. Vezzani, V. Bonato, H. Azarbonyad, J. Kamps, L. Ermakova, Overview of the CLEF 2024 SimpleText task 2: Identify and explain difficult concepts, in: [15], 2024.

[2] L. Ermakova, V. Laimé, H. McCombie, J. Kamps, Overview of the CLEF 2024 SimpleText task 3: Simplify scientific text, in: [15], 2024.

[3] J. D'Souza, S. Kabongo, H. B. Giglou, Y. Zhang, Overview of the CLEF 2024 SimpleText Task 4: SOTA? Tracking the State-of-the-Art in Scholarly Publications, in: [15], 2024.

[4] L. Ermakova, E. SanJuan, S. Huet, H. Azarbonyad, G. M. Di Nunzio, F. Vezzani, J. D'Souza, J. Kamps, Overview of the CLEF 2024 SimpleText track: Improving access to scientific texts for everyone, in: L. Goeuriot, G. Q. Philippe Mulhem, D. Schwab, L. Soulier, G. M. D. Nunzio, P. Galuščáková, A. G. S. de Herrera, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024), Lecture Notes in Computer Science, Springer, 2024.

[8] https://simpletext-project.com/

[9] https://anr.fr/Project-ANR-22-CE23-0019

[10] https://www.madics.fr/ateliers/simpletext/

[5] 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.
[6] E. SanJuan, S. Huet, J. Kamps, L. Ermakova, Overview of the CLEF 2023 simpletext task 1: Passage selection for a simplified summary, in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), Thessaloniki, Greece, September 18th to 21st, 2023, volume 3497 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 2823–2834. URL: https://ceur-ws.org/Vol-3497/paper-238.pdf.
[7] D. P. Varadi, A. Bartulović, SimpleText 2024: Scientific Text Made Simpler Through the Use of AI, in: [15], 2024.
[8] S. M. Ali, H. Sajid, O. Aijaz, O. Waheed, F. Alvi, A. Samad, Improving Scientific Text Comprehension: A Multi-Task Approach with GPT-3.5 Turbo and Neural Ranking, in: [15], 2024.
[9] R. Mann, T. Mikulandric, CLEF 2024 SimpleText Tasks 1-3: Use of LLaMA-2 for text simplification, in: [15], 2024.
[10] R. Elagina, P. Vučić, AI Contributions to Simplifying Scientific Discourse in SimpleText 2024, in: [15], 2024.
[11] N. Largey, R. Maarefdoust, S. Durgin, B. Mansouri, AIIR Lab Systems for CLEF 2024 SimpleText: Large Language Models for Text Simplification, in: [15], 2024.
[12] B. Vendeville, L. Ermakova, P. De Loor, UBO NLP report on the SimpleText track at CLEF 2024, in: [15], 2024.
[13] J. Bakker, G. Yüksel, J. Kamps, University of Amsterdam at the CLEF 2024 SimpleText Track, in: [15], 2024.
[14] A. Capari, H. Azarbonyad, G. Tsatsaronis, Z. Afzal, Enhancing Scientific Document Simplification through Adaptive Retrieval and Generative Models, in: [15], 2024.
[15] G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024: Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS.org, 2024.