Complexity-Aware Scientific Literature Search
Searching for Relevant and Accessible Scientific Text

Liana Ermakova[†], **Jaap Kamps**[‡]
[†]Université de Bretagne Occidentale, HCTI, France, liana.ermakova@univ-brest.fr
[‡]University of Amsterdam, The Netherlands, kamps@uva.nl

Abstract

We conduct a series of experiments on ranking scientific abstracts in response to popular science queries issued by laypersons. We show that standard IR ranking models optimized on topical relevance are indeed ignoring the individual user's context and background knowledge. We also demonstrate the viability of complexity-aware retrieval models that retrieve more accessible relevant documents or ensure these are ranked prior to more complex documents on the topic. More generally, our results help remove some of the barriers to consulting scientific literature by laypersons and hold the potential to promote science literacy in the general public.

Lay Summary: *In a world of misinformation and disinformation, access to objective evidence-based scientific information is crucial. The general public ignores scientific information due to its perceived complexity, resorting to shallow information on the web or in social media. We analyze the complexity of scientific texts retrieved for a layperson's topic, and find a great variation in text complexity. A proof of concept complexity-aware search engine is able to retrieve both relevant and accessible scientific information for a layperson's information need.*

Keywords: Complexity-Aware Information Retrieval, Text Complexity and Readability, Lay Access to Scientific Text.

1. Introduction

The internet and social media drastically altered both the process of generating information and the way we consume it. The internet gives us far easier access to objective scientific information, which is a natural antidote against the pervasive misinformation and disinformation on the Web. In reality, only a small number of non-specialists refer to scientific sources, opting instead for superficial information disseminated on the internet and social media. One of the primary motives for avoiding the scientific literature is its perceived complexity. Even in developed countries, up to 30% of the population can only comprehend texts written with a basic vocabulary (Štajner et al., 2022).

Traditionally Information Retrieval (IR) systems are evaluated according to their efficiency in retrieving documents topically related to a query but this paradigm ignores the widely varying backgrounds and expertise levels of individual users, who may strictly prefer more accessible information on the topic over highly advanced documents. Specialized scholarly search engines, such as Google Scholar, DBLP, or PubMed, are designed to assist experts in scientific literature review (Gusenbauer and Haddaway, 2020) and thus do not target the accessibility of retrieved documents to laypersons. However, retrieved scientific documents might be too difficult for a user who might not understand these documents. As a result, these documents might be completely useless for a user even if they are relevant to the query.

We assume an information retrieval or retrieval augmented generation setting with a closed collection. Despite promising results of LLMs for multiple NLP tasks, including the application of ChatGPT for biomedical QA (Jahan et al., 2023; Ateia and Kruschwitz, 2023), these models still suffer from problems such as hallucinations (Ji et al., 2023; Ateia and Kruschwitz, 2023; Ermakova et al., 2023a) or non-determinism and its potential cascading effect (Ateia and Kruschwitz, 2023). For example, ChatGPT provides correct or partially correct answers in half of the cases but the provided references only exist in a small fraction of the answers (Zuccon et al., 2023). This model's instability and hallucinations reduce the reliability of the provided answers for a scientific request. Arguably, these generative models even increase the need for grounded scientific evidence to validate generated responses.

In this paper, our main aim is to investigate the viability of complexity-aware retrieval models aiming to retrieve scientific information for non-expert users. Specifically, we aim to answer the following research questions:

- How difficult are scientific abstracts?
- Are current retrieval models sensitive to text complexity?
- How effective are complexity-aware retrieval models?

To answer these research questions, we conducted a series of experiments on ranking scientific abstracts in response to popular science queries. As traditional ad-hoc retrieval benchmarks, such as TREC collections, are not aimed

Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context @ LREC-COLING 2024, pages 16–26

at evaluating the complexity of the retrieved documents, we conducted our experiments on a specialized scientific retrieval corpus for a broad audience. The CLEF SimpleText track (Ermakova et al., 2021, 2022, 2023b) was the first to investigate the barriers that ordinary citizens face when accessing scientific literature head-on, by making available corpora and tasks to address different aspects of the problem. The CLEF SimpleText track studies both the initial ranking of scientific abstracts in response to a popular science query and the use of emerging text simplification (e.g., Wu and Huang, 2022; Laban et al., 2021) approaches to rewrite complex text in order to make them accessible. This paper investigates whether the initial ranking stage can already be made aware of the text complexity of retrieved abstracts, and attempts to rank more accessible literature first.

The rest of this paper is structured in the following way. In Section 2, we discuss related work on ranking scientific text and related work on quantifying text complexity. In Section 3, we analyze the difficulty of scientific abstracts. In Section 4, we discuss traditional lexical and neural ranking models and analyze both their retrieval effectiveness as well as the text complexity of retrieved results. In Section 5, we introduce two complexity-aware ranking approaches and analyze the trade-offs between retrieval effectiveness and complexity of the retrieved results. We end in Section 6 with discussion and conclusions.

2. Related Work

This section discusses related work. First, we discuss prior work on retrieving scientific text with particular emphasis on the data used in the experiments of this paper. Second, we discuss prior work on quantifying text complexity, with particular emphasis on the common readability measures used in our analysis.

2.1. Scientific Text Retrieval

The origins of the field of IR and its Cranfield/TREC evaluation paradigm are based on searching academic literature (Cleverdon, 1962, 1967). The constantly growing number of scientific publications makes the use of automatic tools necessary, including information retrieval or summarization (Guo et al., 2021). Although specialized scientific documents have long been considered by IR systems (Jones and Van Rijsbergen, 1976), they are not sensitive to the complexity of the text. Moreover, academic search systems, including Google Scholar, PubMed, and Web of Science, are traditionally designed for scientific domain experts to assist them in doing systematic reviews, meta-analyses (Gusenbauer and Haddaway, 2020). Knowledge extraction from published scholarly literature for business and research applications is another popular area of research but it also targets specialists in a particular domain rather than laypersons (Thakur and Kumar, 2022).

Given the escalating worries about public misinformation in various countries and the rise of disinformation campaigns orchestrated by organizations, addressing how to effectively educate a wide audience about the progress in technology and science is a major concern (Scheufele and Krause, 2019).

The CLEF SimpleText track shifted the focus to laypersons searching scientific literature (Ermakova et al., 2021, 2022, 2023b). The track covers a wider range of topics on automatic scientific text simplification, from language simplification to terminology extraction and explanation. For the analysis in this paper, we use the data of the CLEF SimpleText Track's Task 1 retrieving scientific abstracts in response to a popular science query:

Corpus The Corpus consists of 4.9 million bibliographic records, including 4.2 million academic abstracts with corresponding detailed information about authors, affiliations, and citations from the Citation Network Dataset (12th version released in 2020)[1] (Tang et al., 2008).

Context There are 40 popular science articles, with 20 from *The Guardian*[2] and 20 from *Tech Xplore*.[3] These journalistic articles were used to construct search requests on popular science topics.

Requests There are 114 queries with 1-4 queries per context article, 47 queries are based on *The Guardian* and 67 on *Tech Xplore*.

Train Data The SimpleText organizers provide relevance judgments for 29 queries (corresponding to 15 Guardian articles, G01–G15), with 23 queries having more than 10 relevant abstracts. The approaches of this paper haven't been trained on this data, but it can serve as an additional evaluation for unsupervised approaches.

Assessments For the evaluation, we used the relevance assessments released for the SimpleText test data for 34 queries associated with the 5 articles from The Guardian (G16–G20, 17 queries) and 5 articles from Tech Xplore (T01–T05, 17 queries).

[1] https://www.aminer.cn/citation
[2] https://www.theguardian.com/science
[3] https://techxplore.com/

Table 1: Text complexity: readability in school grade levels

Grade Level	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
School	*Elementary*						*Jr. High*		*High School*				*Undergrad.*				*Grad.*	*PhD*		
	Primary						*Secondary*						*University*					*PhD*		
	Compulsory												*Higher Edu.*							
Age	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25

Table 2: Flesch-Kincaid Grade Level of CLEF SimpleText data

Data	**Size**	**Length**		**FKGL**	
		Mean	**Median**	**Mean**	**Median**
Corpus (scientific abstracts)	4,894,063	901	913	13.87	13.90
News (popular science)	40	5,504	5,540	12.69	12.80

For details of the exact task setup and results, we refer the reader to the detailed overview of the track in (Ermakova et al., 2023b).

2.2. Text Complexity

This paper performs an initial analysis of the complexity of the scientific abstracts retrieved for a popular science query. The most used way to quantify text complexity is by using readability measures (Zamanian and Heydari, 2012). To quantify the complexity, we use the popular Flesch-Kincaid Grade Level (FKGL) measure based on lexical and grammatical complexity (Flesch, 1948). This is a simple measure based on word length and sentence length, which may not be the most accurate for a single abstract but a reasonable approximation when averaging over larger sets of data. Readability measures have been criticized ever since their invention (e.g., Štajner et al., 2012), but are the most used initial indicators of text complexity in NLP and IR.

The FKGL score is calibrated to correspond to the readability level suitable for a given school level in the U.S. school system, as shown in Table 1. While literacy levels vary in the population, even among adults, one may assume that an average layperson would have finished compulsory education, corresponding to a high school diploma at a grade level of 12.

3. Corpus Analysis

In this section, we will investigate our first research question: *How difficult are scientific abstracts?* Specifically, we apply readability measures to analyze the text complexity of the scientific data used in our experiments.

Table 2 shows an analysis of the text complexity of the corpus and of popular science context. As shown in Table 2, the average (median) length of the abstracts is 901 (913) tokens, and the average (median) complexity of the abstracts is 13.87 (13.9) FKGL.

How complex are scientific abstracts? We can immediately confirm that scientific literature is indeed complex: the scale is the U.S. grade levels in years, with 12 being the exit level of compulsory education (high school diploma), hence the observed complexity of 14-15 is translating to students halfway in undergraduate or college education.

What is the target level of complexity? Recall that the track also provides 40 popular science articles from The Guardian and TechXplore, which are written by professional science journalists for a general audience. As also shown in Table 2, the average (median) length of these articles is 5,504 (5,540) tokens, and the average (median) complexity of the articles is 12.69 (12.8) FKGL, confirming that a FKGL around 12, translating to the readability level of a high school diploma, is appropriate for laypersons.

Is every single abstract too complex for an average citizen? We down-sampled the corpus by taking every 500th article, resulting in an arbitrary sample of 8,513 non-empty abstracts. Figure 1 (top) shows the distribution of FKGL readability levels, which show a striking variation ranging from 5 (elementary school, 10-year-old children) to 25 (graduate school domain expert). Figure 1 (bottom) visualizes this extreme variation, plotted against the length of the abstracts. There is in fact a weak correlation between text complexity and length (r=0.1059, highly significant, regression line with slope 0.0007 in red), but for any length, we find abstracts on any level of readability.

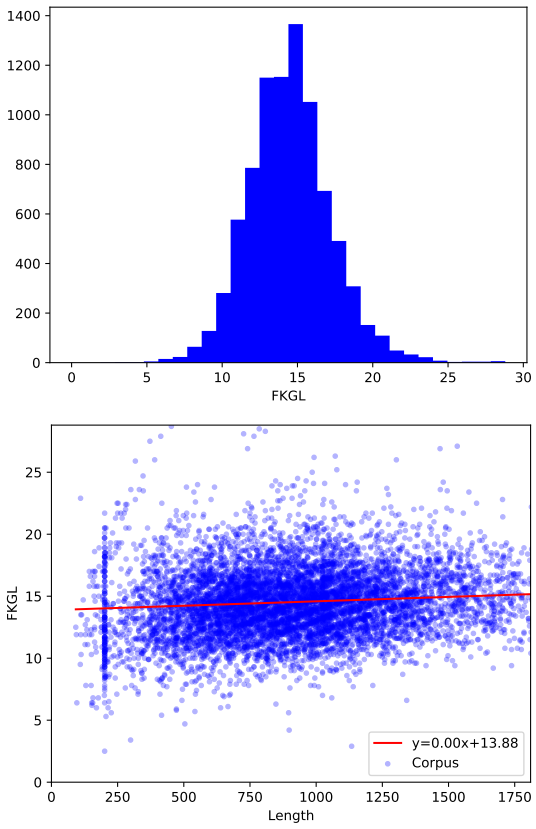

Figure 1: Distribution of text complexity in Flesch-Kincaid Grade Levels (top) and by length (bottom).

Our corpus analysis confirms the common assumption that scientific literature is complex, and a large fraction of abstracts would be very challenging for a layperson. However, our analysis also reveals that a significant fraction of abstracts is within the readability levels of most adult citizens. In the rest of this paper, we will investigate how information retrieval approaches can be made aware of the text complexity and prioritize the retrieval of relevant and accessible abstracts for the request at hand.

4. Effectiveness and Text Complexity

In this section, we will study our second research question: *Are current retrieval models sensitive to text complexity?* Specifically, we will use traditional and neural rankers for scientific text. First, we will evaluate the results in terms of retrieval effectiveness. Second, we will analyze the retrieved results in terms of their text complexity.

4.1. Lexical and Neural Ranking Models

We first conduct a standard IR evaluation of scientific text retrieval, using the corpus of scientific abstracts and popular science requests from the CLEF SimpleText track (Ermakova et al., 2023b).

First, we use a representative traditional ranker BM25 which is based on TF-IDF and normalized document length (Robertson et al., 2009). BM25 is commonly used in traditional search engines, including ElasticSearch,[4] Apache Solr,[5] and Terrier.[6] We used the ElasticSearch implementation of BM25 to retrieve 1,000 results for each keyword query which serves as a first-stage retrieval for the neural re-ranking models. Second, we use a representative neural cross-encoder re-ranker which is a re-implementation of BERT for query-based passage re-ranking (Nogueira and Cho, 2019). This model has shown effective retrieval performance even when applied in zero-shot to new data. Specifically, we apply an MSMARCO-trained model available from Hugging Face.[7] We use this neural cross-encoder re-ranker in a zero-shot way to re-rank either the top 100 or the top 1k retrieved abstracts by the BM25 run.

4.2. Retrieval Effectiveness

We first look at the retrieval effectiveness in the same way as in any other IR evaluation based on topical relevance judgments. Table 3 shows the performance of the three retrieval models on the train and test data, and we make a number of observations. We use standard IR evaluation measures:

- MRR (Mean Reciprocal Rank), which shows a harmonic mean of the ranks;
- Precision@k aiming to compute the share of relevant documents in the top-k retrieved results;
- NDCG (Normalized Discounted Cumulative Gain) considering both the relevance of the items and their position in the list;
- Bpref, preference-based metric that considers whether relevant documents are ranked above irrelevant ones;
- MAP (Mean Average Precision), the mean of the average precision scores for each query.

Comparing the BM25 and the neural re-rankers on the test data, we see that the cross-encoders lead to considerable improvement in retrieval effectiveness, on all evaluation measures. In particular, NDCG@10 increases from 0.3911 up to 0.4782 for

[4] https://www.elastic.co/
[5] https://opensolr.com/
[6] http://terrier.org/
[7] https://huggingface.co/cross-encoder/ms-marco-MiniLM-L-12-v2

Table 3: Retrieval effectiveness on CLEF SimpleText train (top) and test (bottom)

Run	MRR	Precision			NDCG			Bpref	MAP
		5	10	20	5	10	20		
BM25 1k	0.5605	0.4345	0.3655	0.3161	0.3606	0.3627	0.4385	0.4226	0.4072
CE 100	0.5252	0.3241	0.3034	0.2448	0.2701	0.2947	0.3472	0.4012	0.3033
CE 1k	0.4608	0.2759	0.2379	0.1701	0.2312	0.2307	0.2582	0.3335	0.2001
BM25 1k	0.6424	0.4353	0.4059	0.2990	0.4165	0.3911	0.3315	0.2502	0.1895
CE 100	0.7050	0.5118	0.4912	0.3657	0.5004	0.4782	0.4007	0.2616	0.2011
CE 1k	0.6329	0.4765	0.4735	0.3578	0.4502	0.4448	0.3816	0.2797	0.2051

Table 4: Analysis of output (over all 114 queries)

Run	Queries	Top	Year		Length		FKGL	
			Avg	Med	Avg	Med	Avg	Med
BM25 1k	114	10	2012.0	2014	1000.0	995.5	14.0	13.9
CE 100	114	10	2011.7	2013	1102.3	1041.5	14.2	14.1
CE 1k	114	10	2011.8	2014	1142.3	1047.0	14.2	14.1

the CE 100 run, suggesting that the relevant documents have higher ranks, especially in the top positions. The results on the train data are less impressive, but inspection reveals very high fractions of unjudged documents at the top of the neural runs, as no neural IR system contributed to the pools of the train data. Hence, the test data reflects the quality of these runs.

4.3. Text Complexity

We saw that modern IR models perform well in terms of retrieval effectiveness, but how complex are the retrieved abstracts? Table 4 shows an analysis of text complexity of the top 10 results of the lexical and neural models.

We see that the top 10 of the traditional BM25 model retrieves texts of a similar complexity level as the corpus (shown in Table 2 above) with an FKGL of around 14 (with a mean of 14.0, and a median of 13.9). The neural re-rankers also retrieve abstracts with this complexity level, with a slightly higher mean of 14.2 and median of 14.1. To remind, FKGL level 14 corresponds to university-level education, higher than can be taken for granted by a layperson user. Our results indicate that both traditional lexical rankers and modern neural re-rankers focus indeed solely on the topical relevance of abstracts—is the abstract on the topic of the request—and ignore other aspects such as the text complexity.

In this section, we saw that lexical and in particular neural rankers are highly effective in retrieving scientific text. This observation is consistent with the retrieval effectiveness of these models in other domains, and it's reassuring that their effectiveness extends to the domain of scientific text ranking. Their increased effectiveness is already making important potential contributions to the findability of scientific literature, and hence the UNESCO SGDs, at least for expert searchers who have sufficient expertise and language proficiency levels.

5. Complexity-Aware Search

In this section, we explore our third research question: *How effective are complexity-aware retrieval models?* We are interested in making the IR approach aware of the complexity of the text, with the intent to retrieve relevant and accessible texts to our layperson user. We first analyze the distribution of complexity in the retrieved set of abstracts. We then propose straightforward approaches to combine evidence for relevance and readability into the ranking and evaluate these approaches in terms of retrieval effectiveness and in terms of the resulting text complexity. Can we trade-off between these two requirements in ways more suitable for laypersons searching scientific text?

5.1. Analysis of Complexity

What subset of abstracts is selected by a general query based on the popular science newspaper articles? We use the default ElasticSearch engine, retrieve the top 100 scientific articles for each request, and analyze the text complexity of each retrieved abstract. Over the 114 queries, this results in a sample of 11,400 abstracts. As shown also in Table 2, the average (median) length of the retrieved abstracts is 948 (928) tokens, and the aver-

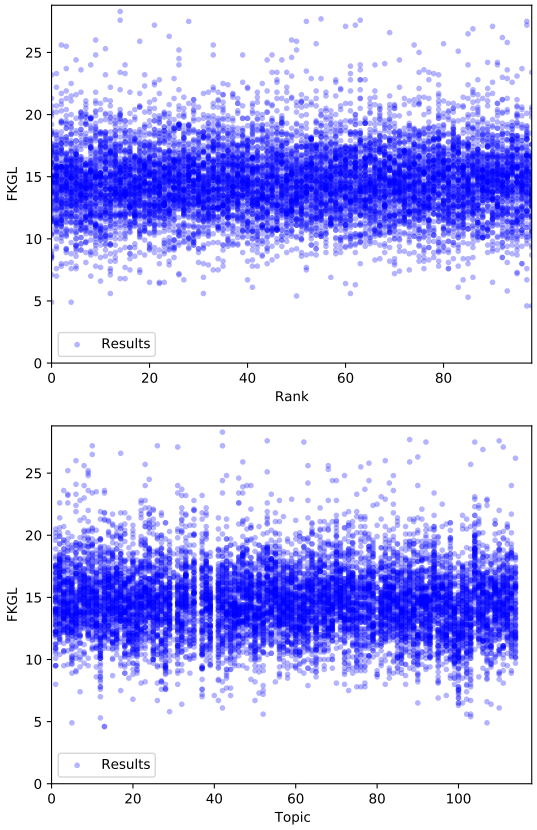

Figure 2: Distribution of text complexity: Top 100 results BM25 over 114 queries by rank (top) and topic (bottom).

age (median) complexity of the abstracts is 13.79 (14.4) FKGL. Hence, the retrieved abstracts are comparable to the corpus statistics, both in terms of length and text complexity, and the distribution of FKGL (not shown) is very similar.

Figure 2 shows the distribution of FKGL readability levels over the rank of retrieval (top) and over each individual query (bottom). In both cases, we see that the standard retrieval engine is completely blind to the text complexity and exclusively focuses on the topical relevance of the abstract. As a result, for any rank and any topic, we see again a striking variation in FKGL, ranging from 10 (starting high school, 15-year-old children) to 20 (doctoral/PhD candidate).

5.2. Complexity-Aware Retrieval Models

Based on the observations above, we explore the viability of complexity-aware retrieval (CAR) models that combine both the relevance and text complexity of a given abstract.

Complexity-Aware Retrieval Filter Our first approach is based on a straightforward global filter, that will only allow the retrieval of abstracts with a favorable readability level. In reality, we use a fudge factor to ensure all selected abstracts receive a higher relevance score than those filtered out.[8] In pseudo-code for FILTER:

if (fkgl <= median_fkgl)
  then combined_score = relevance_score + 10
  else combined_score = relevance_score

We use a global median FKGL of 14 to create interpretable experimental conditions where we prioritize the more accessible half of the corpus and actively demote the less accessible half.

Complexity-Aware Retrieval Combine The neural cross-encoder provides a well-behaved score distribution with a small fraction of documents per topic receiving a positive relevance score. We invert the FKGL level so that lower FKGL levels are more desirable, in a way that the median FKGL level becomes a zero score. In pseudo-code for COMBINE:

if (relevance_score > 0)
  then combined_score = relevance_score
    + (median_fkgl - fkgl)
  else combined_score = relevance_score

Unlike in the rigorous filter, here a high relevance score can still overturn a less desirable FKGL, and a very desirable FKGL can overturn a low relevance score.

We opt for simple and straightforward approaches where we are in full control of the experimental parameters and obtain clear and interpretable outcomes. For the experiments in the rest of this section, we focus on the cross-encoder reranking model.

5.3. Effectiveness and Text Complexity

How will promoting readability fare? Will this be sufficient to retrieve accessible abstracts? And at what cost in performance, as we are trading off against standard retrieval effectiveness?

5.3.1. Text Complexity

Let us first look at whether our complexity-aware retrieval approaches are indeed factoring in the text complexity of the retrieved abstracts. Table 5 shows the text complexity of the top 10 results for all of the 114 queries.

[8] This is following William S. Cooper, ACM SIGIR Salton winner in 1994, who promoted both strict mathematical rigor but also the use of simple experimental stimuli to test controllable and interpretable outcomes. We choose a boost factor of 10 based on the distributional analysis before, which ensures a cohort ranking in which our filter pushes below median FKGL abstracts to the top of the ranking while preserving the internal ranking of each cohort.

Table 5: Analysis of complexity-aware retrieval results (over all 114 queries)

Run	Queries	Top	Year		Length		FKGL	
			Avg	Med	Avg	Med	Avg	Med
CE 1k	114	10	2011.8	2014	1142.3	1047.0	14.2	14.1
CE 1k CAR combine	114	10	2011.6	2014	992.9	909.0	11.2	11.2
CE 1k CAR filter	114	10	2011.5	2014	1056.8	982.0	12.2	12.4

Table 6: Complexity-Aware Retrieval effectiveness on train(top) and test (bottom)

Run	MRR	Precision			NDCG			Bpref	MAP
		5	10	20	5	10	20		
CE 100	0.5252	0.3241	0.3034	0.2448	0.2701	0.2947	0.3472	0.4012	0.3033
CE 100 CAR combine	0.4371	0.3172	0.3069	0.2466	0.2190	0.2489	0.2795	0.3998	0.2838
CE 100 CAR filter	0.5946	0.3517	0.3138	0.2655	0.3008	0.3041	0.3241	0.3906	0.3009
CE 1k	0.4608	0.2759	0.2379	0.1701	0.2312	0.2307	0.2582	0.3335	0.2001
CE 1k CAR combine	0.3182	0.2000	0.1966	0.1655	0.1423	0.1633	0.2240	0.3211	0.1714
CE 1k CAR filter	0.4952	0.2759	0.2414	0.1563	0.2390	0.2431	0.2531	0.3249	0.1934
CE 100	0.7050	0.5118	0.4912	0.3657	0.5004	0.4782	0.4007	0.2616	0.2011
CE 100 CAR combine	0.6779	0.4529	0.3971	0.3456	0.4415	0.4016	0.3642	0.2658	0.1792
CE 100 CAR filter	0.7349	0.5294	0.4353	0.3309	0.5252	0.4511	0.3716	0.2597	0.1790
CE 1k	0.6329	0.4765	0.4735	0.3578	0.4502	0.4448	0.3816	0.2797	0.2051
CE 1k CAR combine	0.5880	0.4412	0.4147	0.3098	0.3854	0.3706	0.3250	0.2700	0.1865
CE 1k CAR filter	0.6403	0.5000	0.4765	0.2941	0.4754	0.4533	0.3334	0.2727	0.1936

We observe that our complexity-aware rankers are indeed returning more accessible scientific abstracts to our lay users. The CAR Filter approach retrieves abstracts of FKGL around 12 (mean 12.2, median 12.4) and the CAR Combine approach FKGL around 11 (mean and median 11.2). To put these text complexity levels in context, an FKGL of 11-12 corresponds to the final years of compulsory education and even lower than the journalistic text used as context for the search requests.

That is, the complexity-aware retrieval approaches are indeed effective in retrieving more accessible scientific abstracts corresponding to the reading level of the targeted lay user.

5.3.2. Retrieval Effectiveness

Let us now look at the performance in terms of retrieval effectiveness. Recall that our baselines are highly effective cross-encoder rankers exhibiting competitive zero-shot performance on many collections and domains. Our CAR approaches try to avoid retrieving complex, but potentially relevant abstracts, so we may observe a trade-off in terms of retrieval effectiveness. Table 6 shows the results. First, we observe that the CAR Combine approach leads to a loss of performance, with NDCG@10 on the train data dropping 16% to 28%. Recall this may still be a reasonable trade-off approach: CAR Combine reduces the FKGL considerably to 11 and strictly focuses on retrieving only accessible content, and still obtains an effectiveness that can exceed the BM25 model. It is reasonable to assume our lay user would prefer to see more accessible abstracts first. Second, the CAR Filter approach fares even better. We would expect some trade-off between retrieval effectiveness and text complexity, and see indeed some small drop at higher recall levels. However, we see a gain in performance on early precision. On the main measure NDCG@10 however, we even observe small gains in retrieval effectiveness up to +5% on the train data and up to +2% on the test data.

In this section, we investigated the viability of complexity-aware rankers aiming to retrieve relevant and accessible abstracts for lay users. First, in line with our analysis of the distribution of text complexity per topic, We observed that we can factor text complexity into the ranking models, and created different types of rankers that promote relevant and accessible text to the front of the ranking. Second, we expected some trade-off in effectiveness between pure-relevance rankers and complexity-aware rankers. However, our experiments demonstrate that the cost can be quite small: it can even lead to minor gains in retrieval effectiveness.

Third, more generally, perhaps most important is the potential positive effect on the user experience of these models by retrieving abstracts fitting the background and education level of our users. This, in turn, holds great promise to increase science literacy and broaden the audience of objective scientific information to the general public.

6. Discussion and Conclusions

The main aim of this paper was to investigate the viability of complexity-aware retrieval models aiming to retrieve scientific information for non-expert users. Scientific literacy is crucial for all citizens, yet traditional IR systems and specialized scholarly search engines seem to cater to expert users.

We conducted an extensive analysis of both relevance and complexity and made a number of observations. Our first research question was: *How difficult are scientific abstracts?* We found that scientific abstracts had high complexity levels on average, confirming the common assumption that scientific literature is complex, but also a remarkable spread of complexity levels. Our second research question was: *Are current retrieval models sensitive to text complexity?* We found that current lexical and neural retrieval models focus exclusively on topical relevance and retrieve scientific abstracts with a complexity similar to the overall corpus. Our third research question was: *How effective are complexity-aware retrieval models?* We found that complexity-aware retrieval models combining relevance and text complexity are effective in reducing the text complexity of retrieved results. One of the more effective strategies is a straightforward filter that demotes those abstracts with undesirable text complexity in the ranking. We expected to have to trade off the retrieval effectiveness with the accessibility of scientific abstract, however, we observed no loss of retrieval effectiveness.

More generally our experiments demonstrate the viability of building complexity-aware rankers sensitive to the background expertise and language proficiency levels of our searchers. This has the potential to greatly improve the user experience of lay users searching scientific literature. Complexity-aware retrieval is a step to make IR more inclusive and sustainable by making scientific knowledge and health-related information more accessible to a wider audience including people with a lower level of education or learning disabilities and thus reducing inequality.

Our conclusions prompt the need for further study of complexity-aware IR. In the future, we plan to investigate in-depth more advanced techniques to evaluate the complexity of texts as well as the accessibility of scientific texts from the perspective of users with different backgrounds.

7. Ethics and Limitations

Complexity-aware ranking is an important step forward to more quality education by making scientific research really open, accessible, and understandable for everyone. Difficult scientific texts are less accessible for non-native speakers (Siddharthan, 2002), young readers, people with reading disabilities (Gala et al., 2020; Chen et al., 2016), needed for reading assistance (e.g. congenitally deaf people) (Inui et al., 2003) or lower level of education. Thus, complexity-aware models could help to reduce inequality and contribute to the inclusiveness and sustainability of natural language processing and information retrieval. Complexity-aware retrieval models can help to make science results accessible for anyone, promoting equal access to education, and health-related information, and ultimately more equal employment opportunities.

The popularization of science is one of UNESCO's oldest programs (UNESCO, b). Education is at the core of UNESCO programs to reach its sustainable development goals (UNESCO, a). This paper investigates how IR can promote science literacy, making significant direct contributions to SGD 4 (quality education), and SGDs 5 and 10 (reduced inequalities), and SDG 3 (increasing well-being), see Figure 3. Moreover, through education it has an indirect impact on all the 17 sustainable development goals (SDGs) of UNESCO.

Figure 3: UNESCO Sustainable Development Goals, with particular contributions to SDG 4, as well as SDG 3, SDG 5, and SGD 10. Based on https://en.unesco.org/sustainabledevelopmentgoals.

The current paper presents a proof of concept of the viability of complexity-aware search. For this reason, we opted for technically simple and interpretable manipulation of very standard classical and modern neural retrieval rankings. This ensures that our results hold for entire classes of systems, but presents no final claims on what would constitute an optimal approach.

Similarly, we equate perceived text complexity with the very crude approximations provided by traditional readability measures. These readability measures have been widely studied and widely used in the literature, ensuring that our results can be directly compared. An additional advantage of these readability measures is that they are clearly interpretable in terms of grammatical and lexical

complexity, strengthening the general conceptual results of the paper.

However, the perceived complexity of scientific text, and the real-world barriers to accessing scientific documents, as well as the key science literacy we may need to provide to lay users, is far more complex. This would need to address missing background knowledge and vernacular, including terminological explanations aiming for the laypersons. For example, explaining a medical condition as *angina pectoris* in precise medical terms may be less helpful than its imprecise relation to heart attacks. Similarly, a technical definition of an advanced term like *differential privacy* may be less helpful than explaining that this is a soft precondition for protecting a lay user's privacy. Such lay explanations seem more general and categorical (this is a type of cancer, privacy protection, ...).

We hope and expect that our general results showing that search engines can be made sensitive to text complexity, will inspire a novel research line in NLP and IR, developing different search technology that can avoid overly complex search results, and appropriate NLP technology that can help laypersons understand the retrieved scientific information. Such future technology should empower lay users, and let them interactively explore scientific information rather than become another single gatekeeper to information. This involves attention to learning aspects, and improving their science literacy, in ways that positive reinforcement of laypersons interest and use of objective science. This can be a natural antidote against shallow information on the web and in social media, often published for their monetary or political value and not their information value or lay user's interests.

8. Acknowledgments

We want to thank in particular the colleagues and the students who participated in data construction, evaluation and reviewing. Liana Ermakova is supported in part by the MaDICS (https://www.madics.fr/ateliers/simpletext/) research group and the French National Research Agency (project *ANR-22-CE23-0019-01*). Jaap Kamps is funded in part by the Netherlands Organization for Scientific Research (NWO CI # CISC.CC.016, NWO NWA # 1518.22.105), and the University of Amsterdam (AI4FinTech program). Views expressed in this paper are not necessarily shared or endorsed by those funding the research.

9. Bibliographical References

Samy Ateia and Udo Kruschwitz. 2023. Is ChatGPT a Biomedical Expert? – Exploring the Zero-Shot Performance of Current GPT Models in Biomedical Tasks. ArXiv:2306.16108 [cs].

Ping Chen, John Rochford, David N. Kennedy, Soussan Djamasbi, Peter Fay, and Will Scott. 2016. Automatic Text Simplification for People with Intellectual Disabilities. In *Artificial Intelligence Science and Technology*, pages 725–731. WORLD SCIENTIFIC.

C. W. Cleverdon. 1962. Report on the testing and analysis of an investigation into the comparative efficiency of indexing systems. Technical report, College of Aeronautics, Cranfield UK.

C. W. Cleverdon. 1967. The Cranfield tests on index language devices. *Aslib*, 19:173–192.

Liana Ermakova, Patrice Bellot, Pavel Braslavski, Jaap Kamps, Josiane Mothe, Diana Nurbakova, Irina Ovchinnikova, and Eric SanJuan. 2021. Overview of simpletext 2021 - CLEF workshop on text simplification for scientific information access. In *CLEF'21: Proceedings of the Twelfth International Conference of the CLEF Association*, volume 12880 of *Lecture Notes in Computer Science*, pages 432–449. Springer.

Liana Ermakova, Sarah Bertin, Helen McCombie, and Jaap Kamps. 2023a. Overview of the CLEF 2023 SimpleText Task 3: Scientific text simplification. In *Working Notes of CLEF 2023: Conference and Labs of the Evaluation Forum*, volume 3497 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Liana Ermakova, Eric SanJuan, Stéphane Huet, Hosein Azarbonyad, Olivier Augereau, and Jaap Kamps. 2023b. Overview of the CLEF 2023 SimpleText Lab: Automatic simplification of scientific texts. In *CLEF'23: Proceedings of the Fourteenth International Conference of the CLEF Association*, volume 14163 of *Lecture Notes in Computer Science*, pages 482–506. Springer.

Liana Ermakova, Eric SanJuan, Jaap Kamps, Stéphane Huet, Irina Ovchinnikova, Diana Nurbakova, Sílvia Araújo, Radia Hannachi, Élise Mathurin, and Patrice Bellot. 2022. Overview of the CLEF 2022 SimpleText Lab: Automatic simplification of scientific texts. In *CLEF'22: Proceedings of the Thirteenth International Conference of the CLEF Association*, volume 13390 of *Lecture Notes in Computer Science*, pages 470–494. Springer.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32(3):p221 – 233.

Núria Gala, Anaïs Tack, Ludivine Javourey-Drevet, Thomas François, and Johannes C Ziegler. 2020. 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)*.

Yue Guo, Wei Qiu, Yizhong Wang, and Trevor Cohen. 2021. Automated lay language summarization of biomedical scientific reviews. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 160–168.

Michael Gusenbauer and Neal R. Haddaway. 2020. Which academic search systems are suitable for systematic reviews or meta-analyses? Evaluating retrieval qualities of Google Scholar, PubMed, and 26 other resources. *Research Synthesis Methods*, 11(2):181–217. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jrsm.1378.

Kentaro Inui, Atsushi Fujita, Tetsuro Takahashi, Ryu Iida, and Tomoya Iwakura. 2003. Text simplification for reading assistance: a project note. In *Proc. of the second international workshop on Paraphrasing - Volume 16*, PARAPHRASE '03, pages 9–16, USA. ACL.

Israt Jahan, Md Tahmid Rahman Laskar, Chun Peng, and Jimmy Huang. 2023. Evaluation of ChatGPT on biomedical tasks: A zero-shot comparison with fine-tuned generative transformers. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 326–336, Toronto, Canada. Association for Computational Linguistics.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

K Sparck Jones and Cornelis Joost Van Rijsbergen. 1976. Information retrieval test collections. *Journal of documentation*.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2021. Keep it simple: Unsupervised simplification of multi-paragraph text. In *ACL/IJCNLP'21: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 6365–6378. Association for Computational Linguistics.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with BERT. *CoRR*, abs/1901.04085.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Dietram A Scheufele and Nicole M Krause. 2019. Science audiences, misinformation, and fake news. *Proceedings of the National Academy of Sciences*, 116(16):7662–7669.

Advaith Siddharthan. 2002. An architecture for a text simplification system.

Sanja Štajner, Richard Evans, Constantin Orasan, and Ruslan Mitkov. 2012. What can readability measures really tell us about text complexity. In *Proceedings of workshop on natural language processing for improving textual accessibility*, pages 14–22. Citeseer.

Sanja Štajner, Daniel Ferrés, Matthew Shardlow, Kai North, Marcos Zampieri, and Horacio Saggion. 2022. Lexical simplification benchmarks for English, Portuguese, and Spanish. *Frontiers in Artificial Intelligence*, 5.

Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. 2008. ArnetMiner: Extraction and mining of academic social networks. In *KDD'08*, pages 990–998.

Khusbu Thakur and Vinit Kumar. 2022. Application of Text Mining Techniques on Scholarly Research Articles: Methods and Tools. *New Review of Academic Librarianship*, 28(3):279–302. Publisher: Routledge _eprint: https://doi.org/10.1080/13614533.2021.1918190.

UNESCO. 1950b. Impact of science on society.

UNESCO. 2017a. *Education for Sustainable Development Goals: learning objectives*. Unesco.

Shih-Hung Wu and Hong-Yi Huang. 2022. CYUT Team2 SimpleText Shared Task Report in CLEF-2022. In *Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022*, CEUR Workshop Proceedings, Bologna, Italy. CEUR-WS.org.

Mostafa Zamanian and Pooneh Heydari. 2012. Readability of texts: State of the art. *Theory & Practice in Language Studies*, 2(1).

Guido Zuccon, Bevan Koopman, and Razia Shaik. 2023. ChatGPT Hallucinates when Attributing Answers. ArXiv:2309.09401 [cs].

10. Language Resource References

Our experiments are based on the corpus (Liana Ermakova and Eric SanJuan and Stéphane Huet and Jaap Kamps, 2023a), the lay search requests (Liana Ermakova and Eric SanJuan and Stéphane Huet and Jaap Kamps, 2023b), and the relevance judgments (Liana Ermakova and Eric SanJuan and Stéphane Huet and Jaap Kamps, 2023c,d).

Liana Ermakova and Eric SanJuan and Stéphane Huet and Jaap Kamps. 2023a. *CLEF 2023 SimpleText Corpus*. CLEF/ANR SimpleText Project. SimpleText Project, CLEF SimpleText Task 1, 1.0. PID https://simpletext-project.com/.

Liana Ermakova and Eric SanJuan and Stéphane Huet and Jaap Kamps. 2023b. *CLEF 2023 SimpleText Popular Science Queries*. CLEF/ANR SimpleText Project. SimpleText Project, CLEF SimpleText Task 1, 1.0. PID https://simpletext-project.com/.

Liana Ermakova and Eric SanJuan and Stéphane Huet and Jaap Kamps. 2023c. *CLEF 2023 SimpleText Relevance Judgments (Test)*. CLEF/ANR SimpleText Project. SimpleText Project, CLEF SimpleText Task 1, 1.0. PID https://simpletext-project.com/.

Liana Ermakova and Eric SanJuan and Stéphane Huet and Jaap Kamps. 2023d. *CLEF 2023 SimpleText Relevance Judgments (Train)*. CLEF/ANR SimpleText Project. SimpleText Project, CLEF SimpleText Task 1, 1.0. PID https://simpletext-project.com/.