University of Amsterdam at the CLEF 2023 SimpleText Track

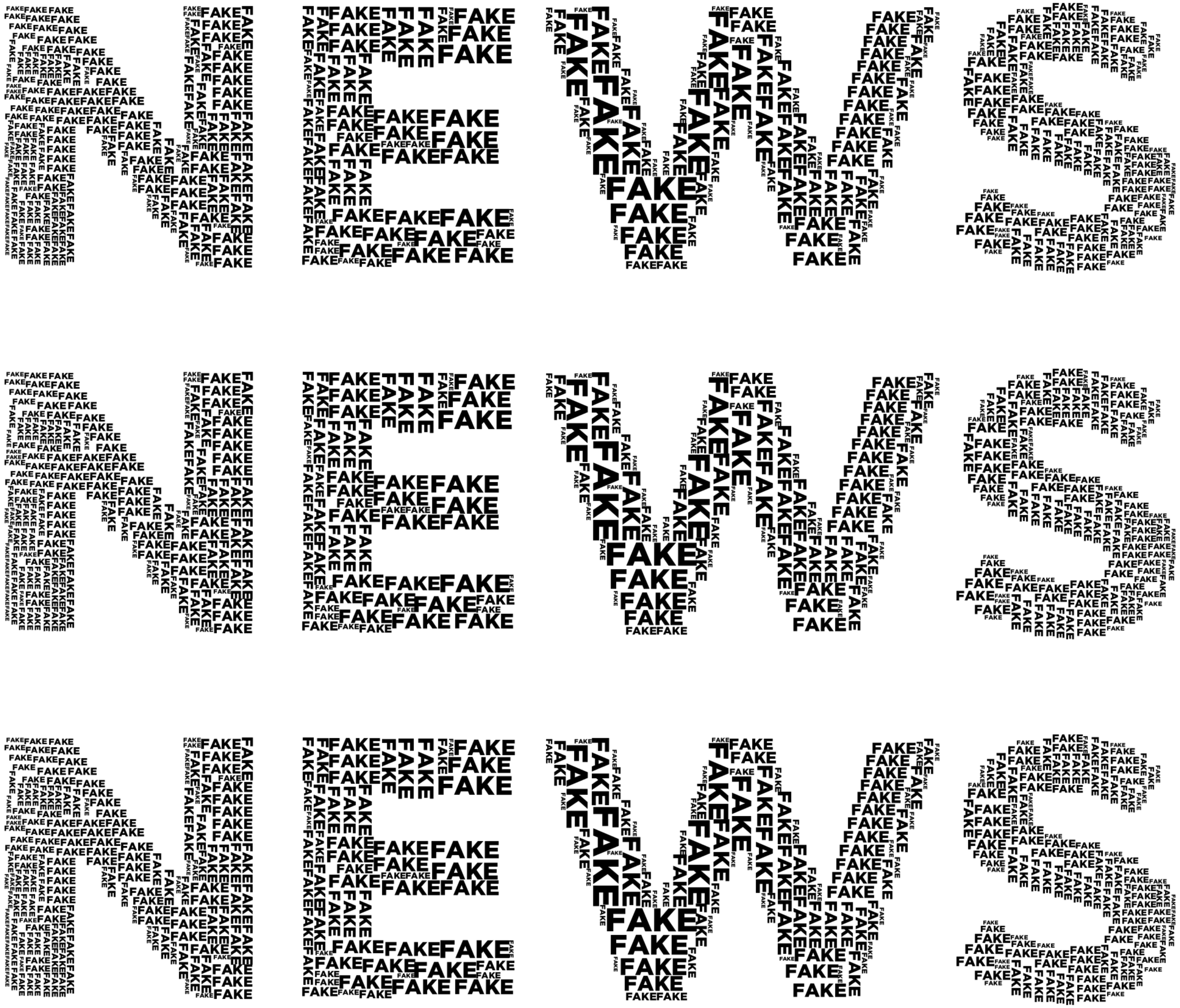
Jaap Kamps
University of Amsterdam

CLEF 2023 SimpleText Track, September 19, 2023, Thessaloniki, Greece

Motivation

Misinfo /Disinfo / Fake News

- Everyone agrees on the importance of **objective** and **reliable** information
- Citizens avoid scientific information as they assume it is **too complex**
- Can we better understand **barriers to access**? even remove them?



What Happens When Laypersons Search Scientific Articles?

- Experiments Complexity-Aware Search and Scientific Text Simplification

Task	Run	Description
1	UAms_Task_1_Elastic	Vanilla elastic run (queries without quotes)
1	UAms_Task_1_CE100	Minilm12 full BERT based crossencoder reranker on top 100
1	UAms_Task_1_CE1k	Minilm12 full BERT based crossencoder reranker on top 1k
1	UAms_Task_1_CE1k_Combine	Neural ranker combining relevance and readability (comb)
1	UAms_Task_1_CE1k_Filter	Neural ranker filtering relevance for readability (comb)
2	UAms_Task_2_RareIDF	IDF baseline using single word terms only
3	UAms_Task_3_Large_KIS150	GPT-2 based text simplification
3	UAms_Task_3_Large_KIS150_Clip	GPT-2 TS with post-processing removing hallucination

How Complex is Science?

#1 Scientific Corpus Analysis

Scientific Text Complexity

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

- Analyze Scientific abstracts, Popular science News articles, and Top 100 results
  - Using standard readability level measures (Flesch-Kincaid Grade Levels)
  - Target level is ~ 12 (high school diploma, exit compulsory education)

Corpus, Context, and Requests

Data	Sample Size	Length		FKGL	
		Mean	Median	Mean	Median
Corpus (scientific abstracts)	8,513	951	905	14.55	14.40
News (popular science)	40	5,504	5,540	12.53	12.70
Retrieved results (top 100)	11,400	948	928	13.79	14.40

- Corpus is too complex, corresponding to university level education
- Popular science news is indeed the target level of 12!
- In response to a general query, the top 100 is as complex as the corpus...

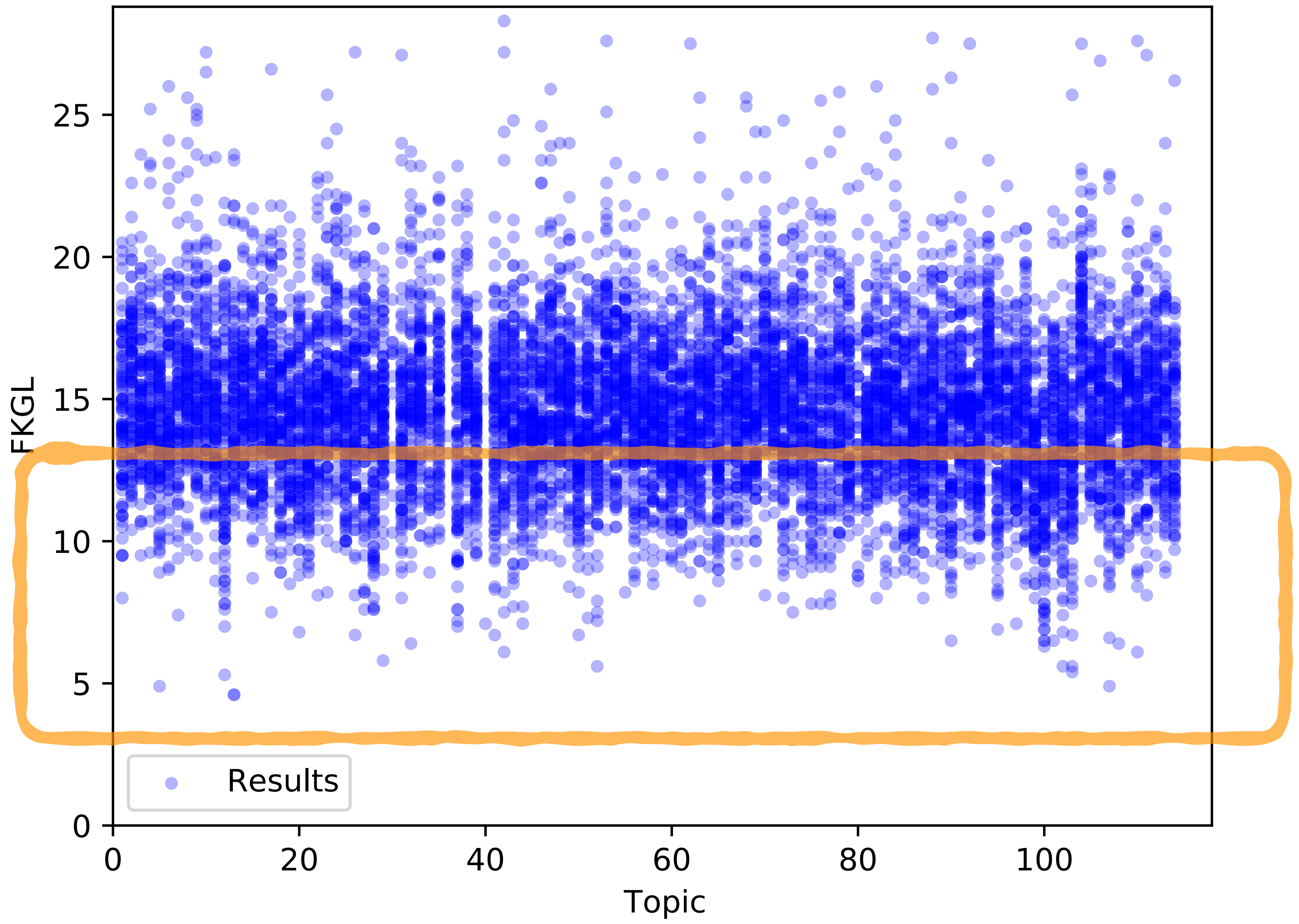
#1 Scientific texts are too Complex

Negative findings explaining why laypersons avoid science...

Can we Avoid Complexity?

#2 Complexity-Aware Retrieval

Complexity Variation per Topic



- For every request there are abstracts with the desirable readability level!

Rel+Read: Complexity-Aware Ranking (1)

Run	MRR	Precision			NDCG			Bpref	MAP
		5	10	20	5	10	20		
Elastic	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
CE 1K Rel+Read combine	0.5880	0.4412	0.4147	0.3098	0.3854	0.3706	0.3250	0.2700	0.1865
CE 1K Rel+Read filter	0.6403	0.5000	0.4765	0.2941	0.4754	0.4533	0.3334	0.2727	0.1936

- As observed in 2022: zero shot neural rankers outcompete lexical
  - NCDG@10 increase from 39% to 48%.
- Our Rel+Read runs very competitive in retrieval effectiveness
  - NDCG@10 even increases from 44% to 45%!

Rel+Read: Complexity-Aware Ranking (2)

Run	Queries	Top	Year		Citations		Length		FKGL	
			Avg	Med	Avg	Med	Avg	Med	Avg	Med
Elastic	114	10	2012.0	2014	13.1	3.0	1000.0	995.5	14.0	13.9
CE 100	114	10	2011.7	2013	25.2	4.0	1102.3	1041.5	14.2	14.1
CE 1k	114	10	2011.8	2014	21.6	3.0	1142.3	1047.0	14.2	14.1
CE 1K Rel+Read combine	114	10	2011.6	2014	16.9	3.0	992.9	909.0	11.2	11.2
CE 1K Rel+Read filter	114	10	2011.5	2014	20.8	3.0	1056.8	982.0	12.2	12.4

- Standard rankers insensitive to text complexity
  - FKGL@10 of ~ 14 similar to the corpus as a whole
- Our Rel+Read runs retrieve more accessible abstracts
  - FKGL@10 drops to the desirable level of 11-12!

#2 Complexity-aware retrieval works

We can avoid abstracts with high text complexity!

Can we Simplify Scientific Text?

#3 Generative AI models for Scientific Text Simplification

Zero-shot Text Simplification

Run	#Snt	FKGL	SARI	BLEU	Comp.	Split	L.Sim.
Train UAms_Task_3_Large_KIS150	648	11.58	36.26	28.60	1.20	1.45	0.81
Train UAms_Task_3_Large_KIS150_Clip	648	12.18	36.61	32.29	0.99	1.23	0.87
Test UAms_Task_3_Large_KIS150	245	10.70	33.41	18.06	1.32	1.51	0.77
Test UAms_Task_3_Large_KIS150_Clip	245	11.98	33.92	21.43	1.01	1.22	0.86

- GPT-2 based "Keep it Simple" (ACL/IJCNLP'21)
  - Used zero-shot, but can be trained *unsupervised* for scientific text
  - Brings FKGL to the desirable level of 11-12
- Evaluation against human simplifications
  - SARI 33%/36% on test/train (cmp. SARI on Wikipedia ~ 26-43%)

#3 Text simplification reduces complexity

We can reduce text complexity of scientific text!

The Truth, the Whole Truth and Nothing but the Truth

#4 Generative AI Models Hallucinate

Generative AI Models for Text Simplification

Topic G07.1, Document 2111507945

The growth of social media provides a convenient ~~communication scheme~~ way for people to communicate , but at the same time it becomes a hotbed of misinformation . ~~The~~ This wide spread of misinformation over social media is injurious to public interest . It is difficult to separate fact from fiction when talking about social media . We design a framework , which ~~integrates~~ combines collective intelligence and machine intelligence , to help identify misinformation . The basic idea is : (1) automatically index the expertise of users according to their microblog ~~contents~~ posts ; and (2) match ~~the~~ experts with the same information given to suspected misinformation . By sending the suspected misinformation to appropriate experts , we can ~~collect~~ gather the ~~assessments of experts~~ relevant data to judge the credibility of the information , and help refute misinformation . In this paper , we ~~focus on~~ look at expert finding for misinformation identification . We ask experts to identify the source of the misinformation , and how it is spread . We propose a tag-based ~~method~~ approach to ~~index~~ indexing the expertise of microblog users ~~with social tags~~ . Our approach will allow us to identify which posts are most relevant and which are not . Experiments on a real world dataset ~~demonstrate~~ show the effectiveness of our ~~method~~ approach for expert finding with respect to misinformation identification in microblogs .

- LLMs used in generative mode:
  - Generate the text simplification as text (prompt) completion
  - But may easily generate additional content!

Quantify and Remove Hallucination

Input	# Input Sentences	# Spurious Content	Fraction Spurious Content
Train	648	126	0.1944
Test Large	152,072	40,449	0.2660

Run	#Snt	FKGL	SARI
Train UAms_Task_3_Large_KIS150	648	11.58	36.26
Train UAms_Task_3_Large_KIS150_Clip	648	12.18	36.61
Test UAms_Task_3_Large_KIS150	245	10.70	33.41
Test UAms_Task_3_Large_KIS150_Clip	245	11.98	33.92

- TS+Clip: Removing hallucination by comparing with input alignment
  - Extremely useful for users: hallucination main problem in LLMs
  - Evaluation measures almost blind — need new TS measures

#4 Need to quantify and remove hallucination

Addressing one of the main challenges in generative AI!

From Sentences to Entire Documents

#5 Generative AI models Hallucinate

Paragraph Level Text Simplification

Run	#Snt	FKGL	SARI	BLEU	Comp.	Split	L.Sim.
Train Sentence level	137	11.60	36.82	29.92	1.11	1.47	0.80
Train Sentence level (clipped)	137	12.23	37.24	33.73	0.97	1.28	0.84
Train Paragraph level	137	12.73	34.55	19.07	0.71	0.90	0.64
Train Paragraph level (clipped)	137	12.76	34.62	19.04	0.67	0.87	0.66
Test Sentence level	38	10.75	34.44	19.36	1.17	1.51	0.79
Test Sentence level (clipped)	38	11.61	34.78	22.78	0.98	1.20	0.85
Test Paragraph level	38	13.05	36.04	8.26	0.51	0.55	0.57
Test Paragraph level (clipped)	38	13.03	36.11	8.29	0.51	0.55	0.57

- Also benchmark data for passage level text simplification
  - Passage level simplification (long input) outperforms sentence level
  - Issues many left out sentences, needs training/finetuning on long input

#5 Models can simplify entire passages directly

SimpleText offers a unique benchmark for passage level text simplification!

What Happens When Laypersons Search Scientific Articles?

#1 Scientific texts are too complex (FKGL 14-15)

#2 Complexity-aware retrieval works (FKGL ~ 12)

#3 Text simplification reduces complexity (FKGL ~12)

#4 Need to quantify and remove hallucination

#5 Models can simplify entire passages directly

Q&A

Thanks to Roos Hutter, Mary Adib, Jop Sutmuller, and David Rau!