University of Amsterdam at the CLEF 2024 SimpleText Track

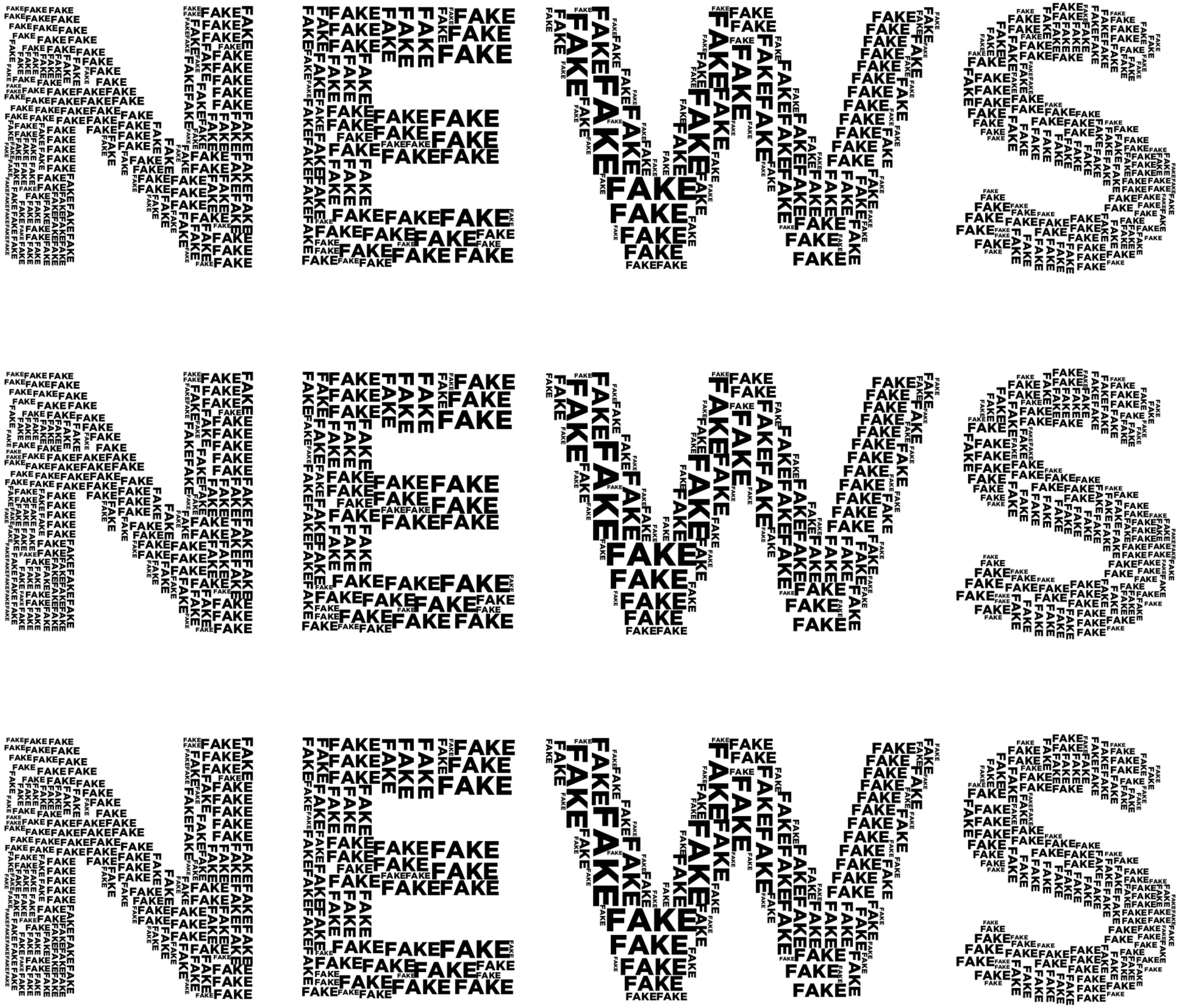
Jaap Kamps, Jan Bakker, Göksenin Yüksel
University of Amsterdam

CLEF 2024 SimpleText Track, September 9, 2024, Grenoble, France

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_Task1_Anserini_bm25	BM25 baseline (Anserini, stemming)
1	UAms_Task1_Anserini_rm3	RM3 baseline (Anserini, stemming)
1	UAms_Task1_CE100	Cross-encoder top 100
1	UAms_Task1_CE1K	Cross-encoder top 1,000
1	UAms_Task1_CE100_CAR	Cross-encoder top 100 + Complexity filter
1	UAms_Task1_CE1K_CAR	Cross-encoder top 1,000 + Complexity filter
2.1	UAms_Task2-1_RareIDF	Up to 5 rarest terms on idf from test-large 2023
2.3	UAms_Task2-3_Anserini_bm25	BM25 baseline (Anserini, stemming)
2.3	UAms_Task2-3_Anserini_rm3	RM3 baseline (Anserini, stemming)
3.1	UAms_Task3-1_GPT2	GPT-2 Sentence level
3.1	UAms_Task3-1_GPT2_Check	GPT-2 Sentence level, Source checked
3.2	UAms_Task3-2_GPT2_Check_Snt	GPT-2 Sentence level, Source checked, merged into abstracts
3.2	UAms_Task3-2_GPT2_Check_Abs	GPT-2 Abstract level, Source checked
3.1	UAms_Task3-1_Wiki_BART_Snt	Wikiauto trained BART sentence level simplification
3.1	UAms_Task3-1_Cochrane_BART_Snt	Cochrane trained BART sentence level simplification
3.2	UAms_Task3-2_Wiki_BART_Par	Wikiauto trained BART paragraph level simplification
3.2	UAms_Task3-2_Cochrane_BART_Par	Cochrane trained BART paragraph level simplification
3.2	UAms_Task3-2_Wiki_BART_Doc	Wikiauto trained BART document level simplification
3.2	UAms_Task3-2_Cochrane_BART_Doc	Cochrane trained BART document level simplification

Search for Scientific Text?

#1 Unsupervised Domain Adaptation

Domain Adaptation: Scientific Text Representations

Run	MRR	Precision			NDCG			Bpref	MAP
		5	10	20	5	10	20		
GPL Base†	0.3752	0.2333	0.2100	0.1611	0.1823	0.1642	0.1465	0.3192	0.0654
GPL Domain Adapt†	0.5169	0.2733	0.2667	0.2233	0.2389	0.2240	0.2075	0.3600	0.0983
GPL Domain Adapt Remining†	0.5011	0.3133	0.3033	0.2467	0.2560	0.2412	0.2285	0.3732	0.1084

† Post-submission experiment.

- Zero shot neural rankers outcompete lexical, but is not tailored to domain
  - Unsupervised domain adaptation creates scientific text representations
- Base (zero shot) can be improved by domain adaptation!
  - NDCG@10 increases from 16% to 22% (GPL), even 24% (new R-GPL)!
  - Training: query generation/fine-tuning, same inference time complexity

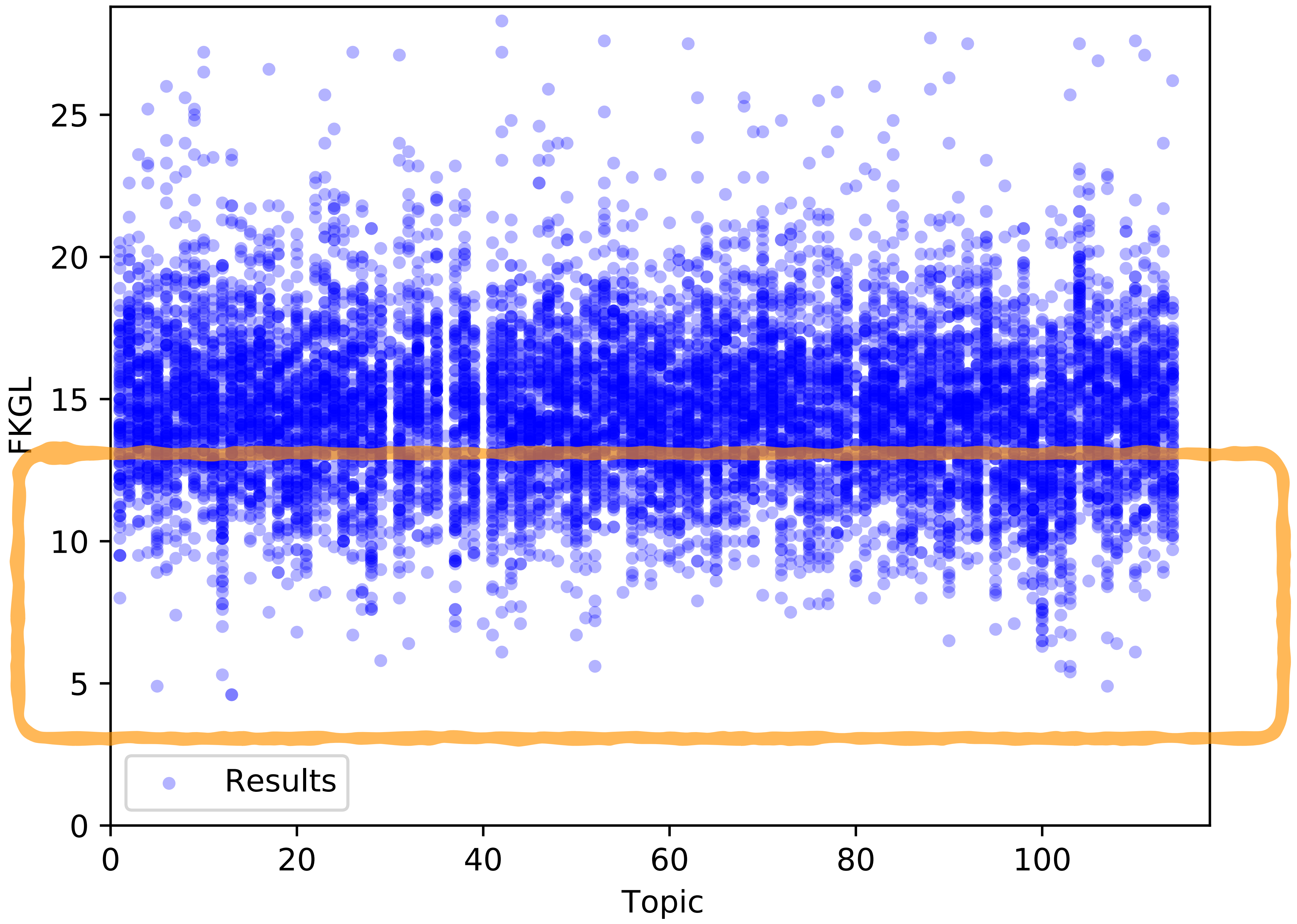
#1 Scientific Text Representations Matter

We can improve models by unsupervised domain adaptation!

Can we Avoid Complexity?

#2 Complexity-Aware Retrieval

Complexity Variation per Topic



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

Complexity-Aware Ranking (1)

Run	MRR	Precision			NDCG			Bpref	MAP
		5	10	20	5	10	20		
UAms_Task1_Anserini	0.7187	0.5600	0.5500	0.4078	0.3867	0.3750	0.3507	0.3994	0.1973
UAms_Task1_Anserini_rm3	0.7878	0.5933	0.5700	0.3611	0.4039	0.3924	0.3282	0.4010	0.1824
UAms_Task1_CE100	0.6618	0.4800	0.5300	0.4044	0.3419	0.3654	0.3452	0.2657	0.1579
UAms_Task1_CE1K	0.5950	0.5133	0.5333	0.4033	0.3571	0.3672	0.3505	0.4031	0.1939
UAms_Task1_CE100_CAR	0.6420	0.5333	0.4700	0.3133	0.3435	0.3199	0.2741	0.2657	0.1321
UAms_Task1_CE1K_CAR	0.6611	0.5467	0.5133	0.2911	0.3800	0.3603	0.2778	0.2676	0.1348

- As observed since 2022: zero shot neural rankers outcompete lexical
  - NCDG@10 increase 39% to 42% on train, but drops 38% to 37% on test.
  - New baseline Anserini performs better than Elastic Search dominating the pool
- Our Complexity-Aware runs very competitive in retrieval effectiveness
  - NDCG@10 only slightly decreases from 36.7% to 36.0%!

Rel+Read: Complexity-Aware Ranking (2)

Run	Queries	Top	Year		Citations		Length		FKGL	
			Avg	Med	Avg	Med	Avg	Med	Avg	Med
UAms_Anserini_bm25	176	10	2012.9	2015	16.5	3.0	1355.9	1249.0	14.5	14.3
UAms_Anserini_rm3	176	10	2013.2	2015	16.8	3.0	1376.6	1272.5	14.5	14.4
UAms_CE100	176	10	2012.6	2015	20.5	3.0	1192.5	1115.0	14.5	14.4
UAms_CE100_CAR	176	10	2012.6	2015	18.0	3.0	1151.4	1081.0	12.5	12.8
UAms_CE1K	176	10	2012.5	2015	19.4	3.0	1147.0	1061.0	14.5	14.4
UAms_CE1K_CAR	176	10	2012.3	2015	18.5	3.0	1083.2	1009.0	12.4	12.7

- Standard rankers insensitive to text complexity
  - FKGL@10 of ~ 14 similar to the corpus as a whole
- Our Complexity-Aware Ranking runs retrieve more accessible abstracts
  - FKGL@10 drops to the desirable level of 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

Scientific Text Simplification (1/3)

run_id	count	FKGL	SARI	BLEU	Compression ratio	Sentence splits	Levenshtein similarity	Exact copies	Additions proportion	Deletions proportion	Lexical complexity score
Source	578	13.65	12.02	19.76	1.00	1.00	1.00	1.00	0.00	0.00	8.80
Reference	578	8.86	100.00	100.00	0.70	1.06	0.60	0.01	0.27	0.54	8.51
UAms_GPT2_Check	578	11.47	29.91	15.10	1.02	1.23	0.87	0.14	0.17	0.14	8.68
UAms_GPT2	578	10.91	29.73	13.07	1.30	1.50	0.79	0.06	0.29	0.12	8.63
UAms_Wiki_BART_Snt	578	12.13	27.45	21.56	0.85	0.99	0.89	0.32	0.02	0.16	8.73
UAms_Cochrane_BART_Snt	578	13.22	18.45	19.21	0.95	0.99	0.96	0.59	0.02	0.07	8.77
Source	103	13.64	12.81	21.36	1.00	1.00	1.00	1.00	0.00	0.00	8.88
Reference	103	8.91	100.00	100.00	0.67	1.04	0.60	0.00	0.23	0.53	8.66
UAms_GPT2_Check_Abs	103	12.85	36.47	13.12	0.91	0.92	0.59	0.00	0.18	0.45	8.73
UAms_Cochrane_BART_Doc	103	14.46	33.51	9.39	0.65	0.58	0.54	0.04	0.06	0.53	8.80
UAms_Cochrane_BART_Par	103	16.53	31.58	15.40	1.08	0.80	0.67	0.04	0.15	0.32	8.81
UAms_GPT2_Check_Snt	103	11.57	30.71	15.24	1.54	1.70	0.78	0.00	0.27	0.13	8.77
UAms_Wiki_BART_Doc	103	15.68	26.50	15.11	1.51	1.14	0.76	0.01	0.25	0.11	8.79
UAms_Wiki_BART_Par	103	13.11	23.92	19.49	1.39	1.37	0.81	0.01	0.11	0.10	8.86

- Lot’s of runs….
  - TL;DR: it “works” FKGL as low as 11% and SARI as high as 36%…

Scientific Text Simplification (2/3)

run_id	count	FKGL	SARI	BLEU	Compression ratio	Sentence splits	Levenshtein similarity	Exact copies	Additions proportion	Deletions proportion	Lexical complexity score
Source	578	13.65	12.02	19.76	1.00	1.00	1.00	1.00	0.00	0.00	8.80
Reference	578	8.86	100.00	100.00	0.70	1.06	0.60	0.01	0.27	0.54	8.51
UAms_GPT2_Check	578	11.47	29.91	15.10	1.02	1.23	0.87	0.14	0.17	0.14	8.68
UAms_GPT2	578	10.91	29.73	13.07	1.30	1.50	0.79	0.06	0.29	0.12	8.63
UAms_Wiki_BART_Snt	578	12.13	27.45	21.56	0.85	0.99	0.89	0.32	0.02	0.16	8.73
UAms_Cochrane_BART_Snt	578	13.22	18.45	19.21	0.95	0.99	0.96	0.59	0.02	0.07	8.77
Source	103	13.64	12.81	21.36	1.00	1.00	1.00	1.00	0.00	0.00	8.88
Reference	103	8.91	100.00	100.00	0.67	1.04	0.60	0.00	0.23	0.53	8.66
UAms_GPT2_Check_Abs	103	12.85	36.47	13.12	0.91	0.92	0.59	0.00	0.18	0.45	8.73
UAms_Cochrane_BART_Doc	103	14.46	33.51	9.39	0.65	0.58	0.54	0.04	0.06	0.53	8.80
UAms_Cochrane_BART_Par	103	16.53	31.58	15.40	1.08	0.80	0.67	0.04	0.15	0.32	8.81
UAms_GPT2_Check_Snt	103	11.57	30.71	15.24	1.54	1.70	0.78	0.00	0.27	0.13	8.77
UAms_Wiki_BART_Doc	103	15.68	26.50	15.11	1.51	1.14	0.76	0.01	0.25	0.11	8.79
UAms_Wiki_BART_Par	103	13.11	23.92	19.49	1.39	1.37	0.81	0.01	0.11	0.10	8.86

- Document level text simplification outcompetes sentence level
  - TL;DR: long input can be risky, but context and discourse structure helps

Scientific Text Simplification (3/3)

run_id	count	FKGL	SARI	BLEU	Compression ratio	Sentence splits	Levenshtein similarity	Exact copies	Additions proportion	Deletions proportion	Lexical complexity score
Source	578	13.65	12.02	19.76	1.00	1.00	1.00	1.00	0.00	0.00	8.80
Reference	578	8.86	100.00	100.00	0.70	1.06	0.60	0.01	0.27	0.54	8.51
UAms_GPT2_Check	578	11.47	29.91	15.10	1.02	1.23	0.87	0.14	0.17	0.14	8.68
UAms_GPT2	578	10.91	29.73	13.07	1.30	1.50	0.79	0.06	0.29	0.12	8.63
UAms_Wiki_BART_Snt	578	12.13	27.45	21.56	0.85	0.99	0.89	0.32	0.02	0.16	8.73
UAms_Cochrane_BART_Snt	578	13.22	18.45	19.21	0.95	0.99	0.96	0.59	0.02	0.07	8.77
Source	103	13.64	12.81	21.36	1.00	1.00	1.00	1.00	0.00	0.00	8.88
Reference	103	8.91	100.00	100.00	0.67	1.04	0.60	0.00	0.23	0.53	8.66
UAms_GPT2_Check_Abs	103	12.85	36.47	13.12	0.91	0.92	0.59	0.00	0.18	0.45	8.73
UAms_Cochrane_BART_Doc	103	14.46	33.51	9.39	0.65	0.58	0.54	0.04	0.06	0.53	8.80
UAms_Cochrane_BART_Par	103	16.53	31.58	15.40	1.08	0.80	0.67	0.04	0.15	0.32	8.81
UAms_GPT2_Check_Snt	103	11.57	30.71	15.24	1.54	1.70	0.78	0.00	0.27	0.13	8.77
UAms_Wiki_BART_Doc	103	15.68	26.50	15.11	1.51	1.14	0.76	0.01	0.25	0.11	8.79
UAms_Wiki_BART_Par	103	13.11	23.92	19.49	1.39	1.37	0.81	0.01	0.11	0.10	8.86

- Scientific text simplification can outcompete generic models
  - Trained on Cochrane plain English summaries (biomedical).

#3 Document level text simplification improves

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

Run	# Input Sentences/Abstracts	Spurious Content	
		Number	Fraction
UAms-1_GPT2	4,797	1,390	0.29
UAms-1_GPT2_Check	4,797	3	0.00
UAms-1_Wiki_BART_Snt	4,797	14	0.00
UAms-1_Cochrane_BART_Snt	4,797	25	0.01
UAms-2_GPT2_Check_Snt	782	111	0.14
UAms-2_GPT2_Check_Abs	782	1	0.00
UAms-2_Wiki_BART_Par	782	46	0.06
UAms-2_Wiki_BART_Doc	782	74	0.09
UAms-2_Cochrane_BART_Par	782	28	0.04
UAms-2_Cochrane_BART_Doc	782	2	0.00

- Hallucination main problem in LLMs: Generative models give more than asked, even for up to 29%!
  - Our “Check” removes hallucination by comparing with input alignment.
  - Standard evaluation measures are “blind” for hallucination: key to quantify and remove.

#4 Need to quantify and remove hallucination

Addressing one of the main challenges in generative AI!

Complex Term Spotting

#5 What term is (not) hard to understand?

Lay Users exhibit Great Variation

Sentence	G06.2_2810968146_2
Source	The model is a ResNet-18 variant, which is fed in images from the front of a simulated F1 car, and outputs optimal labels for steering, throttle, braking.
Reference	[’ResNet-18 variant’, ’braking’, ’braking’, ’f1 car’, ’front’, ’image’, ’model’, ’optimal label’, ’resnet-18’, ’simulated F1 car’, ’steering’, ’steering’, ’throttle’, ’throttle’, ’to be fed’, ’to output’]
Difficulty	[’d’, ’e’, ’e’, ’e’, ’e’, ’e’, ’e’, ’e’, ’d’, ’d’, ’e’, ’e’, ’e’, ’e’, ’e’, ’m’]
Source "d"	The model is a *ResNet-18* variant, which is fed in images from the front of a simulated F1 car, and outputs optimal labels for steering, throttle, braking.
Source "m"	The model is a ResNet-18 variant, which is fed in images from the front of a simulated F1 car, and *outputs* optimal labels for steering, throttle, braking.
Source "e"	The model is a ResNet-18 variant, which is *fed* in *images* from the front of a simulated F1 car, and outputs *optimal labels* for steering, throttle, braking.
Prediction	[’resnet-18’, ’throttle’, ’braking’, ’f1’, ’fed’]

- Lay User see lots of difficult terms (and each different ones)!
  - Simple baseline base on corpus IDF makes reasonable choices

Lay Users also “hallucinate”?

Terms/Sentence	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**29**
Frequency (train)	53	99	90	100	44	55	23	22	16	20	3	5	4	4	1	7	2	2	1
Frequency (test)	18	31	61	65	45	32	26	16	10	3	2	4							

Source	**Number of Terms**	**Occurs in Sentence**	**Not in Sentence**
Train	2,579	2,098	481
Train (case folding)	2,579	2,334	245
Test	1,440	1,312	128
Test (case folding)	1,440	1,347	93

- Up to 29 different terms/concepts, per sentence!
  - And many “spotted terms” don’t literally occur in the sentence!

Evaluation Requires Careful Analysis...

Run	Precision					Recall					F1 Score				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
Train	0.16	0.14	0.13	0.13	0.12	0.04	0.07	0.10	0.13	0.15	0.06	0.09	0.11	0.11	0.12
Test	0.18	0.16	0.14	0.13	0.12	0.05	0.08	0.10	0.12	0.14	0.07	0.10	0.11	0.12	0.12

Run	Rouge				BERTScore		
	1	**2**	**L**	**Lsum**	**P**	**R**	**F1**
Train	0.3729	0.0946	0.3723	0.3733	0.92	0.93	0.92
Test	0.3825	0.0957	0.3810	0.3825	0.93	0.93	0.92

- We return max. 5 single terms per sentence:
  - Exact match P/R/F not high (12%), Top 1 Rouge-1 38%, but BERTScore 92%!

#5 Complex term spotting is complex...

Tricky to evaluate due to user and per sentence variation

What Happens When Laypersons Search Scientific Articles?

#1 Scientific text representations improve

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

#3 Scientific text simplification reduces complexity

#4 Need to quantify and remove hallucination

#5 Complex term spotting is complex...

Q&A

Thanks to Jan Bakker, Göksenin Yüksel, and David Rau!

More details in the paper https://ceur-ws.org/Vol-3740/paper-310.pdf