CLEF 2024 SimpleText Task 3

Simplify Scientific Text

Liana Ermakova Jaap Kamps
Valentin Laimé Helen McCombie

CLEF 2024, September 9, 2024, Grenoble, France

Motivation

- Improving Access to Scientific Texts for Everyone
  - Everyone agrees on the importance of objective scientific information
  - But scientific documents are inherently complex...
- Can we improve accessibility for everyone?
  - Experts
  - Students
  - Lay persons
- Useful for:
  - Scientific communication
  - Science journalism
  - Political communication
  - Education

Generative Text Simplification

CLEF 2024 GRENOBLE
TURN IT SIMPLE

- *Scientific Abstract (FKGL 17.0 – University grad. school)*
  Searching scientific literature and understanding technical scientific documents can be very difficult for users as there are a vast number of scientific publications on almost any topic and the language of science, by its very nature, can be complex. Scientific content providers and publishers should have mechanisms to help users with both searching the content in an effective way and understanding the complex nature of scientific concepts. . . .
- *GPT revisions (FKGL 12.9 – High school diploma)*
  Searching for scientific literature ~~and understanding technical scientific documents~~ can be very ~~difficult~~ time-consuming for users as there are a vast number of scientific publications on almost any topic and the language of science , by its very nature , can be ~~complex~~ very confusing . Scientific content providers and publishers should have mechanisms to help users ~~with both searching~~ find the ~~content~~ right information in an effective way , and understanding the ~~complex~~ nature of scientific concepts . . . .

CLEF 2024 SimpleText Track

- *Task 1: Content Selection: retrieving passages to include in a simplified summary*
  - topical relevance
  - + text complexity scores (e.g., readability)
- *Task 2: Complexity Spotting: identifying and explaining difficult concepts*
  - difficult term detection *and* explanation
- *Task 3: Text Simplification: simplify scientific text*
  - expand the training and automatic evaluation data
  - + both sentence and passage level simplification
  - + analysis of information distortion ("*hallucination?*")
- *Task 4: SOTA?:tracking the state-of-the-art in scholarly publications*
  - Extracting information on system performance from papers
  - Automatically generate leader-boards

Task 3: Text Simplification

- *Task 3: Simplify Scientific Text*
  - This task aims to provide a simplified version of scientific abstracts
- Train data (manually simplified sentences/abstracts)
  - Sentence-level corpus of 648 (2022) and 245 (2023) sentences
  - Paragraph-level corpus of 137 (2022) and 38 (2023) abstracts
- Evaluation
  - Large-scale automatic evaluation measures (SARI, BLEU, ...)
  - Prevalence of spurious content
- Example (human reference simplifications):
  - Source *With the ever increasing number of unmanned aerial vehicles getting involved in activities in the civilian and commercial domain, there is an increased need for autonomy in these systems too.*
  - Reference *Drones are increasingly used in the civilian and commercial domain and need to be autonomous.*

Task 3: New Abstract-level Data

Complex document

Silvano "Nano" Campeggi (] ; January 23, 1923 – August 29, 2018) was an Italian artist who designed and produced the artwork for the posters of many classic Hollywood films. His iconic images are associated with the golden era of Hollywood and Campeggi is now generally regarded as the most important graphic artist and poster designer in the history of American cinema.

In the following decades, Campeggi designed and produced the poster and advertising graphics for over 3000 films, working not only under contract with the MGM studios, but also with Warner Brothers, Paramount, Universal, Columbia Pictures, United Artists, RKO, Twentieth-Century Fox and several other movie studios. Sixty- four of the films he illustrated won Oscars, including "Casablanca", "Ben-Hur", "Singin' in the Rain", "An American in Paris", "West Side Story", "Exodus", "Breakfast at Tiffany's", and "Gigi".

Campeggi died on 29 August 2018, at the age of 95.

Simplification plan-guided sentence-level BART model

Silvano "Nano" Campeggi (]; January 23, 1923 – August 29, 2018) was an Italian artist. He designed and produced the artwork for the posters of many classic Hollywood movies. His iconic images are associated with the golden era of Hollywood and Campeggi is generally regarded as the most important graphic artist and poster designer in the history of American cinema.

Sixty-four of the films he illustrated won Oscars, including "Casablanca", "Ben-Hur", "Singin' in the Rain", "An American in Paris", "West Side Story", "Exodus", "Breakfast at Tiffany's", and "Gigi".

Campeggi died on 29 August 2018 in Milan at the age of 95.

Simplification plan-guided paragraph-level BART model

Silvano "Nano" Campeggi (]; January 23, 1923 – August 29, 2018) was an Italian artist. He designed and produced the artwork for the posters of many classic Hollywood movies. His iconic images are associated with the golden era of Hollywood.

Campeggi illustrated over 3000 movies, including "Casablanca", "Ben-Hur", "Singin' in the Rain", "An American in Paris", "West Side Story", "Exodus", "Breakfast at Tiffany's", and "Gigi".

Campeggi died on 29 August 2018, at the age of 95.

Task 3: Participation

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
3.1	4		4	8			11	1		1	1	1	1	1	1	1	4	2		11	52
3.2	4		4	2			10	1							1	1	6	2			31

- Growing steadily: 14 teams submitted 81 runs.
  - Almost doubled submissions
  - Mostly due to new document-level sub-task

Task 3: Evaluation

Task	Level	Role	Source	Reference
3.1	Sentence	Train	893 sentences	958 simplified sentences
3.1	Sentence	Test	578 sentences	578 simplified sentences
3.1	Sentence	Combined	1,471 sentences	1,536 simplified sentences
3.2	Document	Train	175 abstracts	175 simplified abstracts
3.2	Document	Test	103 abstracts	103 simplified abstracts
3.2	Document	Combined	278 abstracts	278 simplified abstracts

- Created valuable test and train data over 2022–2024
    - 2024 also document-level text simplification
    - Human reference simplifications of abstracts is more labor intensive than sentences...
    - Move to aligned text and simplifications in 2025?

Task 3: Sentence-level Test Results

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
Elsevier_run1	578	10.33	43.63	10.68	0.87	1.06	0.59	0.00	0.45	0.53	8.39
AIIRLab_llama-3-8b_run1	578	8.39	40.58	7.53	0.90	1.37	0.56	0.00	0.48	0.58	8.45
UZHPandas_simple_cot	578	13.74	39.59	3.38	3.44	2.67	0.41	0.00	0.76	0.12	8.61
Sharingans_finetuned	578	11.39	38.61	18.18	0.83	1.07	0.77	0.11	0.16	0.32	8.70
UBO_Phi4mini-s	578	8.74	36.78	0.58	18.23	23.48	0.47	0.00	0.66	0.29	8.89
RubyAiYoungTeam	578	8.76	34.40	15.37	0.60	1.22	0.69	0.03	0.05	0.44	8.71
SONAR_SONARnonlinreg	578	13.14	32.12	18.41	0.97	1.01	0.93	0.13	0.11	0.13	8.73
UAms_GPT2_Check	578	11.47	29.91	15.10	1.02	1.23	0.87	0.14	0.17	0.14	8.68
Arampatzis_T5	578	13.18	28.92	10.66	1.12	1.10	0.72	0.03	0.34	0.37	9.06

- Sentence-level TS high SARI scores throughout up to 44%
  - Larger models seem to perform better (Llama/Mistral vs GPT2/T5)
  - Compression from 60% but up to 1,800% *possible "hallucinations"?*

Task 3: Sentence-level Train Results

run_id	count	FKGL	SARI	BLEU	Compression ratio	Sentence splits	Levenshtein similarity	Exact copies	Additions proportion	Deletions proportion	Lexical complexity score
Source	893	14,30	19,18	38,95	1,00	1,00	1,00	1,00	0,00	0,00	8,72
Reference References	893	11,70	100,00	100,00	0,84	1,07	0,72	0,04	0,21	0,37	8,63
Sharingans_finetuned	714	11,69	64,75	52,53	0,82	1,07	0,73	0,05	0,19	0,37	8,61
Elsevier@SimpleText_run3	714	11,78	46,78	25,55	0,76	0,99	0,68	0,00	0,23	0,47	8,62
Tomislav&Rowan_LLAMA	25	11,84	40,67	4,27	3,94	2,86	0,41	0,00	0,73	0,28	8,36
AIIRLab_Mistral_7B_Instruct_V0.2	893	10,64	39,36	14,07	0,74	1,05	0,58	0,00	0,32	0,58	8,62
UBO_Phi4mini-s	714	8,60	39,27	1,15	17,05	22,28	0,48	0,00	0,65	0,30	8,85
UZH_Pandas_simple_with_cot	714	13,81	38,73	4,62	3,42	2,74	0,41	0,00	0,77	0,12	8,57
PiTheory_T5	97	9,94	36,53	11,02	1,37	1,53	0,63	0,00	0,48	0,30	8,51
team1_Petra_and_Regina_task3_ST	893	8,42	36,19	19,72	0,58	1,29	0,66	0,03	0,05	0,47	8,66
RubyAiYoungTeam	893	8,42	36,19	19,72	0,58	1,29	0,66	0,03	0,05	0,47	8,66
SONAR_SONARnonlinreg	714	13,61	36,01	29,89	0,96	1,02	0,92	0,12	0,10	0,13	8,65
UAms_GPT2_Check	714	11,87	35,21	27,35	1,02	1,22	0,87	0,11	0,17	0,14	8,59
FRANE_AND_ANDREA_t5	893	8,57	34,20	33,58	0,87	1,72	0,82	0,17	0,11	0,24	8,73
Dajana&Kathy_t5	893	8,57	34,20	33,58	0,87	1,72	0,82	0,17	0,11	0,24	8,73
Arampatzis_T5	893	12,15	33,12	21,85	1,09	1,25	0,72	0,03	0,35	0,38	9,07

- Sentence-level train data: broadly similar + signs of overfitting

Task 3: Document-level Test Results

CLEF 2024 GRENOBLE

TURN IT SIMPLE

run_id	count	FKGL	SARI	BLEU	Compression ratio	Sentence splits	Levenshtein similarity	Exact copies	Additions proportion	Deletions proportion	Lexical complexity score
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
AIIRLab_llama-3-8b_run1	103	9.07	43.44	11.73	1.01	1.38	0.51	0.00	0.37	0.56	8.57
Elsevier_run2	103	11.01	42.47	10.54	1.04	1.22	0.51	0.00	0.38	0.55	8.60
Sharingans_finetuned	103	11.53	40.96	18.29	1.20	1.39	0.65	0.00	0.24	0.34	8.80
UBO_Phi4mini-ls	103	8.45	38.79	5.53	1.21	1.75	0.43	0.00	0.40	0.63	8.53
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

- Document-level TS similarly high SARI scores up to 44%
  - Fairly uniform compression of 100% but human reference shorter
  - Some process per sentence (like above), others feed the entire abstract
  - Discourse structure seems to help

Task 3: Document-level Train Resul

run_id	count	FKGL	SARI	BLEU	Compression ratio	Sentence splits	Levenshtein similarity	Exact copies	Additions proportion	Deletions proportion	Lexical complexity score
Source	175	14,30	19,53	39,95	1,00	1,00	1,00	1,00	0,00	0,00	8,88
Reference References	175	11,80	100,00	100,00	0,80	1,04	0,70	0,00	0,20	0,40	8,75
Sharingans_finetuned	119	11,36	60,65	45,74	0,78	1,07	0,68	0,00	0,20	0,41	8,71
Mistral-7B-Instruct-V0.2	175	12,85	40,66	16,52	0,79	0,92	0,60	0,00	0,29	0,51	8,83
AIIRLab_llama-3-8b_run3	119	9,77	40,62	15,04	0,70	1,03	0,55	0,00	0,31	0,57	8,59
Elsevier@SimpleText_run5	119	12,16	40,30	14,23	0,71	0,84	0,55	0,00	0,30	0,57	8,62
UBO_Phi4mini-l	119	9,39	39,95	14,41	1,87	3,23	0,56	0,00	0,18	0,56	8,95
Tomislav&Rowan_LLAMA	20	10,48	37,61	15,26	1,13	1,70	0,53	0,00	0,45	0,48	8,73
UAms_GPT2_Check_Abs	119	12,75	36,68	16,48	0,59	0,66	0,60	0,01	0,11	0,50	8,61
Arampatzis_T5	175	11,39	33,94	9,61	0,48	0,60	0,53	0,00	0,07	0,59	8,90

- Document-level train data
  - Broadly similar findings: trained models overfit?

Task 3: Additional Manual Evaluation

We looked at the document-level simplifications of the entire abstract, and it's coherence and discourse structure.

- Something is working: impressive quality in terms of fluency and language variation
- Changes can be unnecessary nor helpful: Sometimes the vocabulary does not change but is simply rearranged
- Discourse structure matters: the resulting text is incoherent. E.g. start of a first sentence with "*However, . . .*" or with "*For example, . . .*" when the source presented the subject
- Brevity is not always clearer: Some examples shorten the sentences within an abstract, thus technically simplifying, their interrelation is not necessarily maintained, producing a choppy style.
- Gratuitous additions are problematic. For example, adding an off-topic sentence,"*It includes advice on how to manage consultant work, such as research and development,*" with no equivalent in the source text...

Issues in Generative LLMs

- Spurious text completion after each source sentence is simplified

Example of SimpleText Task 3 output versus input: ~~deletions~~, insertions, and whole sentence insertions

As various kinds of output devices emerged , such as highresolution printers or a display of PDA (Personal Digital Assistant) ~~, the~~ . The importance of high-quality resolution conversion has been increasing . | This paper proposes a new method for enlarging an image with high quality . It will involve using a combination of high-speed imaging and high-resolution video . | One of the ~~largest~~ biggest problems on image enlargement is the exaggeration of the jaggy edges . This is especially true when the image is enlarged , as in this case . | To remedy this problem , we propose a new interpolation method ~~, which~~ . This method uses artificial neural network to determine the optimal values of interpolated pixels . | The experimental results are shown and evaluated . The results are compared to other studies and found to be inconclusive . | The effectiveness of our methods is discussed by comparing with the conventional methods . Our methods are designed to help people with mental health problems , not just as a way to cure them . |

Issues in Generative LLMs

- Fraction of sentences with hallucination varies from 0 to 100%
- Existing evaluation measures insensitive to hallucination!

Run	# Input Sentences	Spurious Content	
		Number	Fraction
AB/DVP_SequentialLSTM	4797	4788	1.00
AIIRLab_Mistral_7B_Instruct_V0	779	23	0.03
AIIRLab_llama-3-8b_run3	4797	489	0.10
Dajana/Kathy_t5	779	80	0.10
Elsevier@SimpleText_run1	4797	50	0.01
Elsevier@SimpleText_run4	4795	32	0.01
FRANE_AND_ANDREA_t5	779	80	0.10
SONAR_SONARnonlinreg	4797	15	0.00
Sharingans_finetuned	4797	51	0.01
UAms-1_GPT2	4797	1390	0.29
UAms-1_GPT2_Check	4797	3	0.00
UBO_Phi4mini-s	4797	2055	0.43
UBO_Phi4mini-sl	4797	1822	0.38
RubyAiYoungTeam	4797	1051	0.22
UZHPandas_5Y_target_cot	4797	3383	0.71
UZHPandas_simple_intermediate_defs	4797	79	0.02
Arampatzis_DistilBERT	5576	5575	1.00
Arampatzis_T5	5576	336	0.06
Petra_and_Regina_ST	779	169	0.22

Task 3: Main Findings

- Every participant uses LLMs
- Larger models tend to perform better (in particular on test)
    - Document-level simplification can outperform sentence-level.
    - Very high scores (in particular SARI ~ 0.45)
    - Very good zero-shot performance, even on scientific text
- Output quality looks very good, useful in practice
    - $+$ Lexical/grammatical issues very minor
    - $-$ Text complexity higher than human simplification
    - $-$ Information loss/distortion issues remain
    - $-$ Complex scientific terminology issues remain
    - $-$ Evaluation measures need to factor in hallucination

TURN IT SIMPLE

CLEF 2024 GRENOBLE

cnrs GDR Groupement de recherche
MaDICS Masses de données, informations et connaissances en sciences

anr

Questions?

Fully funded PostDoc available!

Website : https://simpletext-project.com
E-mail : contact@simpletext-project.com
Twitter : https://twitter.com/SimpletextW
Google group : https://groups.google.com/g/simpletext