Overview of the CLEF 2023 SimpleText Track

Automatic Simplification of Scientific Text

Liana Ermakova Éric SanJuan Stéphane Huet
Hosein Azarbonyad Olivier Augereau Jaap Kamps

CLEF 2023, Thessaloniki, Greece, September 18, 2023

SimpleText Track Motivation

- Track focuses on Scientific Text Simplification
- In a world of fake news and misinformation
  - Objective, evidence based science is more important than ever before
  - But hard to find and hard to understand for non-experts
  - Text simplification can remove some of these barriers!
- Also useful for:
  - Promoting science literacy
  - Science journalism and communication
  - Education

Zero-shot Text Simplification

- *Scientific Abstract (FKGL 17.0 ∼ University graduate 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-2 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 . . . .

Goals

- **Use case**: non-expert searching for popular science topics in the scientific literature
- **Overall task**:
  - provide an overview of the scientific literature on this topic
  - create a simplified summary of multiple scientific documents
- We break down this problem in **three IR+NLP subtasks** and provide appropriate reusable **data** and **benchmarks**.
- Track provides a forum to collaborate on **technical & evaluation** challenges of scientific text simplification

CLEF 2023 SimpleText: Three Tasks

Three IR and NLP tasks forming together a pipeline:

- Task 1 (*What is in, or out?*):
  - Selecting passages to include in a simplified summary
  - → Retrieval of scientific articles in response to a popular science query
- Task 2 (*What is unclear?*):
  - Difficult concept identification and explanation
  - → Complex terminology spotting in sentences from scientific abstracts
- Task 3 (*Rewrite this!*):
  - Rewriting scientific text
  - → Rewriting/simplifying sentences from scientific abstracts

SimpleText'23 Submission Stats

Team	Task 1	Task 2.1	Task 2.2	Task 3	Total runs
Elsevier	10				10
Maine (Aiirlab)	10	3	3	2	18
uninib_DoSSIER	2				2
UAms	10	1		2	13
LIA	7				7
MiCroGerk		4	4	3	11
Croland		2	2		
NLPalma		1	1	1	3
Pandas				6	6
QH				3	3
SINAI		4	2		
irgc				4	4
CYUT				4	4
UOL-SRIS		1			1
Smroltra		10	10	1	21
TeamCAU		3	3	1	7
TheLangVerse		1	1	1	3
ThePunDetectives		2	2	2	6
UBO		7	1	1	9
RT				1	1
Total runs	39	39	29	32	139

Task 1: What is in, or out?

- *Task 1: Selecting passages to include in a simplified summary*
- Given a popular science article targeted to a general audience, this task aims to retrieve passages that can help understand this article from a large corpus of academic abstracts and bibliographic metadata
- Citation Network Dataset: DBLP+Citation, ACM Citation network
  - 4,232,520 abstracts in English
- Topics based on 40 press articles + 114 manually extracted queries
  - 20 articles from *The Guardian*
  - 20 articles from *Tech Xplore*

Task 1: Examples

- Text of news articles as context (the **topic**)
  1. Patient data from GP surgeries sold to US companies
  2. Baffled by digital marketing? Find your way out of the maze
- Input: **query** based on these articles
  1. `patient data`
  2. `digital marketing`
  2. `advertising`
- Output:
  - Given the corpus of 4M articles (metadata+abstracts)
  - rank a list of abstracts relevant to the topic/query
  - in JSON format ($\sim$ trec_eval + passage)

Task 1: Evaluation

Qrels	Topics	#Queries	# 0	# 1	# 2
2023 train	G01-G15	29	728	338	237
2023 test	G16-G20, T01-T05	34	2,260	357	1,218

- Train data for system development:
  - 15 popular news topics, with 29 specific queries.
  - Judgments on top 50 abstracts retrieved by ElasticSearch
- Test data:
  - 10 popular news topics, with 34 specific queries
  - Judgments on top 10 abstracts retrieved by all runs
- Evaluation measures:
  - traditional IR metrics (relevance): NDCG, MAP, ...
  - additional complexity/credibility aspect evaluation
  - can we rank relevant and accessible results first?

Task 1: Results on Test Data

Run	MRR	Precision		NDCG		Bpref	MAP
		10	20	10	20		
ElsevierSimpleText_run8	0.8082	0.5618	0.3515	0.5881	0.4422	0.2371	0.1633
ElsevierSimpleText_run7	0.7136	0.5618	0.4103	0.5704	0.4627	0.2626	0.1915
maine_CrossEncoder1^{rel}	0.8106	0.5382	0.4456	0.5675	0.4908	0.3317	0.2810
maine_CrossEncoderFinetuned1^{rel}	0.7691	0.5559	0.4441	0.5542	0.4840	0.3433	0.2572
maine_CrossEncoder1^{comb}	0.7309	0.5265	0.4500	0.5455	0.4841	0.3337	0.2754
ElsevierSimpleText_run5	0.6600	0.4765	0.3838	0.4826	0.4186	0.2542	0.1828
UAms_CE100^{rel}	0.7050	0.4912	0.4044	0.4782	0.4236	0.2616	0.2011
UAms_CE1k_Filter	0.6403	0.4765	0.3559	0.4533	0.3743	0.2727	0.1936
UAms_CE1k^{rel}	0.6329	0.4735	0.4044	0.4448	0.4049	0.2797	0.2051
Elastic baseline	0.6424	0.4059	0.3456	0.3910	0.3541	0.2501	0.1895
unimib_DoSSIER_2	0.5201	0.2853	0.2515	0.2980	0.2683	0.1898	0.1141
unimib_DoSSIER_4	0.5202	0.2853	0.2441	0.2972	0.2632	0.1873	0.1111
run-LIA.bm25	0.4536	0.1912	0.1338	0.2192	0.1700	0.1384	0.0515
run-LIA.all-MiniLM-L6-v2.query	0.3505	0.2000	0.1662	0.2019	0.1767	0.1956	0.0667
run-LIA.all-MiniLM-L6-v2.query-topic	0.3655	0.1765	0.1485	0.1912	0.1647	0.2043	0.0591

- Neural rankers outcompete lexical systems by a large margin
- In particular precision gains, some also recall
- Some submissions prioritized other aspects than relevance

Task 1: Results on Train Data

Run	MRR	Precision		NDCG		Bpref	MAP
		10	20	10	20		
ElsevierSimpleText_run5	0.4156	0.2103	0.1862	0.2097	0.2277	0.3305	0.1742
ElsevierSimpleText_run7	0.3512	0.2241	0.1828	0.1871	0.1986	0.3725	0.1498
ElsevierSimpleText_run8	0.2691	0.1828	0.1500	0.1526	0.1673	0.3585	0.1281
UAms_CE100^{rel}	0.5252	0.3034	0.2690	0.2947	0.3145	0.4012	0.3033
UAms_CE1k^{rel}	0.4608	0.2379	0.1948	0.2307	0.2421	0.3335	0.2001
UAms_CE1k_Filter^{comb}	0.4952	0.2414	0.1879	0.2431	0.2423	0.3249	0.1934
Elastic baseline	0.5605	0.3655	0.3345	0.3627	0.3924	0.4226	0.4072
maine_CrossEncoder1^{rel}	0.7102	0.4448	0.4086	0.4604	0.5017	0.4629	0.5064
maine_CrossEncoder1^{comb}	0.7165	0.4414	0.4155	0.4597	0.5084	0.4619	0.5023
maine_CrossEncoderFinetuned1^{rel}	0.9418	0.7517	0.6086	0.6861	0.7272	0.8730	0.7821
unimib_DoSSIER_2	0.4802	0.2310	0.2086	0.2492	0.2625	0.2568	0.2449
unimib_DoSSIER_4	0.4462	0.2241	0.2069	0.2451	0.2596	0.2514	0.2384
run-LIA.all-MiniLM-L6-v2.query-topic	0.2327	0.1103	0.0983	0.0995	0.1068	0.2279	0.0685
run-LIA.all-MiniLM-L6-v2.query	0.4036	0.1793	0.1397	0.1602	0.1585	0.2025	0.1044
run-LIA.bm25	0.2174	0.1103	0.0948	0.1156	0.1232	0.1626	0.0810

- Many unjudged abstracts in top of train topics based on lexical pool
- Systems with higher recall perform best
- Trained systems overfit on the train topics

Task 1: Text Analysis

Run	# Qry	Top n	Length		FKGL	
			Mean	Median	Mean	Median
UAms_CE1k_Combine	114	10	992.9	909.0	11.2	11.2
UAms_CE1k_Filter	114	10	1056.8	982.0	12.2	12.4
maine_PI2TFIDF	114	10	963.0	964.0	13.5	13.6
ElsevierSimpleText_run5	114	10	994.6	973.5	13.8	13.8
ElsevierSimpleText_run7	114	10	1102.5	1076.5	13.9	13.8
Elastic baseline	114	10	1000.0	995.5	14.0	13.9
maine_CrossEncoder1	114	10	1037.9	999.5	14.0	14.1
ElsevierSimpleText_run8	114	10	1091.3	1046.5	14.1	14.0
run-LIA.bm25	114	10	1112.4	1109.0	14.1	14.2
UAms_CE100	114	10	1102.3	1041.5	14.2	14.1
UAms_CE1k	114	10	1142.3	1047.0	14.2	14.1
uninib_DoSSIER_2	114	10	1102.1	1063.0	14.2	14.2
run-LIA.all-MiniLM-L6-v2.query-topic	114	10	1134.5	1078.5	14.3	14.1
run-LIA.all-MiniLM-L6-v2.query	114	10	1134.5	1078.5	14.3	14.1
uninib_DoSSIER_4	114	10	1099.7	1066.5	14.3	14.2
maine_CrossEncoderFinetuned1	114	10	1076.9	1067.0	14.4	14.4

- The baseline returns FKGL 14 (university level, same as the corpus)
- Some runs return even more complex abstracts
- Some runs return FKGL 11-12 (end of high school, average adult)

Task 1: Complexity Aware Ranking?

Run	NDCG		FKGL	
	10	20	Mean	Median
ElsevierSimpleText_run8	0.5881	0.4422	14.1	14.0
UAms_CE1k_Filter	0.4533	0.3743	12.2	12.4
Elastic baseline	0.3910	0.3541	14.0	13.9
UAms_CE1k_Combine	0.3706	0.3334	11.2	11.2

- Runs targeting relevant and more accessible abstracts
  - Performing competitive on retrieval effectiveness
  - readability level from "university" to "high school"
  - $\rightarrow$ avoiding very complex (but relevant) abstracts
  - *What would our user prefer?*

Task 1: Findings

- Scientific passage retrieval test collection constructed in 2022-2023
  - High pooling diversity
  - Reusable with limited pooling bias
- Almost all submissions based on neural rankers
  - Crossencoders and biencoders popular and very effective
  - Training on scientific text helps
  - Small set of labeled train data can lead to overfitting (use with caution)
- Promising results for runs prioritizing credibility/complexity
  - Possible to factor the text complexity into the ranking
  - Guide users to accessible content first, and more complex text later

Task 2: What is unclear?

- *Task 2: Difficult concept identification and explanation*
- Identify **key concepts** that need to be contextualized with a definition, example, and/or use-case and provide useful and understandable explanations for them.
- Two subtasks:
  1. to retrieve up to 5 difficult terms in a given passage from a scientific abstract;
  2. to provide an explanation of these difficult terms (e.g. definition, abbreviation deciphering, example, etc.).

Task 2.1: Difficult term detection

- Goal:
  - retrieve up to 5 difficult terms in a given passage from a scientific abstract
  - evaluate the difficulty of the retrieved terms
- Data
  - 116,763 sentences from the DBLP scientific abstracts used in Task 1
  - Train data: 203 pairs sentence/term + term definitions
  - Test data: 5,142 distinct pairs sentence/term pooled from the participants' runs (1,262 distinct sentences)
  - Lay annotations of difficult terms
- Metrics
  - correctness of detected term limits: this metric reflects whether the retrieved difficult terms is well limited or not (binary label)
  - difficulty scores: how difficult the term is in the context for an average user and how necessary it is to provide more context about the term:
    - 0 - easy term (no explanation required)
    - 1 - somewhat difficult (explanation could help)
    - 2 - very difficult (explanation is necessary)

Task 2.1: Example

- Input:

  {"query_id":"G14.2",
  "query_text":"end to end encryption",
  "doc_id":"2884788726",
  "snt_id":"G14.2_2884788726_2",
  "source_snt":"However, in information-centric networking (ICN) the end-to-end encryption makes the content caching ineffective since encrypted content stored in a cache is useless for any consumer except those who know the encryption key."}

- Output (1):

  {"snt_id":"G14.2_2884788726_2",
  "term":"content caching",
  "difficulty":1.0,
  "run_id":"team1_task_2.1_TFIDF",
  "manual":0}

Participants approaches

- NLPalma (1 run): BLOOMZ
- UAms (1 run): TF-IDF
- Smroltra (10 runs): KeyBERT, RAKE, YAKE!, BLOOM, SimpleT5, TextRank
- SINAI (4 runs): GPT-3
- TeamCAU (3 runs): SimpleT5, AI21, BLOOM
- MicroGerk (4 runs): YAKE!, TextRank, BLOOM, GPT-3
- ThePunDetectives (2 runs): SimpleT5, GPT-3
- Aiirlab (3 runs): YAKE!, KBIR, IDF
- UBO (7 runs): FirstPhrase, TF-IDF, YAKE!, TextRank, SingleRank, TopicRank, PositionRank
- Croland (2 runs): GPT-3, TF-IDF
- UOL-SRIS (1 run): KeyBERT
- TheLangVerse (1 run): GPT-3

Task 2.1: Results for the official runs

		Test		Train	
	Total	Evaluated	Score	Evaluated	Score
SINAI_task_2.1_PRM_ZS_TASK2_1_V1	11081	1185	507	94	56
UAms_Task_2_RareIDF	675090	1145	241	40	21
SINAI_task_2.1_PRM_FS_TASK2_1_V1	10768	1122	405	81	40
Smroltra_task_2.1_keyBERT_FKgrade	11099	1061	341	41	4
Smroltra_task_2.1_keyBERT_F	11099	1061	171	41	7
UOL-SRIS_2.1_KeyBERT	23757	1061	0	42	0
MiCroGerk_task_2.1_TextRank	21516	1002	391	87	61
Smroltra_task_2.1_TextRank_FKgrade	10056	1002	363	87	29
SINAI_task_2.1_PRM_ZS_TASK2_1_V2	10952	965	330	94	53
SINAI_task_2.1_PRM_FS_TASK2_1_V2	8836	915	316	76	41
Smroltra_task_2.1_YAKE_D	11112	905	422	71	21
MiCroGerk_task_2.1_YAKE	23790	905	362	71	51
Smroltra_task_2.1_YAKE_Fscore	11112	905	209	71	32
MiCroGerk_task_2.1_GPT-3	15892	889	459	79	43
UBO_task_2.1_FirstPhrases	14088	831	161	49	19
UBO_task_2.1_PositionRank	13881	825	181	71	29
UBO_task_2.1_SingleRank	14088	748	151	67	19
UBO_task_2.1_TfIdf	14340	740	187	50	13
UBO_task_2.1_TextRank	14088	722	139	67	16
Smroltra_task_2.1_RAKE_AUI	10660	713	288	48	25
Smroltra_task_2.1_RAKE_F	10660	713	170	48	21
UBO_task_2.1_TopicRank	13912	663	144	61	21
UBO_task_2.1_YAKE	14337	576	116	44	11
MiCroGerk_task_2.1_BLOOM	9600	535	218	64	34
Aiirlab_task_2.2_KBIR	4797	429	135	38	11
TeamCAU_task_2.1_ST5	2234	418	201	90	79
Smroltra_task_2.1_SimpleT5	2234	406	239	82	74

Conclusions from Task 2.1

- Used methods:
  - LLMs
  - unsupervised methods
- Results of difficult term detection by LLMs are comparable to RareIDF, TextRank and YAKE!
- Results of the same methods depend heavily on implementation
- Term difficulty scores are quite different from the lay annotations - further research is needed
- Many partial runs due to token/time constraints of LLMs

Task 2.2: Difficult term explanation

- Goal:
  - Provide a short (one/two sentence) explanation/definition for the term extracted in Task 2.1
  - For the abbreviations, the definition would be the extended abbreviation
- Data
  - Terms extracted in Task 2.1
  - Ground truth
    - 1,000 definitions collected by Elsevier
    - 5,000 mined abbreviations
- Metrics
  - **BLEU**
  - **ROUGE L F-measure** (Longest Common Subsequence).
  - *all-mpnet-base-v2* **semantic match** (sentence transformer based similarity)
  - **Exact match** only for the task of **abbreviation extension**
  - **Partial match** (number of non-identical reference and predicted extensions with a Levenshtein distance < 4 chars)

Task 2.2: Example

- Output (2):

{"snt_id":"G14.2_2884788726_2",
"term":"content caching",
"difficulty":1.0,
"definition":"Content caching is a performance optimization mechanism in which data is delivered from the closest servers for optimal application performance.",
"run_id":"team1_task_2.2_TFIDF_BLOOM",
"manual":0}

Participants' approaches

- NLPalma (1 run): BLOOMZ
- Smroltra (10 runs): BERT, BLOOMZ SimpleT5, Wikipedia
- SINAI (2 runs): GPT-3
- TeamCAU (3 runs): SimpleT5, AI21, BLOOM
- MicroGerk (4 runs): Wikipedia, SimpleT5, BLOOMZ, GPT-3
- ThePunDetectives (2 runs): SimpleT5, GPT-3
- Aiirlab (3 runs): definition detection in top-ranked documents based on a trained classifier
- UBO (1 run): Wikipedia
- Croland (2 runs): GPT-3, Wikipedia
- TheLangVerse (1 run): GPT-3

Term explanation results

Run	Evaluated	BLEU	ROUGE	Semantic match
UBO_task_2.1_FirstPhrases_Wikipedia	393	29.73	0.41	0.80
Croland_task_2_PKE_Wiki	43	33.68	0.46	0.70
MiCroGerk_task_2.2_GPT-3_Wikipedia	932	26.38	0.41	0.75
Smroltra_task_2.2_Text_Wiki	547	17.59	0.33	0.75
Smroltra_task_2.2_RAKE_Wiki	337	16.95	0.32	0.74
Smroltra_task_2.2_YAKE_Wiki	436	16.94	0.32	0.73
TeamCAU_task_2.1_BLOOM	10	10.46	0.27	0.48
MiCroGerk_task_2.2_GPT-3_BLOOMZ	1,108	9.07	0.40	0.83
Smroltra_task_2.2_keyBERT_Wiki	302	8.60	0.23	0.69
MiCroGerk_task_2.2_GPT-3_GPT-3	1,108	7.73	0.38	0.83
NLPalma_task_2.2_BERT_BLOOMZ	537	7.22	0.39	0.76
Smroltra_task_2.2_Bloomz	23	7.15	0.30	0.69
TeamCAU_task_2.1_AI21	22	6.38	0.31	0.78
TheLangVerse_task_2.2_openai-curie-finetuned	444	5.03	0.25	0.74
Croland_task_2_GPT3	69	4.83	0.27	0.77
SINAI_task_2.1_PRM_FS_TASK2_2_V1	649	4.23	0.21	0.78
MiCroGerk_task_2.2_GPT-3_simpleT5	1,108	4.22	0.28	0.77
TeamCAU_task_2.1_ST5	379	3.33	0.20	0.60
Smroltra_task_2.2_SimpleT5	392	3.09	0.22	0.72
SINAI_task_2.1_PRM_ZS_TASK2_2_V1	649	3.08	0.19	0.69
Smroltra_task_2.2_keyBERT_dict	120	2.07	0.14	0.51
Smroltra_task_2.2_YAKE_WN	48	1.88	0.15	0.44
Aiirlab_task_2.2_KBIR	556	1.62	0.15	0.50
Smroltra_task_2.2_keyBERT_WN	328	1.33	0.14	0.45
Aiirlab_task_2.2_YAKEIDF	179	1.13	0.14	0.41
Aiirlab_task_2.2_YAKE	165	1.10	0.15	0.43
Smroltra_task_2.2_RAKE_WN	70	0.00	0.14	0.46

Abbreviation expansion results

Run	Evaluated	BLEU	ROUGE	Semantic	Exact	Partial
MiCroGerk_task_2.2_GPT-3_BLOOMZ	854	13.87	0.68	0.76	326	185
MiCroGerk_task_2.2_GPT-3_GPT-3	855	11.86	0.64	0.73	294	166
MiCroGerk_task_2.2_GPT-3_Wikipedia	855	4.68	0.43	0.60	205	109
MiCroGerk_task_2.2_GPT-3_Wikipedia	618	5.01	0.56	0.64	198	109
NLPalma_task_2.2_BERT_BLOOMZ	345	6.83	0.39	0.52	50	47
Smroltra_task_2.2_SimpleT5	185	0.00	0.12	0.39	8	7
TeamCAU_task_2.1_ST5	141	1.48	0.14	0.40	6	3
TheLangVerse_task_2.2_openai-curie-finetuned	204	1.60	0.14	0.42	1	2
SINAI_task_2.1_PRM_ZS_TASK2_2_V1	228	1.61	0.13	0.55	1	0
TeamCAU_task_2.1_AI21	10	1.87	0.14	0.38	0	0
SINAI_task_2.1_PRM_FS_TASK2_2_V1	228	1.35	0.10	0.53	0	0
UBO_task_2.1_FirstPhrases_Wikipedia	116	5.09	0.19	0.47	0	0
Aiirlab_task_2.2_KBIR	202	1.17	0.07	0.44	0	0
Smroltra_task_2.2_RAKE_Wiki	27	0.54	0.04	0.14	0	0
Smroltra_task_2.2_Bloomz	4	0	0.22	0.61	0	0
Aiirlab_task_2.2_YAKEIDF	19	0	0.10	0.40	0	0
Smroltra_task_2.2_keyBERT_WN	188	0	0.04	0.27	0	0
Smroltra_task_2.2_keyBERT_Wiki	163	0.21	0.02	0.13	0	0
Smroltra_task_2.2_keyBERT_dict	46	0	0.04	0.34	0	0
Smroltra_task_2.2_RAKE_WN	21	0	0.04	0.24	0	0
Smroltra_task_2.2_YAKE_WN	32	0	0.02	0.21	0	0
Smroltra_task_2.2_YAKE_Wiki	31	0	0.03	0.11	0	0
Smroltra_task_2.2_Text_Wiki	50	0	0.02	0.10	0	0
Aiirlab_task_2.2_YAKE	9	0	0.13	0.36	0	0
TeamCAU_task_2.1_BLOOM	3	0	0	0.14	0	0

Conclusions from Task 2.2

- High semantic similarity between participants' definitions and the ground truth
- Wikipedia-based runs have the highest similarity
- LLMs (BLOOMz, GPT-3) have the best performance for abbreviation expansion
- Evaluation problems to solve in 2024:
  - Many partial runs lead to less fair results
  - Evaluation results depend on the performance of term extraction
  - Our evaluation does not take into account explanation usefulness and its difficulty

Task 3: Rewrite this!

- *Task 3: Rewriting scientific text*
- This task aims to provide a simplified version of sentences extracted from scientific abstracts.
- Data
  - Corpus of 2,234 (S), 4,797 (M), and 152,072 (L) sentences from Task 1
  - Train data: 648 manually simplified sentences
  - Test data: 245 manually simplified sentences
- Evaluation
  - Large-scale automatic evaluation measures (SARI, ROUGE, compression, readability)
  - Small-scale detailed human evaluation of other aspects, including information distortion

Task 3: Examples

- Input format: *{"snt_id":"G11.1_2892036907_2",*
  "source_snt":"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.",
  "doc_id":2892036907,
  "query_id":"G11.1",
  "query_text":"drones"}
- Output format: *{"run_id":"BTU_task_3_run1",*
  "manual":1,
  "snt_id":"G11.1_2892036907_2",
  "simplified_snt":"Drones are increasingly used in the civilian and commercial domain and need to be autonomous."}

Zero-shot GPT-2 Text Simplification

- **G01.1, 2463945949** DIANE is a digital assistant system that ~~aims to fasten~~ allows the doctor a faster access ~~to various informations at the~~ patient and hospital ~~such as health care facility , medical records , and also human resource data~~ information . ~~The fasten access~~ This could be ~~achieved by implementing~~ done with face recognition and live streaming ~~as part of the digital assistant system~~ .
- **G01.1, 2797641221** Digital assistants are ~~emerging to become more prevalent~~ becoming popular in our daily lives . ~~In interacting with these assistants ,~~ It will allow users ~~may engage in~~ to do multiple tasks ~~within~~ in a ~~short period of time~~ faster way .
- **G01.2, 1448624402** ~~As extensive experimental research has shown individuals~~ Research showed that people suffer from diverse biases (disproportionate weight in favor of or against an idea) in decision-making . . . .

Task 3: Results on Test data

run_id	count	FKGL	SARI	BLEU	Compression ratio	Sentence splits	Levenshtein similarity	Exact copies	Additions proportion	Deletions proportion	Lexical complexity score
Identity_baseline	245	13.64	15.09	26.22	1.00	1.00	1.00	1.00	0.00	0.00	8.64
Reference	245	12.03	100.00	100.00	0.95	1.10	0.66	0.00	0.33	0.40	8.64
CYUT_run1	245	9.63	47.98	14.81	0.87	1.14	0.56	0.0	0.47	0.55	8.35
AiirLab_davinci	243	11.17	47.10	18.68	0.75	1.00	0.68	0.0	0.20	0.45	8.59
irgc_ChatGPT_2stepTurbo	245	12.31	46.98	16.86	0.94	1.04	0.63	0.0	0.37	0.46	8.46
MiCroGerk_GPT-3	245	10.74	46.90	16.98	0.72	1.01	0.67	0.0	0.19	0.47	8.67
ThePunDetectives_GPT-3	245	7.52	41.56	6.10	0.46	0.97	0.50	0.0	0.16	0.68	8.46
Pandas_submission_ensemble	245	10.51	40.25	17.40	0.77	1.09	0.73	0.0	0.15	0.40	8.52
NLPalma_BLOOMZ	245	9.61	35.66	5.76	0.68	1.00	0.51	0.0	0.35	0.66	8.26
UAms_Large_KIS150_Clip	245	11.12	33.47	16.59	1.01	1.23	0.82	0.0	0.24	0.23	8.48
UBO_SimpleT5	245	12.33	30.89	21.08	0.88	1.05	0.89	0.0	0.10	0.22	8.51
TheLangVerse_openai-curie-finetuned	245	12.21	30.78	18.92	0.86	1.00	0.86	0.0	0.11	0.24	8.49
QH_run3	245	12.74	27.56	20.24	0.90	1.01	0.91	0.0	0.09	0.19	8.50
TeamCAU_ST5	245	12.77	27.19	21.06	0.90	1.00	0.91	0.0	0.10	0.20	8.52
Smroltra_SimpleT5	245	12.88	26.25	21.43	0.90	1.00	0.91	0.0	0.09	0.19	8.54

Task 3: Results on Train data

run_id	count	FKGL	SARI	BLEU	Compression ratio	Sentence splits	Levenshtein similarity	Exact copies	Additions proportion	Deletions proportion	Lexical complexity score
Identity_baseline	648	14.54	20.50	43.24	1.0	1.0	1.0	1.0	0.0	0.0	8.74
Reference	648	11.58	100.00	100.00	0.80	1.05	0.74	0.05	0.16	0.35	8.63
UBO_SimpleT5	469	11.87	89.94	77.98	0.79	1.06	0.75	0.0	0.16	0.36	8.62
TheLangVerse_openai-curie-finetuned	648	11.80	89.31	79.34	0.80	1.03	0.75	0.0	0.17	0.36	8.57
QH_run1	648	12.09	79.56	68.63	0.85	1.09	0.79	0.0	0.17	0.31	8.62
TeamCAU_task_2.2_ST5	648	12.30	64.99	59.61	0.81	1.01	0.83	0.0	0.10	0.28	8.63
Smroltra_GPT	100	12.14	44.04	22.79	0.70	0.99	0.68	0.0	0.14	0.44	8.78
AiirLab_davinci	469	12.40	42.01	24.52	0.74	1.00	0.68	0.0	0.18	0.44	8.73
ThePunDetectives_SimpleT5	648	13.39	41.40	45.18	0.89	0.99	0.91	0.0	0.07	0.17	8.68
MiCroGerk_GPT-3	469	12.23	40.49	21.53	0.69	0.99	0.66	0.0	0.15	0.47	8.79
Pandas_submission_ensemble	648	11.36	40.44	28.47	0.69	1.00	0.71	0.0	0.12	0.42	8.66
Croland_GPT3	50	8.77	39.72	10.40	0.43	1.0	0.50	0.0	0.12	0.70	8.68
irgc_t5_noaron	648	9.87	38.70	29.68	0.75	1.42	0.72	0.0	0.15	0.38	8.70
UAms_Large_KIS150_Clip	648	11.92	36.65	28.68	0.98	1.22	0.84	0.0	0.21	0.21	8.58
CYUT_run4	469	10.29	36.58	9.65	0.77	1.01	0.58	0.0	0.43	0.59	8.31
NLPalma_BLOOMZ	469	10.00	33.63	10.86	0.64	0.99	0.50	0.0	0.37	0.69	8.43

Task 3: Additional Manual Evaluation

We manually evaluate binary errors:

- Incorrect syntax;
- Unresolved anaphora due to simplification;
- Unnecessary repetition/iteration (lexical overlap);
- Spelling, typographic or punctuation errors;
- Information distortion by type;
- Information distortion by severity (1-7).

→ Very few linguistic issues, but information distortion issues remain

Task 3: Manual Evaluation

- Comparison of manually simplified and source sentences

Metric (Avg)	Source snt	Simplified snt
FKGL	15.16	12.12
# Abbreviations	0.24	0.13
# Difficult terms	0.41	0.28

- Difficulty levels and information distortion of simplified sentences

	1	2	3	4	5	6	7
syntax complexity	259	51	9				
lexical complexity	93	119	62	26	19		
information loss severity	34	25	29	15	16	4	9
information loss severity %	8	6	7	4	4	1	2

- Generally low lexical and grammatical complexity (low FKGL)
- Significant information distortion: 15% missing essential details, 10% oversimplification, 6% insertion false/unsupported information, ...

Error Type Statistics

Information distortion type (sample of 22 source sentences, and 319 simplifications)	Instances #	Instances %
Spelling, typographic or punctuational errors	94	30.00
Unresolved anaphora due to simplification	34	11.00
Unnecessary repetition/iteration	16	5.00
Incorrect syntax	10	3.00
Omission of essential details with regard to a query	62	14.59
Oversimplification	41	9.65
Insertion of false or unsupported information	25	5.88
Missinterpretation	20	4.70
Topic shift	10	2.36
Insertion of unnecessary details with regard to a query	7	1.65
Style	5	1.17
Redundancy	4	0.94
Nonsense	2	0.47

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 55%
- Existing evaluation measures insensitive to hallucination!

Run	# Input Sentences	Spurious Content	
		Number	Fraction
AiirLab_run1	245	58	0.24
CYUT_run1	757	12	0.02
MiCroGerk_GPT-3	245	1	0.00
NLPalma_BLOOMZ	245	135	0.55
Pandas_alpaca-lora-both-alpaca-simplifier-tripple_10	245	3	0.01
QH_run3	245	1	0.00
Smroltra_SimpleT5	245	0	0.00
TeamCAU_task_2.2_ST5	245	0	0.00
TheLangVerse_openai-curie-finetuned	245	1	0.00
ThePunDetectives_GPT-3	245	0	0.00
UAms_Large_KIS150	757	213	0.28
UBO_SimpleT5	245	0	0.00
irgc_pegasusTuner007plus_plus	245	57	0.23

Task 3: Main Findings

- Every participant uses LLMs
- Larger models tend to perform better (in particular on test)
  - Very high scores (in particular SARI 0.50)
  - Very good zero shot performance, even on scientific text
- Finetuning/training can easily lead to overfitting on train data
  - Unrealistic high SARI and BLEU scores (0.8-0.9)
- Output quality looks very good, useful in practice
  - + Text complexity similar or lower than human simplification
  - + Lexical/grammatical issues very minor
  - − Information loss/distortion issues remain
  - − Complex scientific terminology issues remain
  - − Evaluation measures need to factor in hallucination

SimpleText Program (Room 4)

- Mon 18 Sep 2023 (TODAY)
  - 16:10 - 17:10 Invited talk by Federica Vezzani (Padua) *Easy peasy term squeeze: A terminological perspective on text simplification*
  - 17:10-17:40 CLEF 2023 SimpleText Task 1-3 Overviews
- Tue 19 Sep 2023
  - 09:30 - 11:00 Participant presentations
- Tue 19 Sep 2023
  - 14:00 - 15:00 Participant presentations
  - 15:00 - 15:30 Round table and SimpleText 2024 discussion
    Any ideas or volunteers are welcome!

CLEF 2024

- Conference and Labs of the Evaluation Forum
- 9-12 September 2024, Grenoble - France
- https://clef2024.imag.fr/
  - Today: register at CLEF (closes April 22)
  - 6 May 2024: Run submission for Tasks 1-4 (new SotA task!)
  - 24 May 2024: Evaluation results
  - 31 May 2024: Submission of CEUR Papers
  - 24 June 2024: Reviews/feedback
  - 8 July 2024: Camera Ready of CEUR Papers
- → Please Join Today!

Thanks!

Fully funded 3-years PhD available!

Website : https://simpletext-project.com[1]
E-mail : contact@simpletext-project.com
Twitter : https://twitter.com/SimpletextW
Google group : https://groups.google.com/g/simpletext

[1]This research was funded, in whole or in part, by the French National Research Agency (ANR) under the project ANR-22-CE23-0019-01.

References

- Liana Ermakova, Eric SanJuan, Stéphane Huet, Hosein Azarbonyad, Olivier Augereau, and Jaap Kamps. "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*. Lecture Notes in Computer Science. Springer, 2023.
- Eric SanJuan, Stéphane Huet, Jaap Kamps, and Liana Ermakova. "Overview of the CLEF 2023 SimpleText Task 1: Passage Selection for a Simplified Summary". In: *Working Notes of CLEF 2023: Conference and Labs of the Evaluation Forum*. CEUR Workshop Proceedings. CEUR-WS.org, 2023.
- Liana Ermakova, Hosein Azarbonyad, Sarah Bertin, and Olivier Augereau. "Overview of the CLEF 2023 SimpleText Task 2: Difficult Concept Identification and Explanation". In: *Working Notes of CLEF 2023: Conference and Labs of the Evaluation Forum*. CEUR Workshop Proceedings. CEUR-WS.org, 2023.
- Liana Ermakova, Sarah Bertin, Helen McCombie, and Jaap Kamps. "Overview of the CLEF 2023 SimpleText Task 3: Scientific Text simplification". In: *Working Notes of CLEF 2023: Conference and Labs of the Evaluation Forum*. CEUR Workshop Proceedings. CEUR-WS.org, 2023.