CLEF 2024 SimpleText Track

Improving Access to Scientific Text for Everyone

Liana Ermakova Éric SanJuan Stéphane Huet
Hosein Azarbonyad Giorgio Di Nunzio Federica Vezzani
Jennifer D'Souza Jaap Kamps

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 Example

- *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

SimpleText 2024 Statistics

CLEF 2024 GRENOBLE

- Growing steadily: 45 registered teams, 20 submitted 207 runs.

Team	Task 1	Task 2			Task 3		Task 4		Total runs
		2.1	2.2	2.3	3.1	3.2	4.1	4.2	
AIIRLab	5	3	3		4	4			19
AMATU							3	9	12
Arampatzis	9	5	5	2	4	4			29
Elsevier	10				8	2			20
L3S							12	12	24
LIA	5								5
PiTheory					11	10			21
Sharigans	1	1	1		1	1			5
SINAI		3	3						6
SONAR					1				1
AB/DPV	1	1	1		1				4
Dajana/Katya		1			1				2
Frane/Andrea		1	1		1				3
Petra/Regina	1	1			1				3
Ruby	1	1			1	1			4
Tomislav/Rowan	2	2			1	1			6
UAmsterdam	6	1		2	4	6			19
UBO	1	1	1		2	2			7
UniPD		3	3						6
UZHPandas					11				11
Total runs	42	24	18	4	52	31	15	21	207

Task 1: Content Selection

- *Task 1: Retrieving Passages to Include in a Simplified Summary*
  - This task aims to retrieve scientific abstracts of relevance to a topic in a popular science news article
- Train data
  - AMINER corpus of 4.2M scientific articles (metadata + abstract)
  - Available training data from 2023 includes 29 (train) and 34 (test) queries with judgments
- Topical keyword queries + new queries
  - Generated with OpenAI GPT 4 and post-edited
  - Longer queries (e.g. "*How AI systems, especially virtual assistants, can perpetuate gender stereotypes?*")
- Evaluation
  - Retrieval effectiveness (e.g., NDCG@10)
  - Additional measures for complexity

Task 1: Evaluation

Qrels	Topics	#Q's	#Assessed			#Avg Ass.
			0	1	2	
2022 test	G1–G20, T2,4,5,10–12,15–16,T18–20	72	192	187	107	6.8
2023 train	G01–G15	29	728	338	237	44.9
2023 test	G16–G20, T01-T05	34	2260	357	1218	112.8
2024 train	G01–G20, T01-T05	63	3,675	768	1,655	95.5
2024 test	G1.C1–G10.C1, T06–T11	30	2,775	1,500	579	128.5
2024 test ext.	G1-G10, T01-T20	96	6,463	2,491	1,036	104.1

- Created valuable test and train data over 2022–2024
  - 2024 test uses only new queries
  - Both long questions (Guardian) and keyword (Tech Explore)

Task 1: Results on Test Data

CLEF 2024 GRENOBLE TURN IT SIMPLE

Run	MRR	Precision 10	Precision 20	NDCG 10	NDCG 20	Bpref	MAP
AIIRLab_Task1_LLaMABiEncoder^{rel}	0.9444	0.8167	0.5517	0.6311	0.5240	0.3559	0.2304
LIA_vir_title	0.8454	0.6933	0.4383	0.5090	0.4010	0.3594	0.1534
UAms_Task1_Anserini_rm3	0.7878	0.5700	0.4350	0.3945	0.3506	0.4010	0.1824
UAms_Task1_Anserini_bm25	0.7187	0.5500	0.4883	0.3774	0.3721	0.3994	0.1972
UAms_Task1_CE1K_CAR^{rel}	0.5950	0.5333	0.4583	0.3726	0.3659	0.2701	0.1605
UAms_Task1_CE100^{comb}	0.6618	0.5300	0.4567	0.3705	0.3579	0.2657	0.1579
Arampatzis_1.GPT2_search^{rel}	0.6986	0.5100	0.2550	0.3522	0.2465	0.0742	0.0577
UBO_Task1_TFIDFT5	0.7132	0.4833	0.3817	0.3506	0.3215	0.2354	0.1274
Elsevier@SimpleText_task_1_run8	0.7123	0.4533	0.3367	0.3152	0.2755	0.1582	0.0906
LIA_elastic	0.6173	0.3733	0.2900	0.2818	0.2442	0.3016	0.1325
AB&DPV_SimpleText_task1_FKGL^{rel}	0.6173	0.3733	0.2900	0.2818	0.2442	0.1966	0.1078
Ruby_Task_1^{rel}	0.5470	0.4233	0.3533	0.2790	0.2688	0.1980	0.1110
Ruby_Task_1^{comb}	0.5910	0.3767	0.3000	0.2641	0.2407	0.1961	0.0980
Tomislav/Rowan&Rowan_SimpleText_T1_1^{rel}	0.5444	0.3733	0.2750	0.2477	0.2201	0.0963	0.0601
Sharingans_Task1_marco-GPT3	0.6667	0.0667	0.0333	0.1167	0.0807	0.0107	0.0107
Petra&Regina_simpleText_task_1	0.0026	0.0000	0.0050	0.0000	0.0035	0.0031	0.0007

- Neural rankers outcompete lexical systems (less than 2023)
- In particular precision gains, some also recall
- Some submissions prioritized other aspects than relevance

Task 1: Text Analysis

CLEF 2024 GRENOBLE TURN IT SIMPLE

Run	Avg #Refs	Avg size of vocabulary	Ratio of long words	Ratio of complex words	FKGL avg	FKGL median
AB/DPV_SimpleText_task1_FKGL^{rel}	9.2	92.9	0.384	0.505	15.3	15.1
AIIRLab_Task1_LLaMABiEncoder^{rel}	8.7	95.8	0.375	0.485	15.3	15.1
Arampatzis_1.GPT2_searchs	10.5	91.9	0.392	0.511	15.7	15.1
Elsevier@SimpleText_task_1_run8	10.3	94.4	0.387	0.504	15.5	15.3
LIA_vir_title	9.8	90.4	0.372	0.483	15.0	14.7
Petra/Reginas_simpleText_task1	5.5	86.1	0.386	0.509	15.4	15.3
Ruby_Task_1^{comb}	9.6	101.2	0.36	0.484	14.0	13.7
Ruby_Task_1^{rel}	9.7	92.9	0.389	0.503	15.9	15.2
Sharingans_Task1_marco-GPT3	9.8	59.8	0.373	0.436	15.5	15.5
Tomislav/Rowan_SimpleText_T1_1^{rel}	9.9	93.2	0.391	0.505	15.9	15.4
Uams_Task1_Anserini_bm25	11.8	111.4	0.385	0.506	16.2	15.3
Uams_Task1_Anserini_rm3	11.9	112.9	0.387	0.508	16.8	16.0
UAms_Task1_CE100_CAR^{comb}	10.6	102.5	0.363	0.485	13.5	13.5
UAms_Task1_CE1K_CAR^{comb}	10.2	98.5	0.363	0.483	13.8	13.5
UBO_Task1_TFIDFT5	10.3	99.2	0.386	0.498	15.4	15.2

- The baseline returns FKGL 15 (university level, same as the corpus)
- Some runs return even more complex abstracts
- Some runs return FKGL 13 (end of high school, average adult)

Task 1: Findings

- Scientific passage retrieval test collection constructed in 2022-2024
  - High pooling diversity
  - Reusable with limited pooling bias
- Top submissions based on neural rankers
  - Crossencoders and bi-encoders popular and quite effective
  - Training on scientific text can help (CLEF Conference paper!)
- Promising results for runs taking into account complexity
  - Possible to factor the text complexity into the ranking
  - Guide users to accessible content first, and more complex text later

Task 2: Complexity Spotting

- *Task 2: Identifying and Explaining Difficult Concepts*
- This task aims
  1. to identify terms in a scientific abstract and their difficulty (easy/medium/difficult)
  2. to generate a definition and an explanation for each difficult term
  3. to retrieve the provided definitions of difficult terms in "correct" order
- Example:
  - "*a Bayesian framework for genotype estimation for mixtures of multiple bacteria, named as Genetic Polymorphisms Assignments (GPA) has reduced the false discovery rate (FDR) and mean absolute error (MAE) in single nucleotide variant (SNV) identification.*"
  - GPA (definition): *the identification and categorization of variations in DNA sequences among individuals or populations.*

Task 2: Evaluation

- Train data
  - 576 train sentences with ground truth complex terms/concepts for a total of 2,579 terms (4.5 per query).
- Test data
  - 317 test sentences with ground truth on complex terms/concepts for a total of 1,440 terms (4.6 per query)
  - Additional 3,815 other sentences (candidate definitions) for Task 2.3
- Evaluation measures
  - Difficult term spotting and difficulty level (recall, precision, F1)
  - Generated definitions based on text overlap with references (BLEU, ..)

Task 2.1 Results (difficult terms)

CLEF 2024 GRENOBLE

runid	Overall			Average		
	Recall	**Prec.**	**F1**	**Recall**	**Prec.**	**F1**
AIIRLab_Task2.2_LLaMA	0.299	0.681	0.415	0.307	0.950	0.465
AIIRLab_Task2.2_LLaMAFT	0.006	1.000	0.012	0.007	1.000	0.014
AIIRLab_Task2.2_Mistral	0.212	0.485	0.295	0.199	0.892	0.326
Dajana&Kathy_SimpleText_Task2.2_LLAMA2_13B_CHAT	0.000	0.000	0.000	0.000	0.989	0.000
FRANE_AND_ANDREA_SimpleText_Task2.2_LLAMA2_13B_CHAT	0.006	0.364	0.012	0.010	0.981	0.020
team1_Petra_and_Regina_Task2_ST	0.000	0.000	0.000	0.000	0.995	0.000
Sharingans_Task2.2_GPT	0.565	0.587	0.576	0.583	0.854	0.693
SINAI_task_2_PRM_ZS_TASK2_V1	0.105	0.538	0.176	0.092	0.935	0.167
SINAI_task_2_PRM_ZS_TASK2_V2	0.149	0.806	0.251	0.134	0.978	0.236
SINAI_task_2_PRM_ZS_TASK2_V3	0.053	0.857	0.101	0.047	0.995	0.090
Tomislav&Rowan_Task2.2_LLAMA2_13B_CHAT_1	0.000	0.000	0.000	0.000	1.000	0.000
Tomislav&Rowan_Task2.2_LLAMA2_13B_CHAT	0.000	0.000	0.000	0.000	1.000	0.000
UAms_Task2-1_RareIDF	0.025	0.091	0.040	0.034	0.780	0.066
UboNLP_Task2.1_phi3-oneshot	0.351	0.387	0.368	0.332	0.737	0.457
unipd_t21t22_chatgpt	0.077	0.612	0.137	0.087	0.979	0.160
unipd_t21t22_chatgpt_mod1	0.226	0.591	0.327	0.234	0.979	0.378
unipd_t21t22_chatgpt_mod2	0.385	0.682	0.492	0.324	0.986	0.488

- Finding the difficult terms selected by human experts is hard
  - LLMs performs very well to spot difficult terms...

Task 2.2 Results

runid	BLEU			
	1	2	3	4
AIIRLab_Task2.2_LLaMA	0.286	0.150	0.047	0.018
AIIRLab_Task2.2_LLaMAFT	0.240	0.117	0.000	0.000
AIIRLab_Task2.2_Mistral	0.259	0.133	0.041	0.014
Sharingans_Task2.2_GPT	0.227	0.106	0.031	0.016
SINAI_task_2_PRM_ZS_TASK2_V1	0.252	0.157	0.082	0.060
SINAI_task_2_PRM_ZS_TASK2_V2	0.276	0.159	0.067	0.049
SINAI_task_2_PRM_ZS_TASK2_V3	0.216	0.112	0.039	0.025
unipd_t21t22_chatgpt	0.309	0.185	0.089	0.049
unipd_t21t22_chatgpt_mod1	0.311	0.181	0.082	0.045
unipd_t21t22_chatgpt_mod2	0.294	0.184	0.091	0.052

- Generate definitions evaluated against human reference definitions
  - ChatGPT definitions match some of those by human experts well...

Task 2 Findings

- Main findings
- Task 2.1 Spotting complex terms
  - Models have high precision, but low recall
  - Largest models (ChatGPT, Llama, Mistral) best
- Task 2.1 Generative definitions of complex terms
  - Generative output difficult to evaluate in reusable ways...
  - ChatGPT definitions match some of those by human experts well...
- Task 2.3 Ranking definitions for complex terms
  - Only two participants submitted runs

Task 3: Text Simplification

CLEF 2024 GRENOBLE

- *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: 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
  - But human simplifications of full abstracts is more labor intensive than sentences...

Task 3: Sentence-level 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	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: Document-level Results

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

Issues in Generative LLMs

CLEF 2024 GRENOBLE

TURN IT SIMPLE

- 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

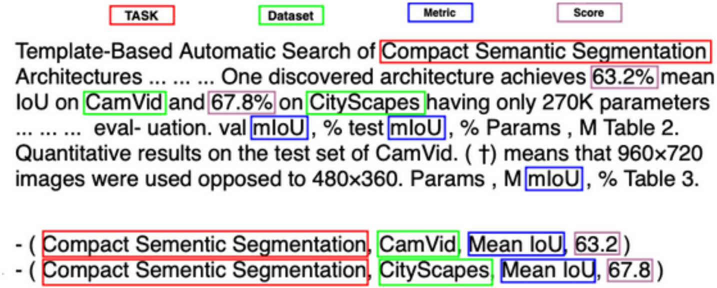
Task 3: Main Findings

- 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

Task 4: SOTA?

- *Task 4: Tracking the State-of-the-Art in Scholarly Publications*
- **Background**: Leaderboards are like scoreboards that display top AI model results for specific tasks, datasets, and metrics. Traditionally community-curated, as seen on paperswithcode.com, text mining could speed up their creation.
- **SOTA Task**: Participants develop systems that recognize if an incoming AI paper reports model performances on benchmark datasets. If it does, the model should extract all related (Task, Dataset, Metric, Score) tuples that are reported in the work.
- **Evaluation**: standard F1 metrics
  - **Few-shot.** Test dataset includes (TDMS)'s seen in training.
  - **Zero-shot.** The test dataset includes (TDMS) with unseen T, D, or M.

Task 4: SOTA Example


TASK Dataset Metric Score

Template-Based Automatic Search of Compact Semantic Segmentation Architectures ... ... One discovered architecture achieves 63.2% mean IoU on CamVid and 67.8% on CityScapes having only 270K parameters ... ... eval- uation. val mIoU, % test mIoU, % Params , M Table 2. Quantitative results on the test set of CamVid. (†) means that 960×720 images were used opposed to 480×360. Params , M mIoU, % Table 3.

- (Compact Sementic Segmentation, CamVid, Mean IoU, 63.2)
- (Compact Sementic Segmentation, CityScapes, Mean IoU, 67.8)

- AI paper with two extracted (Task, Dataset, Metric, Score) tuples

Task 4: SOTA Results

- Filtering papers for leaderboard inclusion

	Few-shot					Zero-shot				
	Rouge				Gen.	Rouge				Gen.
	1	2	L	Lsum	Acc.	1	2	L	Lsum	Acc.
AMATU	**58.34**	12.98	**57.34**	54.4	75.59	**73.72**	6.07	72.72	72.57	85.93
L3S	57.24	**19.67**	56.28	**56.19**	**89.68**	73.54	**12.23**	**73.01**	**72.95**	**95.97**

- Evaluation of individual (Task, Dataset, Metric, Score) tuples

		Few-shot					Zero-shot				
Model	**Mode**	**T**	**D**	**M**	**S**	**Overall**	**T**	**D**	**M**	**S**	**Overall**
AMATU	Exact	27.11	**23.22**	**24.85**	**9.34**	**21.13**	10.01	13.16	11.65	**9.85**	11.16
	Partial	28.08	24.92	25.8	10.86	22.62	16.12	17.12	13.72	11.1	14.52
L3S	Exact	**33.38**	18.51	24.23	1.87	19.50	**26.99**	**14.32**	**22.04**	1.20	**16.14**
	Partial	46.35	32.75	34.16	2.25	28.88	44.90	27.29	32.23	1.41	26.46

Task 4: SOTA Findings

- Bringing information extraction (IE) to CLEF!
- Main findings:
  - Effective prompting paradigms for LLMs out-of-the-box
  - Fine-tuning small-scale models better than larger-scale LLMs for IE task
- Paper context matters:
  - Balancing of length vs specificity of passages containing (T, D, M, S)
  - Context of full paper distracts LLM downstream IE task performance
  - Highly selective passages containing references may harm recall

SimpleText Sessions at CLEF 2024

Date	Event
Sep 9 14:00-15:30	Overview Talks SimpleText Task 1-4
Sep 9 16:00-18:00	*Participant's talks (6x)*
Sep 10 11:10-12:40	Keynote Brian Ondov (Yale/NIH) on TREC PLABA *Participant's talks (3x)*
Sep 10 16:40-18:10	*Participant's talks (3x)* Planning Session: New corpus, new tasks, exciting challenges and opportunities

- Please join the SimpleText sessions in Room 2!

Please join the SimpleText Track

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