Overview of the CLEF 2022 SimpleText Lab: Automatic Simplification of Scientific Texts

Liana Ermakova *et al.*

TURN IT SIMPLE

CLEF 2022 BOLOGNA

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

CLEF - **September 6, 2022**

Motivation

- Scientific documents are difficult to understand
- Accessibility to:
  - Non-native
  - Younger readers
  - Citizens with reading disabilities
- Useful for:
  - Scientific communication
  - Science journalism
  - Political communication
  - Education

Goals

- To create a simplified summary of multiple scientific documents based on a query which provides users with an instant simplified overview on the specific topic they are interested in

---

- **Technical & evaluation** challenges of scientific text simplification
- To provide appropriate reusable **data** and **benchmarks** for text simplification

Organizers

2022 BOLOGNA

TURN IT SIMPLE

- Eric SanJuan, Avignon Université, LIA, France
- Jaap Kamps, University of Amsterdam, The Netherlands
- Stéphane Huet, Avignon Université, LIA, France
- Irina Ovchinnikova, ManPower Language Solution, Israel
- Diana Nurbakova, University of Lyon, INSA Lyon, CNRS, LIRIS, France
- Sílvia Araújo, University of Minho, Portugal
- Radia Hannachi, Université de Bretagne Sud, HCTI, France
- Elise Mathurin, Université de Bretagne Occidentale, HCTI, France
- Patrice Bellot, Aix Marseille Univ, Université de Toulon, CNRS, LIS, France

Shared tasks = pipeline

- *Task 1: What is in (or out)?*
  - Select passages to include in a simplified summary, given a query.
- *Task 2: What is unclear?*
  - Given a passage and a query, rank terms/concepts that are required to be explained for understanding this passage (definitions, context, applications,..).
- *Task 3: Rewrite this!*
  - Given a query, simplify passages from scientific abstracts.

SimpleText run submission statistic

- 62 registered teams
- 40 users downloaded data from the server

Team	Task 1	Task 2	Task 3	Total runs
UAms	2	1		3
NLP@IISERB	3			3
SimpleScientificText		1		1
aaac		1		1
LEA_T5		1	1	2
PortLinguE			1	1
CYUT Team2	1		1	2
HULAT-UC3M			10	10
CLARA-HD			1	1
Total runs	6	4	14	24

Task 1: What is in (or out)?

- To find references in scientific literature that could be inserted as citations in original press articles of general audience for illustration, fact checking or actualization.
- Citation Network Dataset: DBLP+Citation, ACM Citation network
  - 4,232,520 abstracts in English
- Topics = 40 press articles + manually extracted queries (keywords)
  - 20 articles from *The Guardian*
  - 20 articles from *Tech Xplore*

Examples of topics and queries

Topic ID	Query ID	Title or Query
G12		*Patient data from GP surgeries sold to US companies*
	G12.1	`patient data`
G13		*Baffled by digital marketing? Find your way out of the maze*
	G13.1	`digital marketing`
	G13.2	`advertising`

Output formats

- **run_id** Run ID starting with team ID, followed by task1 and run name
- **manual** Whether the run is manual {0,1}
- **topic_id** Topic ID
- **query_id** Query ID used to retrieve the document (if one of the queries provided for the topic was used; 0 otherwise)
- **doc_id** ID of the retrieved document (to be extracted from the JSON output)
- **passage** Text of the selected passage (abstract)

Returned results:

- max 100 distinct DBLP references (_id json field)
- max 1,000 tokens

Evaluation

Passage relevance were evaluated through manual assessment of a pool of passages

- only articles chosen by at least two participants
- relevance score on a scale of 0 to 5
- relevance at the article level
- The abstract was considered as relevant as soon it has a sentence useful to explain the title or the original article

Results

- **#Queries**: the number of queries with at least one result
- **#Docs**: the number of returned documents with a score ≥ 1
- **NDCG@5**: official ranking on this task

Team	#Queries	Avg #Doc.	NDCG		
			5	10	20
CYUT	114	4.9	0.5866	0.5636	0.5536
UAMS	114	95.5	0.3531	0.3776	0.4073
UAMS-MF*	69	2.7	0.3494	0.3328	0.3270
NLP@IISERB 1	30	92.5	0.0605	0.0680	0.0819
NLP@IISERB 2	114	100	0.0503	0.0640	0.0815

* *Manual run.*

Task 2: What is unclear?

- Given a passage and a query, rank terms/concepts that are required to be explained for understanding this passage (definitions, context, applications etc.).
- Passages (sentences) are considered to be independent, i.e. difficult term repetition was allowed.
- max 5 terms per passage
- term difficulty score 1-3 and 1-5

Train dataset

- A master student in Technical Writing and Translation manually annotated each sentence
    - extraction of difficult terms
    - difficulty score on a scale of 1-3 (3 to be the most difficult terms, while the meaning of terms scored 1 can be derived or guessed)
    - difficulty score on a scale of 1-5 (5 to be the most difficult terms)
- 453 annotated examples in total

Test dataset

- 116,763 sentences from the DBLP abstracts according to the queries from Task 1
- manually evaluation of 592 distinct sentences for 11 queries
- 4,167 distinct pairs *sentence-term* in total
- For each evaluated source sentence, the pool contained the results of all participants

Evaluation

- correctness of term limits;
- term difficulty score on the scale 1-3;
- term difficulty score on the scale 1-5;

Results

Table 1: Results for the official runs

	Total	Evaluated		Score_3		Score_5	
			+Limits		+Limits		+Limits
aaac	581,285	2,951	1,388	702	318	415	175
SST	63,027	298	262	48	44	47	42
UAms	263,022	1,315	1,175	105	69	60	49
lea_t5	23,331	5	4	0	0	0	0

Table 2: Results on a subset of 167 common sentences

	Total	Evaluated		Score_3		Score_5	
			+Limits		+Limits		+Limits
aaac	581,285	833	414	200	104	127	67
UAms	263,022	574	514	46	28	25	21
SST	63,027	208	188	33	32	32	29

Task 3: Rewrite this!

- Given a query, simplify passages from scientific abstracts.
- Train dataset
  - parallel corpus of 648 manually simplified sentences
- Test dataset
  - 116,763 sentences retrieved by the ElasticSearch engine from the DBLP dataset, identical to Task 2
  - We manually evaluated 2,276 pairs of sentences for 11 queries

Example (zero-shot simplification)

- *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-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 . . . .

Evaluation

We manually evaluated 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).

Information distortion types

1. Style (distortion severity 1)
2. Insertion of unnecessary details with regard to a query(distortion severity 1)
3. Redundancy (without lexical overlap) (distortion severity 2)
4. Insertion of false or unsupported information (distortion severity 3)
5. Omission of essential details with regard to a query (distortion severity 4)
6. Overgeneralization (distortion severity 5)
7. Oversimplification
8. Topic shift (distortion severity 5)
9. Contra sense / contradiction (distortion severity 6)
10. Ambiguity (distortion severity 6)
11. Nonsense (distortion severity 7)

General results

Run	Total	Unchanged	Truncated	Valid	Longer	Length Ratio	Evaluated	Syntax Err	Anaphora	Minors	Syntax dif	Lexical dif	Information Loss
CLARA-HD	116,763	128	2,292	111,627	201	0.61	851	28	3	68	2.10	2.42	3.84
CYUT Team2	116,763	549	101,104	111,818	49	0.81	126	1		32	2.25	2.30	2.26
PortLinguE_full	116,763	42,189	852	111,589	3,217	0.92	564	7		5	2.94	3.06	1.50
PortLinguE_run1	1,000	359	7	970	30	0.93	80	1			3.63	3.57	2.27
lea_task3_t5	23,360	52	23,201	22,062	24	0.35	.	.	.	.	.	.	.
HULAT-UC3M01	1,000	.	13	973	968	2.46	95	10	1	20	4.69	3.69	2.20
HULAT-UC3M02	2,001	3	58	1,960	1,920	2.53	205	10	1	37	3.60	3.53	2.34
HULAT-UC3M03	1,000	2	13	958	966	2.53	.	.	.	.	.	.	.
HULAT-UC3M04	2,000	.	33	1,827	1,957	37	.	.	.	.	.	.	.
HULAT-UC3M05	2,000	.	56	1,921	1,918	2.38	.	.	.	.	.	.	.
HULAT-UC3M06	2,000	.	47	1,976	1,921	2.45	.	.	.	.	.	.	.
HULAT-UC3M07	1,000	.	56	970	972	2.43	.	.	.	.	.	.	.
HULAT-UC3M08	2,000	.	62	1,964	1,919	2.59	.	.	.	.	.	.	.
HULAT-UC3M09	2,000	.	170	1,964	1,904	2.15	.	.	.	.	.	.	.
HULAT-UC3M10	2,000	.	215	1,963	1,910	2.13	.	.	.	.	.	.	.

Information distortion

Run	Evaluated	Non-Sense	Contresens	Topic Shift	Wrong Synonym	Ambiguity	Essential Detail Loss	Overgeneralization	Oversimplification	Unsupported Information	Unnecessary Details	Redundancy	Style
CLARA-HD	851	162	68	37	20	80	314	59	203	26	10	29	13
CYUT Team2	126	2	1	.	.	4	42	4	5	.	.	.	4
PortLinguE_full	564	9	3	4	3	19	94	9	13	2	2	5	1
PortLinguE_run1	80	.	.	1	.	.	27	5	2	.	.	.	.
lea_task3_t5	.	.	.	.	.	.	.	.	.	.	.	.	.
HULAT-UC3M01	95	1	7	2	.	5	2	.	1	5	38	36	.
HULAT-UC3M02	205	4	9	4	.	9	4	.	.	12	72	61	1

Ranking

Run	Score
PortLinguE_full	0.149
CYUT Team2	0.122
CLARA-HD	0.119

Average harmonic mean of normalized opposite values of *Lexical Complexity (LC)*, *Syntactic Complexity (SC)* and *Distortion Level (DL)*:

$$s_i = \frac{3}{\frac{7}{7-LC} + \frac{7}{7-SC} + \frac{7}{7-DL}} \tag{1}$$

$$Score = \frac{\sum_i \begin{cases} s_i, & \text{if } \mathit{No\ Error} \\ 0, & \text{otherwise} \end{cases}}{n} \tag{2}$$

Conclusions

- CLEF 2022 SimpleText track contains three interconnected shared tasks on scientific text simplification.
- We created a corpus of sentences extracted from the abstracts of scientific publications, with manual annotations of term complexity (Task 2) with regard to the queries from Task 1.
- We introduced a new classification of information distortion types for automatic simplification and we annotated the collected simplifications according to this error classification (Task 3).
- The HULAT-UC3M team submitted runs which combine tasks 2 and 3 which demonstrates strong interconnection of the tasks as often the terminology cannot be removed nor simplified but it needs to be explained to a reader.

Future work ?

- Task 1: topical relevance + text complexity + source authoritativeness
- Task 2: Provide explanations for difficult terms
- Task 3: expand the training and evaluation data + large-scale automatic evaluation measures

---

To discuss at the breakout session TOMORROW, Sep 6, at 9:30

- We want to hear from you!
- What was great about 2022, and what could we improve for you?

SimpleText program (ROOM F)

- Tue 06 Sep 2022 (TODAY)
  - 15:30 - 15:40 Welcome talk
  - 15:40 - 16:40 **Invited talk by Hosein Azarbonyad (Elsevier)**
    "Answers instead of articles: Helping users search and understand scientific content"
  - 16:40 - 18:50: Participants' presentations
  - After 19:00 **Social event** sponsored by Elsevier:
    Informal discussions over drinks and light food – every attendee of the session is invited!
- Wed 07 Sep 2022 (TOMORROW)
  - 8:50 - 9:30: Participants' presentations
  - 9:30 - 10:20 **Round table and SimpleText 2023 discussion**
    Any ideas or volunteers are welcome!

References

- Liana Ermakova, Eric SanJuan, Jaap Kamps, Stéphane Huet, Irina Ovchinnikova, Diana Nurbakova, Silvia Araujo, Radia Hannachi, Elise Mathurin, and Patrice Bellot (2022). Overview of the CLEF 2022 SimpleText Lab: Automatic Simplification of Scientific Texts. In A. Barrón-Cedeño, G. Da San Martino, M. Degli Esposti, F. Sebastiani, C. Macdonald, G. Pasi, A. Hanbury, M. Potthast, G. Faggioli, & N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Thirteenth International Conference of the CLEF Association (CLEF 2022)

TURN IT SIMPLE

CLEF 2023 Thessaloniki

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

Thank you !

We are hiring a PhD student !

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

Task 2: # of evaluated sentences per query

Query		# SNT	# SNT-term pairs
1	*guessing attack*	60	389
2	*end to end encryption*	55	390
3	*imbalanced data*	55	381
4	*distributed attack*	54	385
5	*genetic algorithm*	51	374
6	*quantum computing*	51	385
7	*qbit*	50	363
8	*side-channel attack*	49	340
9	*traffic optimization*	47	344
10	*quantum applications*	42	320
11	*cyber-security*	35	244
12	*conspiracy theories*	23	180
13	*crowsourcing*	15	104
14	*digital assistant*	5	32

Task 2: Term difficulty scale examples (1)

Grade	Non-abbreviated (ordinary) term	Abbreviation
7	The qubit—qutrit pair acts as a closed system and one external qubit serve as the environment for the pair.	We compared XCSFHP to XCSF on several problems.
6	This paper bring forward based on immune genetic algorithm to solve man on board automated storage and retrieval system optimized problem, immune genetic algorithm remains the characteristic which is not ... Tile coding is a well-known function approximator that has been successfully applied to many reinforcement learning tasks.	XCS with computed prediction, namely XCSF, extends XCS by replacing the classifier prediction with a parametrized prediction function. Side-channel attack (SCA) is a very efficient cryptanalysis technology to attack cryptographic devices.
5	Experiment simulation result express: the result of immune genetic algorithm is better than traditional genetic algorithm in the circumstance of the same clusters and the same evolution generation.	This paper presents a simple real-coded estimation of distribution algorithm (EDA) design using x-ary extended compact genetic algorithm (XECGA) and discretization methods.
4	Immune genetic algorithm can shorten storage or retrieval distance in application, and enhance storage or retrieval efficiency . Deep learning has become increasingly popular in both academic and industrial areas in the past years.	This paper presents a simple real-coded estimation of distribution algorithm (EDA) design using x-ary extended compact genetic algorithm (XECGA) and discretization methods.

Term difficulty scale examples (2)

Grade	Non-abbreviated (ordinary) term	Abbreviation
3	The XECGA is then used to build the probabilistic model and to sample a new population based on the probabilistic model.	We evaluate each measure's performance by AUC which is usually used for evaluation of imbalanced data classification.
2	Experiment simulation result express: the result of immune genetic algorithm is better than traditional genetic algorithm in the circumstance of the same clusters and the same evolution generation. Specifically, the real-valued decision variables are mapped to discrete symbols of user-specified cardinality using discretization methods.	Recently NIST has published the second draft document of recommendation for the entropy sources used for random bit generation.
1	video labeling game is a crowdsourcing tool to collect user-generated metadata for video clips. On the other hand, a 3dimensional (3D) map, which is one of major themes in machine vision research, has been utilized as a simulation tool in city and landscape planning, and other engineering fields.	*2D* (2-dimensional), *3D* (3-dimensional) *maps* as in The 3D maps will give more intuitive information compared to conventional 2-dimensional (2D) ones.
0	This device has two work modes: native and remote. The proposed rECGA is simple, making it amenable for further empirical and theoretical analysis.	However, Nam et al. pointed out...

Task 2: Examples of the annotation

Sentence	Term	Limits		Difficulty
		OK	Corrected	
This device has two work modes: ‘native’ and ‘remote’.	remote	YES		1
This device has two work modes: ‘native’ and ‘remote’.	work modes	YES		0
This device has two work modes: ‘native’ and ‘remote’.	modes native	NO	work modes	0
This device has two work modes: ‘native’ and ‘remote’.	device work	NO	device	0
This device has two work modes: ‘native’ and ‘remote’.	native remote	NO	native	1

Task 3: # of evaluated sentences per query

Query		# source SNT	# Simplified SNT
1	*digital assistant*	370	1,280
2	*conspiracy theories*	195	398
3	*end to end encryption*	55	102
4	*imbalanced data*	55	87
5	*genetic algorithm*	51	85
6	*quantum computing*	51	85
7	*qbit*	50	76
8	*quantum applications*	42	73
9	*cyber-security*	28	47
10	*fairness*	18	22
11	*crowsourcing*	14	21