CLEF 2024 SimpleText Track:
Improving Access to Scientific Texts for Everyone

Liana Ermakova Éric SanJuan Stéphane Huet
Hosein Azarbonyad Giorgio Di Nunzio Federica Vezzani
Jennifer D'Souza Salomon Kabongo Hamed Babaei Giglou
Yue Zhang Jaap Kamps

TURN IT SIMPLE CLEF INITIATIVE cnrs GDR Groupement de recherche MaDICS Masses de données, informations et connaissances en sciences anr NFDI4 DataScience

CLEF - **September 21, 2023**

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

Goals

TURN IT SIMPLE

- 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

SimpleText Track at CLEF 2023

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.

- *Task 4: SOTA? Tracking the State-of-the-Art* (NEW in 2024!)
  - Given scientific papers in artificial intelligence, extract related (Task, Dataset, Metric, Score) tuples that are reported in the work.

Benchmarks constructed in 2023

- Citation Network Dataset: DBLP+Citation, ACM Citation network
  - 4,232,520 abstracts in English
  - Topics = 40 press articles (from The Guardian & Tech Xplore)
  - with 114 manually extracted queries (keywords)
  - 152K sentences from retrieved abstracts are used in Task 2/3
- For **Task 1**: relevance judgements for 29 (train) + 34 (test) queries.
- For **Task 2**: 1,262 sentences with 5K sentence-terms pairs with difficulty level (Task 2.1) and 15K sentences with 300 concept explanations and 38K sentences with 5K abbreviation expansion (Task 2.2).
- For **Task 3**: parallel corpus of 648 (train) and 245 (test) manually simplified sentences

SimpleText 2023 Statistics

- Growing steadily: 74 registered teams, 20 submitted runs.

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

Shared Tasks 2024

- *Task 1: Content Selection: Retrieving passages to include in a simplified summary*
  - topical relevance
  - + text complexity scores (e.g., readability)
  - + authoritativeness scores (e.g., bibliometrics and altmetrics)
- *Task 2: Complexity Spotting: Identifying and explaining difficult concepts*
  - difficult term detection
  - + usefulness of the provided explanation with regard to a query
  - + difficulty of the provided explanation
- *Task 3: Text Simplification: Sentence and passage level simplification*
  - manual evaluation of information distortion & text complexity
  - + expand the training and evaluation data
  - + large-scale automatic evaluation measures
- *Task 4: SOTA? Tracking the State-of-the-Art in Scholarly Publications*
  - see next slide.

Task 4: SOTA?

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 Metrics: standard F1 metrics
- Evaluation Settings:
  - **Few-shot.** Test dataset includes (TDMS)'s seen in training.
  - **Zero-shot.** The test dataset includes (TDMS) with unseen T, D, or M.

SOTA as Simplification Task

- It simplifies a scientist's information access needs to easily track the best models on a task dataset based on (T,D,M,S) tuples, as opposed to having to manually scour for this information buried in the text of thousands of papers.

Planning Session

- We want to hear from you!
  - What was great about 2023, and what could we improve for you?
  - Any ideas or volunteers are welcome!
- Plans for 2024:

Task 1 Focus on relevance and text complexity, credibility, topic profiles

Task 2 Focus on explaining and contextualizing complex concepts

Task 3 Focus on sentence-level and paragraph-level text simplification

Task 4 New filtering/routing tracking the state-of-the-art in AI topics

- SimpleText roadmap:
  - Complete the current setup in 2024, new data/tasks/setup in 2025 (e.g. moving to ArXiv?)

Help us Run and Grow the Track!

- New organizers in Task 1:
  - . . . . (*fill in your name!*)
- New organizers in Task 2:
  - *Federica Vezzani*, Università di Padova, Italy
  - *Giorgio Di Nunzio*, Università di Padova, Italy
- New organizers in Task 3:
  - . . . . (*fill in your name!*)
- New organizers in Task 4:
  - *Jennifer D'Souza*, TIB – Leibniz Information Centre for Science and Technology and University Library, Germany
  - *Salomon Kabongo* (TIB), *Hamed Babaei Giglou* (TIB), and *Yue Zhang* (TU Berlin)
- **Internships available for SimpleText and Joker projects!**

TURN IT SIMPLE

CLEF INITIATIVE

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

anr

NFDI4 DataScience

Please join the SimpleText Track

Fully funded 3-years PhD 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