CLEF 2025 SimpleText Track

Simplify Scientific Text (and Nothing More)

Liana Ermakova **Team SimpleText**[1]

CLEF 2024, September 12, 2024, Grenoble, France

[1] There's a large team of organizers and supporters of individual tasks: Hosein Azarbonyad, Hamed Babaei Giglou, Jan Bakker, Stéphane Huet, Jaap Kamps, Benjamin Vendeville, Giorgio Maria Di Nunzio, Eric SanJuan, Jennifer D'Souza, Federica Vezzani, ...

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

Overview

- SimpleText Track setup similar 2021-2024
  - Very successful benchmarks constructed
  - "Finished" original tasks?
  - Major changes in setup and corpora in 2025
- CLEF 2025 SimpleText Track
  - *Simplify Scientific Text (and Nothing More)*
- The following 3+1 tasks:
  1. **Text Simplification**: *simplify scientific text.*
  2. ***Controlled Creativity***: *identify and avoid hallucination.*
  3. ***LeaderBoardQA***: *extract state-of-the-art from scientific text.*
  4. ***SimpleText 2024 Revisited***: *selected tasks by popular request.*

Task 1: Text Simplification

- *Task 1: Simplify Scientific Text*
  - New corpus!
    - Cochrane-auto is true document-level text simplification
    - More variation (sentence merge, order swaps) and discourse structure
    - Paragraph-level and sentence-level data realigned and restricted
  - Biomedical text – free to use, similar to existing TS corpora
    - Sentence-level (T1.1) and Document-level (T1.2) text simplification
    - Large-scale aligned train and test data
    - Free human judge effort for analysis...

	Cochrane-auto	**Newsela-auto**	**Wiki-auto**
Domain	Biomedical	News	General
# Doc Pairs	5,585	18,820	138,095
# Sent Pairs	35,800	813,972	685,769

- Align activities with TREC PLABA track (and CLEF BioASQ?)

Task 2: Controlled Creativity

- *Task 2: Identify and Avoid Hallucination*
- Have created huge collection of spurious/ over-generation content!
    - 2024: 47% of submissions > 10% spurious sentences, 19% > 50%...
    - Task 2.1: *to identify creative generation, at document level*
        - to detect what sentences are fully grounded on source input (a) without and (b) with access to the source sentences
        - -> also labels those introducing significant new content
        - post-hoc identification or explanation task
    - Task 2.2 *to avoid creative generation, and perform grounded generation by design*
        - asks to submit pairs of runs with/without source attribution by design
    - Task 2.3 *text alignment task*?
        - source attribution by optimal alignmentment of source and run output
    - Collab. with CLEF Joker/Eloquent/..., align with SemEval (Mu)Shroom.

Task 3: LeaderBoardQA

- *Task 3: Extract State-of-the-Art from Scientific Text*
- CLEF 2024 was the pilot year of the *SOTA? task* at SimpleText
- Continuation with more clear/attractive use-case (QA)?
- Create human supervised ground truth at scale?
- Interest in participation?

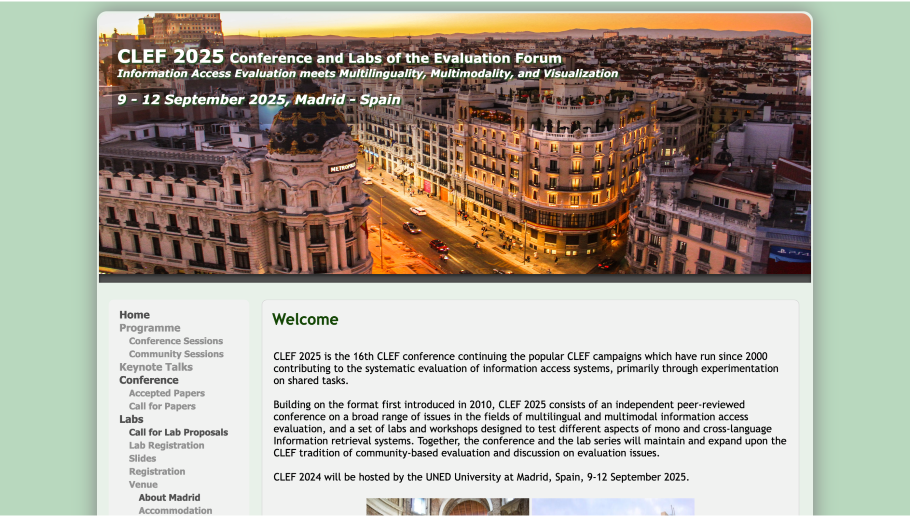
Task 4: SimpleText 2024 Revisited

- *Task 4: Selected Tasks by Popular Request*
- (Re)run selected 2024 tasks:
  - *Content Selection; Complexity Spotting; SOTA?*
- Based (only) on popular request...
  - Ease the transfer to new tasks and data
  - Stimulate reuse of build benchmarks
  - Encourage use and publication of new findings
- Can turn earlier data in CodaLabs leaderboards?
  - Allow for continuous submission
  - Allow for submitting CEUR papers (re)using earlier data
- *Collaborate with other tracks on a "Monster Track" on CLEF 2024?*

CLEF 2025 SimpleText Track

- *Task 1: Text Simplification: simplify scientific text*
  - + New aligned biomedical data (Cochrane-auto)
  - + both sentence, paragraph and document level simplification
  - + analysis of information distortion ("*hallucination?*")
- *Task 2: Controlled Creativity: identify and avoid hallucination*
  - + Real "hallucination" data from CLEF generative text tasks
  - + What output is (not) grounded on source(s)? (w/wo source access)
  - + How to avoid creative generation? (paired submissions)
- *Task 3: LeaderBoardQA: extract state-of-the-art from scientific text*
  - Extracting information on system performance from papers
  - + QA use case, + human ground truth evaluation
- *Task 4: SimpleText 2024 Revisited:selected tasks by popular request*
  - We take submissions of earlier tasks
  - release additional data and evaluation packages

Join us at CLEF 2025 in Madrid!



CLEF 2025 **Conference and Labs of the Evaluation Forum**

Information Access Evaluation meets Multilinguality, Multimodality, and Visualization

9 - 12 September 2025, Madrid - Spain

- **Home**
- Programme
  - Conference Sessions
  - Community Sessions
- Keynote Talks
- **Conference**
  - Accepted Papers
  - Call for Papers
- **Labs**
  - **Call for Lab Proposals**
  - Lab Registration
  - Slides
  - Registration
  - Venue
    - **About Madrid**
    - Accommodation

Welcome

CLEF 2025 is the 16th CLEF conference continuing the popular CLEF campaigns which have run since 2000 contributing to the systematic evaluation of information access systems, primarily through experimentation on shared tasks.

Building on the format first introduced in 2010, CLEF 2025 consists of an independent peer-reviewed conference on a broad range of issues in the fields of multilingual and multimodal information access evaluation, and a set of labs and workshops designed to test different aspects of mono and cross-language Information retrieval systems. Together, the conference and the lab series will maintain and expand upon the CLEF tradition of community-based evaluation and discussion on evaluation issues.

CLEF 2024 will be hosted by the UNED University at Madrid, Spain, 9-12 September 2025.

TURN IT SIMPLE

CLEF INITIATIVE

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

anr

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