CLEF 2023 SimpleText Track Guidelines

Automatic Simplification of Scientific Texts

The general public tends to avoid reliable sources such as scientific literature due to their complex language and lacking background knowledge. Instead, they rely on shallow and derived sources on the web and social media - often published for commercial or political incentives, rather than informational value. Can text simplification help to remove some of these access barriers? The SimpleText track is a part of the CLEF initiative which promotes the systematic evaluation of information access systems, primarily through experimentation on shared tasks. SimpleText addresses the challenges of text simplification approaches in the context of promoting scientific information access, by providing appropriate data and benchmarks. The track uses a corpus of scientific literature abstracts and popular science requests. Our overall use case is to create a simplified summary of multiple scientific documents based on a popular science query which provides a user with an accessible overview of this specific topic.

The track has the following three concrete tasks.

Tasks

- **Task 1:** What is in (or out)? Selecting passages to include in a simplified summary.
- **Task 2:** What is unclear? Difficult concept identification and explanation (definitions, abbreviation deciphering, context, applications,..).
- **Task 3:** Rewrite this! Given a query, simplify passages from scientific abstracts.

In addition, we welcome manual runs and any other type of submission that uses our data as an open task. Manual intervention should be reported.

How to participate & terms of use

In order to participate, you should sign up at the CLEF website (https://clef2023.clef-initiative.eu/index.php). All team members should join the SimpleText mailing list (https://groups.google.com/g/simpletext). The data will be made available to all registered participants.

By downloading and using these data, you agree to the terms of use. Any use of the data for any purpose other than academic research would be in violation of the intended use of these data.

Therefore, by downloading and using these data you give the following assurances with respect to the SimpleText data:

1. You will not use nor permit others to use the data in the SimpleText datasets in any way except for educational use and academic research.
2. You will not at any time disclose, give, or transmit (in any manner or form or for any purpose) the data (or any portion thereof) to any location or person, including but not limited to making the data available on the Internet and copying the data onto any cloud-based storage system.
3. You will not release nor permit others to release the dataset or any part of it to any person.

In case of violation of the conditions for access to the data for scientific purposes, this access may be withdrawn from the research entity and/or from the researcher. The research entity may also be liable to pay compensation for damages to third parties or asked to take disciplinary action against the offending researcher.

You will receive your login information from the address STID Avignon NextCloud admin@termwatch.es with the subject Your STID Avignon NextCloud account is created. Please check your spam folder if it is not the case. The accounts were created with the email addresses used for the registration. Usernames are identical to the usernames of the email addresses used for the registration without domains. You can use it to connect to NextCloud and change the password. Accounts from last year remain active.

You will have access to the shared SimpleText folder with data:

https://guacamole.univ-avignon.fr/nextcloud/index.php/apps/files/?dir=/simpleText

Each participant has the folder Documents:

https://guacamole.univ-avignon.fr/nextcloud/index.php/apps/files/?dir=/Documents

Run submission

For each task, participants can submit up to 10 runs by emailing them to contact@simpletext-project.com and simpletextworkshop@gmail.com until ~~**28 April 2023**~~ **10 May 2023**. The confirmation email will be sent within 2 days after the submission. Please contact the organizers if you do not receive a confirmation email. Besides, participants should put their run results into their personal folder "Documents" created on the data server.

The email subject has to be in the format [CLEF SimpleText TASK <NUMBER>] TEAM_ID (your registration name at CLEF), e.g. *[CLEF SimpleText TASK 1] UBO.*

Runs should be submitted as a ZIP folder of the corresponding JSON or TSV files. Please precise in the email the priority for each run 1-10 (which runs seem to be the best). We will manually evaluate runs depending on their priorities (one run with "priority":1 is guaranteed to be evaluated manually). Automatic evaluation could be possible for other runs.

Task 1: "What is in (or out)?" Select passages to include in a simplified summary, given a query

Given a popular science article targeted to a general audience, this task aims at retrieving passages that can help to understand this article, from a large corpus of academic abstracts and bibliographic metadata. Relevant abstracts should relate to any of the topics in the source article. These passages can be complex and require further simplification to be carried out in tasks 2 and 3. Task 1 focuses on content retrieval.

Corpus: DBLP abstracts

We use the Citation Network Dataset: DBLP+Citation, ACM Citation network (12th version released in 2020: https://www.aminer.org/citation). An ElasticSearch index is provided to participants with access through an API. A JSON dump of the index is also available for participants.

This index can be accessed online through queries, e.g. https://inex:qatc2011@guacamole.univ-avignon.fr/dblp1/_search?q=biases&size=1000 for the query "Biases"

login: inex

password: qatc2011

Topics: Press articles

Topics are a selection of press articles from the tech section of The Guardian newspaper (topics G01 to G20) and the Tech Xplore website (topics T01 to T20). URLs to original articles and textual content of each topic are provided to participants. All passages retrieved from DBLP by participants are expected to have some overlap (lexical or semantic) with the article content.

Queries as facets

Keywords queries are provided with each topic. It has been manually checked that each query allows retrieving relevant passages that could be inserted as citations in the press article.

Qrels

Quality relevance of abstracts w.r.t. topics are given in *Simpletext_2023_task1_train.qrels*. This file extends the qrels released with a significant increase of the depth of judgments of abstracts per query. Relevance annotations are provided on a 0-2 scale (the higher the more relevant) for 29 queries associated with the first 15 articles from the Guardian.

Participants in the previous edition of Task 1, who would like to access qrels used last year to reproduce their results, can contact us.

Expected results

Ad-hoc passage retrieval

Participants should retrieve, for each topic and each query, all passages from DBLP abstracts, related to the query and relevant to be inserted as a citation in the paper associated with the topic. Some passages could require simplification. We encourage participants to take into account passage complexity as well as its credibility/influentialness.

Open passage retrieval (optional)

Participants are encouraged to extract supplementary relevant queries from the titles or content articles and to provide results based on these supplementary queries.

Output format

Results should be provided in a TREC style JSON or TSV format with the following fields:

1. *run_id:* Run ID starting with <team_id>_<task_id>_<method_used>, e.g. *UBO_task_1_TFIDF*
2. *manual:* Whether the run is manual {0,1}
3. *topic_id:* Topic ID
4. *query_id:* Query ID used to retrieve the document (if one of the queries provided for the topic was used; 0 otherwise)
5. *doc_id:* ID of the retrieved document (to be extracted from the JSON output)
6. *rel_score:* Relevance score of the passage (in the [0-1] scale)
7. *comb_score:* General score that may combine relevance and other aspects: readability, citation measures…(in the [0-1] scale)
8. *passage:* Text of the selected passage

For each query, the maximum number of distinct DBLP references (*doc_id* field) must be 100 and the total length of passages should not exceed 1000 tokens.

The idea of taking into account complexity is to have passages easier to understand for non-experts, while credibility score aims at guiding them on the expertise of authors and the value of publication w.r.t. the article topic. For example, complexity scores can be evaluated using readability score and credibility scores using bibliometrics.

Here is an output format example:


```
[{"run_id":"ST_task_1_run1",
"manual":0,
"topic_id":"G01",
"query_id":"G01.1",
"doc_id":1564531496,
"rel_score":0.97,
"comb_score":0.85,
"passage":"A CDA is a mobile user device, similar to a Personal Digital Assistant (PDA). It supports the citizen when dealing with public authorities and proves his rights - if desired, even without revealing his identity."},

{"run_id":"ST_task_1_run1",
```

"manual":0,
"topic_id":"G01",
"query_id":"G01.1",
"doc_id":3000234933,
"rel_score":0.9,
"comb_score":0.9,
"passage":"People are becoming increasingly comfortable using Digital Assistants (DAs) to interact with services or connected objects"},

{"run_id":"ST_task_1_run1",
"manual":0,
"topic_id":"G01",
"query_id":"G01.2",
"doc_id":1448624402,
"rel_score":0.6,
"comb_score":0.3,
"passage":"As extensive experimental research has shown individuals suffer from diverse biases in decision-making."}]

Evaluation

Passage relevance will be assessed based on:

1. lexical and semantic overlap of extracted passages with topic article content
2. manual relevance assessment of a pool of passages (relevance scores provided by participants will be used to measure ranking quality)
3. manual assessment by non-expert users of credibility and complexity

NB: Outputs must be produced by participants **on the complete set of 40 topics**. Evaluation will be performed on a subset of topics including the 15 topics of the provided qrels.

Task 2: "What is unclear?" Difficult concept identification and explanation

The goal of this task is to identify key concepts that need to be contextualized with a definition, example, and/or use-case and provide useful and understandable explanations for them. Thus, there are two subtasks:

1. to retrieve up to 5 difficult terms in a given passage from a scientific abstract;

2. to provide an explanation (one/two sentences) of these difficult terms (e.g. definition, abbreviation deciphering, example, etc.).

For each passage, participants should provide a ranked list of difficult terms with corresponding difficulty scores on a scale of 0-2 (2 to be the most difficult terms, while the meaning of terms scored 0 can be derived or guessed) and definitions (optional). Passages (sentences) are considered to be independent, i.e. difficult term repetition is allowed. Detected concept spans and term and term difficulty will be evaluated. Term pooling and automatic metrics (accuracy of term binary classification, NDCG for term ranking, kappa statistics, and the similarity of the provided definitions to ground-truth definitions…) will be used to evaluate participants' results.

Data

We provide 3 test sets:

- Small
- Medium
- Large

The small dataset is included in the Medium one, while the latter is included in the Large one. The runs will be ranked based on the Small and Medium test sets. Additional scores will be provided for the Large test set.

Input format

The train and the test data are provided in JSON and TSV formats with the following fields:

1. *snt_id*: a unique passage (sentence) identifier
2. *doc_id*: a unique source document identifier
3. *query_id*: a query ID
4. *query_text*: difficult terms should be extracted from sentences with regard to this query
5. *source_snt*: passage text

Input example:


```
[{"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."
},
```

{"snt_id":"G06.2_2548923997_3",
"doc_id":2548923997,
"query_id":"G06.2",
"query_text":"self driving",
"source_snt":"These communication systems render self-driving vehicles vulnerable to many types of malicious attacks, such as Sybil attacks, Denial of Service (DoS), black hole, grey hole and wormhole attacks."}]

Output format

Results should be provided in a TREC style JSON or TSV format with the following fields:

1. *run_id:* Run ID starting with <team_id>_<task_id>_<method_used>, e.g. *UBO_task_**2.1**_TFIDF*
2. *manual:* Whether the run is manual {0,1}
3. *snt_id*: a unique passage (sentence) identifier from the input file
4. *term*: Term or another phrase to be explained
5. *term_rank_snt*: term difficulty rank within the given sentence
6. *difficulty*: difficulty scores of the retrieved term on the scale 0-2 (2 to be the most difficult terms, while the meaning of terms scored 0 can be derived or guessed)
7. *definition (optional - please precise task_2.2 in this case)*: short (one/two sentence) explanations/definitions for the terms. For the abbreviations, the definition would be the extended abbreviation.

Output example Task 2.1

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

Output example Task 2.2

{"snt_id":"G14.2_2884788726_2",
"term":"content caching",
"difficulty":1.0,
"term_rank_snt":1,

"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}

Task 3: Rewrite this! Rewriting scientific text

This task aims to provide a simplified version of sentences extracted from scientific abstracts. We will do a large-scale evaluation of participants' results based on automatic measures (SARI, ROUGE, compression, readability) and a small-scale detailed human evaluation of other aspects, including information distortion.

Data

We provide a parallel corpus of 648 manually simplified sentences as train data.

We provide 3 test sets:

- Small
- Medium
- Large

The small dataset is included in the Medium one, while the latter is included in the Large one. The runs will be ranked based on the Small and Medium test sets. Additional scores will be provided for the Large test set.

Input format

The train and the test data are provided in JSON and TSV formats with the following fields:

1. *snt_id*: a unique passage (sentence) identifier
2. *doc_id*: a unique source document identifier
3. *query_id*: a query ID
4. *query_text*: difficult terms should be extracted from sentences with regard to this query
5. *source_snt*: passage text

Input example

{"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

Results should be provided in a TREC style JSON or TSV format with the following fields:

1. *run_id:* Run ID starting with <team_id>_<task_id>_<method_used>, e.g. *UBO_task_3_BLOOM*
2. *manual:* Whether the run is manual {0,1}
3. *snt_id*: a unique passage (sentence) identifier from the input file
4. *simplified_snt*: simplified passage

Output example

{"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."}

Contacts	Deadlines
SimpleText website: http://simpletext-project.com/	Registration: **28 April 2023**
CLEF website: https://clef2023.clef-initiative.eu/index.php	Run submission: ~~**28 April 2023**~~ **10 May 2023**
Registration: http://clef2023-labs-registration.dei.unipd.it/	Results available: ~~**10 May2023**~~ **20 May 2023**
Email: contact@simpletext-project.com	Draft paper submission: **5 June 2023**
Twitter: https://twitter.com/SimpletextW	Camera-ready: **7 July 2023**
Google group: https://groups.google.com/g/simpletext	CLEF conference: **18-21 September 2023**