SimpleText 2021

Text Simplification for Scientific Information Access
CLEF 2021 Workshop

Liana Ermakova, Eric San-Juan, Josiane Mothe, Jaap Kamps, Pavel Braslavski, Patrice Bellot, Irina Ovchinnikova, Diana Nurbakova

- Scientific publications are difficult to read
- Fight against misinformation
- Faster reading
- Accessibility to
  - Non-native
  - Younger readers
  - Citizens with reading disabilities
- Improving the results of NLP applications for pre-editing or translation
- Useful for:
  - Scientific communication
  - Science journalism
  - Political communication
  - Education

Motivation&Objectives

- Bringing together an interdisciplinary scientific community
- Definition & Methods
- Contribute to the response to challenges:
  - Technical
  - Evaluation
- Open and accessible science

Format & Call for Contributions

Half-day workshop:

- Introduction & Welcome (15 min)
- Invited talk (1 h)
- Presentations of the participants (15 min + 5 min questions)
- Open discussion (30 min)
- Closing remarks (10 min)

Types of contributions:

- **Participation in the pilot tasks!**
- Research & survey papers
- Position, discussion & demo paper
- Extended abstracts of published papers

Topics of interest (not exhaustive)

- Automated or computer-assisted scientific popularization/simplification
- Contextualization, search for background knowledge
- Terminology extraction
- Methods for assessing language complexity
- Methods for assessing information complexity
- Automatic summarization of scientific texts
- Daily digest generation
- Simplification of technical text, computer-assisted pre-editing
- Alteration and distortion of scientific information
- Automatic methods for scientific/data journalism

Pilot Tasks

Guidelines:

https://simpletext-madics.github.io/data/Guideline-SimplText-2021.pdf

PILOT TASK 1 : Content Selection

Select passages to include in a simplified summary, given a query

Queries: titles of scientific journalism articles + keywords

Data: ElasticSearch index of Citation Network Dataset: DBLP+Citation, ACM Citation network

Evaluation: pooling, traditional IR metrics, unresolved anaphora,...

Potential problems:

- The information in a summary designed for an expert is different from those for the general audience
- Relevance of the source
- Unresolved anaphora
- ...

PILOT TASK 1 : Example

Input:

```
<topic>
    <topic_id>1</topic_id>
    <topic_text>Digital assistants like Siri and Alexa entrench gender biases, says UN</topic_text>
    <keywords>
        <keyword>Digital assistant</keyword>
        <keyword>Biases</keyword>
    </keywords>
</topic>
```

Expected output:

```
run_id  manual  topic_id  doc_id  passage  rank
ST_1  1  1  3000234933  People are becoming increasingly comfortable using Digital Assistants (DAs) to interact with services or connected objects.  1
ST_1  1  1  3003409254  big data and machine learning (ML) algorithms can result in discriminatory decisions against certain protected groups defined upon personal data like gender, race, sexual orientation etc.  2
ST_1  1  1  3003409254  Such algorithms designed to discover patterns in big data might not only pick up any encoded societal biases in the training data, but even worse, they might reinforce such biases resulting in more severe discrimination.  3
```

PILOT TASK 2:
Searching for concepts to be explained

Given a passage and a query, rank terms/concepts that are required to be explained for understanding this passage (definitions, context, applications,..)

Queries: titles of scientific journalism articles + keywords

Data: DBLP abstracts

Evaluation: NDCG?,...

Potential extension in future:

- Provide a context
- ...

PILOT TASK 2 : Example

Input:

<topic>
<topic_id>1</topic_id>
<topic_text>Digital assistants like Siri and Alexa entrench gender biases, says UN</topic_text>
<passage_id>1</passage_id>
<passage_text>Automated decision making based on big data and **machine learning** (**ML**) algorithms can result in discriminatory decisions against certain protected groups defined upon personal data like gender, race, sexual orientation etc. Such algorithms designed to discover patterns in big data might not only pick up any encoded **societal biases** in the training data, but even worse, they might reinforce such biases resulting in more severe discrimination.
</passage_text>
</topic>

Expected output:

Run_id	manual	topic_id	passage_id	term	rank
ST_1	1	1	1	machine learning	1
ST_1	1	1	1	societal biases	2
ST_1	1	1	1	ML	3

PILOT TASK 3: Language Simplification

Given a query, simplify passages from scientific abstracts

Queries: titles of scientific journalism articles + keywords

Data: DBLP abstracts

Evaluation: manual? Aggregated metrics?

Potential problems:

- Is it possible to simplify terminology? ⇒ Pilot task 2: background knowledge
- Out of scope of consideration: puns and idioms

PILOT TASK 3: Example

Input:

<topic>

<topic_id>1</topic_id>

<topic_text>Digital assistants like Siri and Alexa entrench gender biases, says UN</topic_text>

<passage_id>1</passage_id>

<passage_text>Automated decision making based on big data and machine learning (ML) algorithms can result in discriminatory decisions against certain protected groups defined upon personal data like gender, race, sexual orientation etc. Such algorithms designed to discover patterns in big data might not only pick up any encoded societal biases in the training data, but even worse, they might reinforce such biases resulting in more severe discrimination.

</passage_text>

</topic>

Expected output:

Run_id	*manual*	*topic_id*	*passage_id*	*simplified_passage*
ST_1	*1*	*1*	*1*	*Automated decision-making may include sexist and racist biases and even reinforce them because their algorithms are based on the most prominent social representation in the dataset they use.*

Organizers

- **Liana Ermakova,** HCTI - EA 4249, Université de Bretagne Occidentale (Brest, France)
- **Eric San-Juan,** Laboratoire d'Informatique d'Avignon,Institut de technologie d' Avignon (Avignon, France)
- **Josiane Mothe,** INSPE, Université de Toulouse, IRIT, UMR5505 CNRS (Toulouse, France)
- **Jaap Kamps,** Faculty of Humanities, University of Amsterdam (Amsterdam, Netherland)
- **Pavel Braslavski,** Combinatorial Algebra Lab, Ural Federal University, (Yekaterinburg, Russia)
- **Patrice Bellot,** Aix-Marseille Université - CNRS (LIS – INS2I) (Marseille, France)
- **Irina Ovchinnikova,** Institute of Linguistics and Intercultural Communication, Sechenov University (Moscow, Russia)
- **Diana Nurbakova,** LIRIS, Institut National des Sciences Appliquées de Lyon, (Lyon, France)

We are open to discuss other ideas

Thank you!

Questions? Suggestions?

Participate in SimpleText@CLEF!

Website: https://www.irit.fr/simpleText/

E-mail: simpletextworkshop@gmail.com

Twitter: https://twitter.com/SimpletextW

Google group: https://groups.google.com/g/simpletext

CLEF website: http://clef2021.clef-initiative.eu/index.php