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

TURN IT SIMPLE

- 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

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

Scenarios

- To create a simplified summary of multiple scientific documents based on a given query which provides the user with an instant simplified summary on a specific topic they are interested in or to generate a daily digest, for example for ArXiv

Simplification pipeline

Select content

Which documents and passages should be included in the simplified summary?

Explain difficult concepts

Which terms should be contextualised by giving a definition and/or application ?

Simplify language

How to reduce vocabulary and syntactic complexity with acceptable rate of information distortion?

TURN IT SIMPLE

Data

- Citation Network Dataset: DBLP+Citation, ACM Citation network (https://www.aminer.org/citation)
- DBLP full dump in the JSON.GZ format
- DBLP abstracts extracted for each topic in the following MD format

1551421219	2010	Online advertising has been fueling the rapid growth
2052650089	2012	Although online product reviews have emerged as an i
1571655962	2014	Purpose – The purpose of this paper is to investigat
2101571056	2005	This paper examines the practice of advertising with
2121552264	2012	The value proposition of mobile technology for educa

TURN IT SIMPLE

CLEF 2021 BUCHAREST

Queries

- Press titles from *The Guardian* with manually extracted keywords
- Each keyword allows to extract at least 5 relevant abstracts
- Full text articles from The Guardian (link, folder query_related_content with full texts in the MD format)

Query 1: Digital assistants like Siri and Alexa entrench gender biases, says UN

https://www.theguardian.com/technology/2019/may/22/digital-voice-assistants-siri-alexa-gender-biases-unesco-says

Topic 1.1: Digital assistant

https://inex:qatc2011@guacamole.univ-avignon.fr/dblp1/_search?q="Digital assistant"&size=1000

Topic 1.2: Biases

https://inex:qatc2011@guacamole.univ-avignon.fr/dblp1/_search?q=biases&size=1000

TURN IT SIMPLE

CLEF 2021 BUCHAREST

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
- ...

TURN IT SIMPLE

CLEF 2021 BUCHAREST

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

TURN IT SIMPLE

CLEF 2021 BUCHAREST

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
- ...

TURN IT SIMPLE

CLEF 2021 BUCHAREST

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

TURN IT SIMPLE

CLEF 2021 BUCHAREST

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

TURN IT SIMPLE

CLEF 2021 BUCHAREST

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.

SimpleText@CLEF'21 overview & lessons learned

- 43 registered teams
- 23 participants subscribed on our Google group
- 24 followers on Twitter
- Data was downloaded from the server by several participants, but no submitted runs → data can be reused
- We enriched data prepared for the pilot tasks & we will provide a baseline
- New data can be released in autumn 2021 → more time to potential participants
- Unshared task → possibility to find new data applications
- New team members
- Advertise more

Programme

22 September

15:30 - 15:50: S.Araújo & R.Hannachi “Multimodal science communication: from documentary research to infographic via mind mapping”

15:50 - 16:30: *Invited talk*: N.Grabar & R.Cardon “Various factors of the evaluation of text simplification”

16:30-17:00: Interactive session: Evaluation of text and terminology difficulty

17:30-18:10: *Invited talk*: Wei Xu “Importance of Data and Controllability in Neural Text Simplification”

18:10 - 18:30: M.Hajjem & E.Sanjuan “Societal trendy multi word term extraction from DBLP”

18:30 - 19:00: Interactive session: How to select information for science simplification?

23 September

17:30-18:00: *Industrial talk*: Mike Unwalla “TechScribe”

18:00 - 18:20: I.Ovchinnikova, D.Nurbakova & L.Ermakova “What Science-Related Topics Need to Be Popularized? A Comparative Study”

18:20 - 19:00: *Invited talk*: John Rochford “Using AI and Crowdsourcing to Simplify COVID-19 Info Worldwide”

TURN IT SIMPLE

Thank you!

Website: https://simpletext-madics.github.io/2021/

E-mail: simpletextworkshop@gmail.com

Twitter: https://twitter.com/SimpletextW

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