Overview of SimpleText CLEF 2021 Workshop and Pilot Tasks

Liana Ermakova[1], Patrice Bellot[2], Pavel Braslavski[3], Jaap Kamps[4], Josiane Mothe[5], Diana Nurbakova[6], Irina Ovchinnikova[7] and Eric San-Juan[8]

[1]HCTI - EA 4249, Université de Bretagne Occidentale, 20 Rue Duquesne, 29200 Brest, France
[2]Université de Aix-Marseille, LIS, France
[3]Ural Federal University, Yekaterinburg, Russia
[4]University of Amsterdam, Amsterdam, The Netherlands
[5]Université de Toulouse, IRIT, France
[6]Institut National des Sciences Appliquées de Lyon, Lyon, France
[7]Sechenov University, Moscow, Russia
[8]Avignon Université, LIA, France

Abstract
Scientific literacy is important for people to make right decisions, evaluate the information quality, maintain physiological and mental health, avoid spending money on useless items. However, since scientific publications are difficult for people outside the domain and so they do not read them at all even if they are accessible. Text simplification approaches can remove some of these barriers to use scientific information, thereby promoting the use of objective scientific findings and avoiding that users rely on shallow information in sources prioritizing commercial or political incentives rather than the correctness and informational value. The CLEF 2021 SimpleText workshop addresses the opportunities and challenges of text simplification approaches to improve scientific information access head-on. This year, we run three pilot tasks trying to answer the following questions: (1) What information should be simplified? (2) Which terms should be contextualized by giving a definition and/or application? (3) How to improve the readability of a given short text (e.g. by reducing vocabulary and syntactic complexity) without significant information distortion?

Keywords
Scientific text simplification, (Multi-document) summarization, Contextualization, Background knowledge

1. Introduction

Scientific literacy, including health related questions, is important for people to make right decisions, evaluate the information quality, maintain physiological and mental health, avoid spending money on useless items. For example, the stories the individuals find credible can determine their response to the COVID-19 pandemic, including the application of social distancing, using dangerous fake medical treatments, or hoarding. Unfortunately, stories in social

CLEF 2021 – Conference and Labs of the Evaluation Forum, September 21–24, 2021, Bucharest, Romania
liana.ermakova@univ-brest.fr (L. Ermakova); kamps@uva.nl (J. Kamps)
0000-0002-7598-7474 (L. Ermakova); 0000-0002-6614-0087 (J. Kamps)
© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)
CEUR-WS.org/Vol-2936/paper-199.pdf

media are easier for people to understand than the research papers. Scientific texts such as scientific publications can also be difficult to understand for non domain-experts or scientists outside the publication domain. Improving text comprehensibility and its adaptation to different audience remains an unresolved problem. Although, there are some attempts to tackle the issue of text comprehensibility, they are mainly based on readability formulas, which are not convincingly demonstrated the ability to reduce the difficulty of text [1].

To put a step forward to automatically reduce understanding difficulty of a text, we propose a new workshop called **SimpleText** which aims to create a community interested in generating simplified summaries of scientific documents. Thus, the goal of this workshop is to connect researchers from different domains, such as Natural Language Processing, Information Retrieval, Linguistics, Scientific Journalism etc. to work together on automatic popularisation of science.

Improving text comprehensibility and its adaptation to different audience bring societal, technical, and evaluation challenges. There is a large range of important *societal challenges* SimpleText is linked to. Open science is one of them. Making the research really open and accessible for everyone implies providing them in a form that can be readable and understandable; referring to the "comprehensibility" of the research results, making science understandable [2]. Another example of those societal challenges is offering means to develop counter-speech to fake news based on scientific results. SimpleText also tackles *technical challenges* related to data (passage) selection and summarisation, comprehensibility and readability of texts.

To face these challenges, SimpleText provides an open forum aiming at answering questions like:

- What information should be simplified (e.g. in terms document and passage selection and summarisation)?
- What kind of background information should be provided (e.g. which terms should be contextualized by giving a definition and/or application)? What information is the most relevant or helpful?
- How to improve the readability of a given short text (e.g. by reducing vocabulary and syntactic complexity) without information distortion?

We will provide data and benchmarks, and address evaluation challenges underlying the technical challenges, including:

- How to measure text simplification?
- How to evaluate background information?
- How to evaluate information selection?

2. Data Selection, Comprehensibility, Readability

2.1. Data Selection

People have to manage the constantly growing amount of information. According to several estimates the number of scientific journals is around 30,000, with about two million articles published per year [3]. According to scholarly information platform Dimensions[1], from January

[1]https://www.dimensions.ai/

2020 to October 2020, about 180,000 articles on Covid-19 were published [4]. To deal with this data volume, one should have a concise overview, i.e. a summary. People prefer to read a short document instead of a long one. Thus, even single-document summarization is already a step of text simplification. Notice, that the information in a summary designed for a scientist from a specific field should be different from that adapted for general public.

Despite recent significant progress in the domains of information retrieval and natural language processing (NLP), the problem of constructing consistent overview has not yet been solved. Automatic text summarization is one of the popular NLP and information access tasks since a pioneering paper by Luhn [5]. Automatic summarization can simplify access to primary scientific documents – the resulting concise text is expected to highlight the most important parts of the document and thus reduces the reader's efforts. Early studies developed automatic summarization methods for scientific and technical documents. Evaluation initiatives in the 2000s such as Document Understanding Conference (DUC) and the Summarization track at the Text Analysis Conference (TAC) have focused primarily on the automatic summarization of news in various contexts and scenarios. Modern methods of automatic summarization are trained and tested on large collections of news [6] or social media texts [7]. Scientific articles are typically provided with a short abstract written by the authors. Thus, automatic generation of an abstract for a stand-alone article does not seem to be a practical task. However, if we consider a large collection of scientific articles and citations between them, we can come to a task of producing an abstract that would contain important aspects of a paper from the perspective of the community. Such a task has been offered to the participants of the TAC 2014 Biomedical Summarization Track[2], as well as of the CL-SciSumm shared task series. In particular, the 2020 edition of CL-SciSumm features LaySummary subtask, where a participating system must produce a text summary of a scientific paper intended for non-technical audience.[3]

Sentence compression can be seen as a middle ground between text simplification and summarization. The task is to remove redundant or less important parts of an input sentence, preserving its grammaticality and original meaning [8]. Thus, the main challenge is to *choose which information* should be included in a simplified text.

2.2. Comprehensibility

Readability, comprehensibility and usability are key points of the information evaluation [9]. The most recent works in the text comprehension field show various approaches to explain stages and strategies of text comprehension in children, bilinguals, and adults with reading / learning disabilities. Comprehensibility of a simple text varies for different readership. Readers of popular science texts have a basic background, are able to process logical connections and recognize a novelty [10]. In the popular science text, a reader looks for rationalization and clear links between well known and new [11]. To adopt the novelty, readers need to include new concepts into their mental representation of the scientific domain.

According to The Free Dictionary, *background knowledge* is "information that is essential to understanding a situation or problem"[12]. Lack of basic knowledge can become a barrier to reading comprehension [13]. In [13], the authors suggested that there is a knowledge threshold

[2]https://tac.nist.gov/2014/BiomedSumm/
[3]https://ornlcda.github.io/SDProc/sharedtasks.html#laysumm

allowing reading comprehension. Background knowledge, along with content, style, location, and some other dimension, are useful for personalised learning [14]. In contrast to newspapers limited by the size of the page, digital technologies provide essentially unbounded capabilities for hosting primary-source documents and background information. However, in many cases users do not read these additional texts. It also is important to remember, that the goal is to keep the text simple and short, not to make it indefinitely long to discourage potential readers.

Entity linking (also known as Wikification) is the task of tying named entities from the text to the corresponding knowledge base items. A scientific text enriched with links to Wikipedia or Wikidata can potentially help mitigate the background knowledge problem, as these knowledge bases provide definitions, illustrations, examples, and related entities. However, the existing standard datasets for entity linking such as [15] are focused primarily on such entities as people, places, and organizations, while a lay reader of a scientific article needs rather assistance with new concepts, methods, etc. Wikification is close to the task of terminology and keyphrase extraction from scientific texts [16].

Thus, the main challenge of the comprehensibility is to *provide relevant background knowledge* to help a reader to understand a complex scientific text.

2.3. Readability

Readability is the ease with which a reader can understand a written text. It is part of the so called *information nutritional label* which aims at helping users to analyze information objectively [17].

Readability is different from legibility, which measures how easily a reader can distinguish characters from each other. Readability indices have been widely used to evaluate teaching materials, news, and technical documents for about a century [18, 19]. For example, Gunning fog index, introduced in 1944, estimates the number of years in a scholar system required to understand a given text on the first reading.Similarly, the Flesch–Kincaid readability tests shows the difficulty of a text in English based on word length and sentence length [20]. Although these two metrics are easy to compute, they are criticized for the lack of reliability [21]. The very structure of the readability indices suggested to authors or editors how to simplify a text: organize shorter and more frequent words into short sentences. Later studies incorporate lexical, syntactic, and discourse-level features to predict text readability [22].

In NLP tasks, readability, coherence, conciseness, and grammar are usually assessed manually since it is difficult to express these parameters numerically [23]. However, several studies were carried out in the domain of automatic readability evaluation, including the application of language models [21, 24, 25, 26] and machine learning techniques [27, 26]. Traditional methods of readability evaluation are based on familiarity of terms [28, 29, 30] or their length [31] and syntax complexity (e.g. sentence length, the depth of a parse tree, omission of personal verb, rate of prepositional phrases, noun and verb groups etc.) [24, 32, 33, 34, 35]. Word complexity is usually evaluated by experts [36, 29, 30]. [37] computed average normalized number of words in valid coherent passages without syntactical errors, unresolved anaphora, and redundant information. Several researches argue also the importance of sentence ordering for text understanding [38, 39].

Automatic text simplification might be the next step after estimation of text complexity. Usually, text simplification task is performed and assessed on the level of individual sentences. To reduce the reading complexity, in [40], the authors introduced a task of sentence simplification through the use of more accessible vocabulary and sentence structure. They provided a new corpus that aligns English Wikipedia with Simple English Wikipedia and contains simplification operations such as rewording, reordering, insertion and deletion. Accurate lexical choice presupposes unambiguous reference to the particular object leading to actualization of its connections with other objects in the domain. Domain complexity concerns the number of objects and concepts in the domain, and connections among them described by the terminology system (see a survey: [41]). Names of the objects are not replaceable in the process of text transformation or simplification due to risk of information distortion [42, 43]. For example, 'hydroxychloroquine' represents a derivative of 'chloroquine', so the substances are connected thanks to belonging to a set 'chloroquine derivatives'. However, it is impossible to substitute 'hydroxychloroquine' by 'chloroquine' while simplifying a medical text about a Covid-19 treatment because of the difference in their chemical composition. A hypernym 'drugs' can refer to the substances. The hyperonym generalizes the information while omitting essential difference between the drugs; however, the generalization allows to avoid misinformation [44]. Science text simplification presupposes facilitation of readers' understanding of complex content by establishing links to basic lexicon, avoiding distortion connections among objects within the domain.

Ideally, the results undergo a human evaluation, since traditional readability indices can be misleading [45]. Automatic evaluation metrics have been proposed for the task: SARI [46] targets lexical complexity, while SAMSA estimates structural complexity of a sentence [47]. Formality style transfer is a cognate task, where a system rewrites a text in a different style preserving its meaning [48]. These tasks are frequently evaluated with BLEU metrics [49] to compare system's output against gold standard.

Thus, the main challenge of the readability improvement is to *reduce vocabulary and syntactic complexity* without information distortion while keeping the target genre.

3. Data set

3.1. Collection

For this edition we use the Citation Network Dataset: DBLP+Citation, ACM Citation network [4]. An elastic search index is provided to participants accessible through a GUI API. This Index is adequate to:

- apply basic passage retrieval methods based on vector or language IR models;
- generate Latent Dirichlet Allocation models;
- train Graph Neural Networks for citation recommendation as carried out in StellarGraph[5] for example;
- apply deep bi directionnal transformers for query expansion;
- and much more …

[4]https://www.aminer.org/citation
[5]https://stellargraph.readthedocs.io/

One of the important problems in manual text simplification is a cognitive bias called the curse of knowledge, which occurs when an individual assumes that their interlocutor has the background to understand them. To leverage this issue, we simplify text passages issued from computer science articles abstracts by a pair of experts. One annotator is a computer scientist who understands the text and simplifies passages. Then each pair of passages (simplified and not) is reread by a professional translator from the University of Western Brittany Translation Office[6] who is an English native speaker but not a specialist in computer science. Each passage is discussed and rewritten multiple times until it becomes clear for non computer scientists. The observation of the obtained simplification examples revealed opposite strategies in making text understandable. On the one hand, shortening passages by eliminating details and generalization seem an efficient strategy. On the other hand, simplified sentences are longer and more concrete, e.g. the sentence from an article on exposing image tampering “The learning classifiers are applied for classification” was simplified as “The machine learning algorithms are applied to detect image manipulation”. For a computer scientist, it is evident that the detection problem is a special case of a binary classification task, but in order to make this sentence understandable for a non computer scientist, the abstract term “classification” should be replaced with a concrete use-case “to detect image manipulation”. Thus, on the one hand our methodology of passage simplification ensures data quality. On the other hand, it provides interesting insights to simplification strategies.

Simplification efficacy depends on the subject: how wide is it spread and whether the field it belongs to is sophisticated or well known thanks to the basic level of education? Nevertheless, every subject can be simplified by improving the text readability and providing background information to improve its comprehensibility. Meanwhile, improving the comprehensibility may bring in distortion of the original content. While selecting the materials for queries, we provide various opportunities to work on simplifications. Thirteen queries are associated with a set of demanded topics including global markets and cryptocurrencies, social media regulation, medicine, technologies and ethical problems caused by AI. The extracted keywords are relevant in retrieving resources that provide relevant information, which is of importance for readers to understand the topic. We retrieved sentences from the documents from the Citation Network Dataset using these keywords to work out manual simplification. Our enhancement aims at readability and comprehensibility of the sources.

In Table 1, the instances of our simplification are shown. The third column of the table contains target sentences. The fourth column includes various types of the simplified sentences:

a. sentences simplifying original construction;
b. sentences generalizing original content;
c. sentences providing basic information;
d. sentences that are easy to read and to comprehend;
e. sentences that provide explanations of terminology.

The (a) case is shown in the first row, where content of the target sentence with the key word ‘misinformation’ is simplified through eliminating heavy construction. It also explains the term "confirmation bias". The (b) case is represented in the second row. Simplification of the

[6] https://www.univ-brest.fr/btu

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

Figure 1: Query example

1564531496	2002	In this short paper we describe the architectural cc
2988211052	2002	In this short paper we describe the architectural cc
3006661050	2003	Modern Personal Digital Assistant (PDA) architecture
1970213811	2006	This demonstration presents a new interaction techni
2797641221	2018	Digital assistants are emerging to become more preva
2158159346	2004	Abstract Mobile devices are significantly changinc
2463945949	2016	DIANE is a digital assistant system that aims to fas

Figure 2: DBLP abstract examples

target sentence with the keyword ‘financial markets’ is reached by eliminating the description of the automatized trading on financial markets. The (c) case in the third row shows how content simplification deletes some original content. The simplified sentence expresses basic information instead of comparison of the sensor based control in managing interaction between a robot and its environment over joint positions and velocities. The (d) case is represented by an omission of particular details about JPEG images that does not damage the original content. The simplified sentence about guest and host virtual machines (e) provides an explanation of the terms ‘guest’ and ‘host’ enhancing comprehensibility of the target sentence, nevertheless simplification leads to enlarging the text volume.

57 manually simplified passages were provided to participants for training.

3.2. Queries

For this edition 13 queries are a selection of recent n press titles from The Guardian enriched with keywords manually extracted from the content of the article. It has been checked that each keyword allows to extract at least 5 relevant abstracts. The use of these keywords is optional.

Input format for all tasks:

- Topics in the MD format (see Fig. 1);
- Full text articles from The Guardian (link, folder query_related_content with full texts in the MD format);
- ElasticSearch index on the data server [7];
- DBLP full dump in the JSON.GZ format;

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

Table 1
Simplification typology examples

Simplification type	Keyword / Title of The Guardians article	Original sentence	Simplified sentence
(a) construction simplification	misinformation / Misinformation runs rampant as Facebook says it may take a week before it unblocks some pages	Simultaneously, they allow the spread of misinformation by empowering individuals to self-select the narratives they want to be exposed to, both through active (confirmation bias) and passive (personalized news algorithms) self-reinforcing mechanisms.	But misinformation is spread via social media because individuals can search for information that confirms their beliefs and personalized news algorithms may supply it.
(b) generalization	financial markets / Bitcoin's market value exceeds $1tn after price soars	Construction of BSE was motivated by the fact that most of the world's major financial markets have automated, with trading activity that previously was the responsibility of human traders now being performed by high-speed autonomous automated trading systems.	BSE was built because most of the financial markets became automated.
(c) basic information	humanoid robots / Robots on the rise as Americans experience record job losses amid pandemic	Furthermore, for service and manipulation tasks, it is more suitable to study the interaction between the robot and its environment at the contact point using the sensor based control, rather than specifying the joint positions and velocities required to achieve them.	Interaction between the robot and its environment using the sensor based control is important.
(d) omission of details	forensics / Forensic Architecture: detail behind the devilry	As the most popular multimedia data, JPEG images can be easily tampered without leaving any clues; therefore, JPEG-based forensics, including the detection of double compression, interpolation, rotation, etc., has become an active research topic in multimedia forensics.	JPEG images can be easily manipulated without leaving any clues. This is why researchers are trying to develop methods for JPEG image manipulation detection.
(e) terminology explanation	forensics / Forensic Architecture: detail behind the devilry	Guest virtual machines are especially vulnerable to attacks coming from their (more privileged) host.	Guest virtual machines use computing resources provided by a physical machine called a host. Guest virtual machines are especially vulnerable to attacks coming from their host.

- DBLP abstracts extracted for each topic in the following MD format (doc_id, year, abstract) (see Fig.2).

Table 2
Task 1 output example

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

4. Pilot tasks

To start with, we will develop three pilot tasks that will help to better understand the challenges as well to discuss these challenges and the way to evaluate solutions. Details on the tasks, guideline and call for contributions can be found at www.irit.fr/simpleText, in this paper we just briefly introduce the planed pilot tasks. Note that the pilot tasks are means to help the discussions and to develop a research community around text simplification. Contributions will not exclusively rely on the pilot tasks.

4.1. Task 1: Selecting passages to include in a simplified summary - Content Simplification

Given an article from a major international newspaper general audience, this pilot task aims at retrieving from a large scientific bibliographic database with abstracts, all passages that would be relevant to illustrate this article. Extracted passages should be adequate to be inserted as plain citations in the original paper.

Sentence pooling and automatic metrics will be used to evaluate these results. The relevance of the source document will be evaluated as well as potential unresolved anaphora issues.

Output: A maximum of 1000 passages to be included in a simplified summary in a TSV (Tab-Separated Values) file with the following fields:

- *run_id*: Run ID starting with *team_id_*;
- *manual*: Whether the run is manual 0,1;
- *topic_id*: Topic ID;
- *doc_id*: Source document ID;
- *passage*: Text of the selected passage;
- *rank*: Passage rank.

Table 3
Task 2 output example

run_id	manual	topic_id	passage_text	term	rank
ST_1	1	1	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.	machine learning	1
ST_1	1	1	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.	societal biases	2
ST_1	1	1	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.	ML	3

4.2. Task 2: Searching for background knowledge

The goal of this pilot task is to decide which terms (up to 10) require explanation and contextualization to help a reader to understand a complex scientific text - for example, with regard to a query, terms that need to be contextualized (with a definition, example and/or use-case).

Output: List of terms to be contextualized in a tabulated file TSV with the following fields:

- *run_id*: Run ID starting with *team_id_*;
- *manual*: Whether the run is manual 0,1;
- *topic_id*: Topic ID;
- *passage_text*: Passage text;
- *term*: Term or other phrase to be explained;
- *rank*: Importance of the explanation for a given term.

Term pooling and automatic metrics (NDCG,...) will be used to evaluate these results.

Table 4
Task 3 output example

run_id	manual	topic_id	doc_id	source_passage	simplified_passage
ST_1	1	1	3003409254	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.	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.

4.3. Task 3: Scientific text simplification

The goal of this pilot task is to provide a simplified version of text passages. Participants will be provided with queries and abstracts of scientific papers. The abstracts can be split into sentences as in the exemple. The simplified passages will be evaluated manually with eventual use of aggregating metrics.

Output: Simplified passages in a TSV tabulated file with the following fields:

- *run_id*: Run ID starting with *team_id_*;
- *manual*: Whether the run is manual 0,1;
- *topic_id*: Topic ID;
- *doc_id*: Source document ID;
- *source_passage*: Source passage text;
- *simplified_passage*: Text of the simplified passage.

5. Program overview

43 teams were registered for the SimpleText workshop. However, participants did not submit their runs on our pilot tasks.

SimpleText will host three invited talks:

- *Importance of Data and Controllability in Neural Text Simplification information on submission* by Wei Xu;
- *What if EVERYONE could understand COVID-19 information? EasyCOVID-19 project* by John Rochford;
- *Evaluation of simplification of scientific texts* by Natalia Grabar.

Wei Xu will demonstrate in her talk that creating high-quality training data and injecting linguistic knowledge can lead to significant performance improvements that overshadow gains from many of these model variants. She will present her two recent works on text simplification, both on lexical and syntactic level: 1) a neural conditional random field (CRF) based semantic model to create parallel training data; 2) a controllable text generation approach that incorporates syntax through pairwise ranking and data augmentation.

John Rochford will present the EasyCOVID-19[8] project which aims to simplify textual information from every world's government websites.

Grabar extensively worked on technical and simplified medical texts in French [50, 51] as well as text transformation topology during simplification [52] .

Mike Unwalla will give an industrial talk about TermChecker[9], a solution to check a document for compliance to the ASD-STE100 Simplified Technical English specification.

Sílvia Araújo and Radia Hannachi will present their work on *Multimodal science communication: from documentary research to infographic via mind mapping*. They conducted a pedagogical experiment in a university context. The goal of this experiment was to introduce students to active methodologies through a pedagogical approach in three stages. The students were required to read and understand (scientific) texts in order to extract important information and organise it in a new visual communication format using digital tools.

An overview of the SimpleText workshop at the French conference INFORSID-2021 will be presented by Liana Ermakova, Josiane Mothe and Eric Sanjuan.

We will also discuss *What Science-Related Topics Need to Be Popularized?* via a comparative study.

Malek Hajjem and Eric Sanjuan will present their work on *Societal trendy multi word term extraction from DBLP*. In their experiments, they focus on scientific terms in news articles using the SimpleText corpus and a collection of french political parties press releases. They found that the overlap between journalistic articles and scientific publications is higher than expected. They studied how ongoing scientific research is impacted by ongoing political debate.

The second part of the workshop will be interactive. We are soliciting position statements on opportunities, problems, and solutions in text simplification and its evaluation.

6. Conclusion

The paper introduced the CLEF 2021 SimpleText track, consisting of a workshop and pilot tasks on text simplification for scientific information access. As SimpleText is in the intersection of computer science (namely AI, IR and NLP) and linguistics, the collaboration of the researchers from these domains is necessary. The SimpleText workshop relies on an interdisciplinary community of researchers in automatic language processing, information retrieval, linguistics, sociology, science journalism and science popularization working together to try to solve one of the biggest challenges of today. This diversity is reflected by the previewed presentations.

[8] https://easycovid19.org/
[9] https://www.techscribe.co.uk/

Acknowledgments

We would like to thanks the MaDICS[10] (Masses de Données, Informations et Connaissances en Sciences - Big data, Information and Knowledge in Sciences) research group, Elise Mathurin, Alain Kerhervé and Nicolas Poinsu.

References

[1] G. Leroy, J. E. Endicott, D. Kauchak, O. Mouradi, M. Just, User evaluation of the effects of a text simplification algorithm using term familiarity on perception, understanding, learning, and information retention, Journal of medical Internet research 15 (2013) e144.

[2] B. Fecher, S. Friesike, Open science: one term, five schools of thought, in: Opening science, Springer, Cham, 2014, pp. 17–47.

[3] P. G. Altbach, H. d. Wit, Too much academic research is being published, 2018. URL: https://www.universityworldnews.com/post.php?story=20180905095203579.

[4] "2019-nCoV" OR ... Publication Year: 2020 in Publications - Dimensions, ???? URL: https://covid-19.dimensions.ai/.

[5] H. P. Luhn, The automatic creation of literature abstracts, IBM Journal of research and development 2 (1958) 159–165.

[6] R. Nallapati, B. Zhou, C. dos Santos, Ç. GuÌ‡lçehre, B. Xiang, Abstractive text summarization using sequence-to-sequence RNNs and beyond, in: Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning, Association for Computational Linguistics, Berlin, Germany, 2016, pp. 280–290. URL: https://www.aclweb.org/anthology/K16-1028. doi:10.18653/v1/K16-1028.

[7] M. Völske, M. Potthast, S. Syed, B. Stein, Tl; dr: Mining reddit to learn automatic summarization, in: Proceedings of the Workshop on New Frontiers in Summarization, 2017, pp. 59–63.

[8] K. Filippova, Y. Altun, Overcoming the lack of parallel data in sentence compression, in: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 2013, pp. 1481–1491.

[9] E. Beaunoyer, M. Arsenault, A. M. Lomanowska, M. J. Guitton, Understanding online health information: Evaluation, tools, and strategies, Patient Education and Counseling 100 (2017) 183–189. URL: https://www.sciencedirect.com/science/article/pii/S0738399116304025. doi:https://doi.org/10.1016/j.pec.2016.08.028.

[10] P. B. Jarreau, L. Porter, Science in the Social Media Age: Profiles of Science Blog Readers, Journalism & Mass Communication Quarterly 95 (2018) 142–168. URL: https://doi.org/10.1177/1077699016685558. doi:10.1177/1077699016685558, publisher: SAGE Publications Inc.

[11] K. Molek-Kozakowska, Communicating environmental science beyond academia: Stylistic patterns of newsworthiness in popular science journalism, Discourse & Communication 11 (2017) 69–88. URL: https://doi.org/10.1177/1750481316683294. doi:10.1177/1750481316683294.

[10] https://www.madics.fr/ateliers/simpletext/

[12] background knowledge, ???? URL: https://www.thefreedictionary.com/background+knowledge.
[13] T. O'Reilly, Z. Wang, J. Sabatini, How Much Knowledge Is Too Little? When a Lack of Knowledge Becomes a Barrier to Comprehension:, Psychological Science (2019). URL: https://journals.sagepub.com/doi/10.1177/0956797619862276. doi:10.1177/0956797619862276, publisher: SAGE PublicationsSage CA: Los Angeles, CA.
[14] H. Shi, S. Revithis, S.-S. Chen, An agent enabling personalized learning in e-learning environments, in: Proceedings of the first international joint conference on Autonomous agents and multiagent systems: part 2, AAMAS '02, Association for Computing Machinery, New York, NY, USA, 2002, pp. 847–848. URL: https://doi.org/10.1145/544862.544941. doi:10.1145/544862.544941.
[15] J. Hoffart, M. A. Yosef, I. Bordino, H. Fürstenau, M. Pinkal, M. Spaniol, B. Taneva, S. Thater, G. Weikum, Robust disambiguation of named entities in text, in: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, 2011, pp. 782–792.
[16] I. Augenstein, M. Das, S. Riedel, L. Vikraman, A. McCallum, Semeval 2017 task 10: Scienceie-extracting keyphrases and relations from scientific publications, arXiv preprint arXiv:1704.02853 (2017).
[17] N. Fuhr, A. Giachanou, G. Grefenstette, I. Gurevych, A. Hanselowski, K. Jarvelin, R. Jones, Y. Liu, J. Mothe, W. Nejdl, et al., An information nutritional label for online documents, in: ACM SIGIR Forum, volume 51, ACM New York, NY, USA, 2018, pp. 46–66.
[18] B. L. Zakaluk, S. J. Samuels, Readability: Its Past, Present, and Future, International Reading Association, 800 Barksdale Rd, 1988. URL: https://eric.ed.gov/?id=ED292058.
[19] E. Fry, The Varied Uses of Readability Measurement (1986).
[20] R. Flesch, A new readability yardstick., Journal of Applied Psychology 32 (1948) p221 – 233.
[21] L. Si, J. Callan, A statistical model for scientific readability, in: Proceedings of the Tenth International Conference on Information and Knowledge Management, CIKM '01, ACM, New York, NY, USA, 2001, pp. 574–576. URL: http://doi.acm.org/10.1145/502585.502695. doi:10.1145/502585.502695.
[22] E. Pitler, A. Nenkova, Revisiting readability: A unified framework for predicting text quality, 2008.
[23] L. Ermakova, J. V. Cossu, J. Mothe, A survey on evaluation of summarization methods, Information Processing & Management 56 (2019) 1794–1814. URL: http://www.sciencedirect.com/science/article/pii/S0306457318306241. doi:10.1016/j.ipm.2019.04.001.
[24] K. Collins-Thompson, J. Callan, A Language Modeling Approach to Predicting Reading Difficulty, Proceedings of HLT/NAACL 4 (2004).
[25] M. Heilman, K. Collins-Thompson, M. Eskenazi, An analysis of statistical models and features for reading difficulty prediction, in: Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications, EANL '08, Association for Computational Linguistics, Stroudsburg, PA, USA, 2008, pp. 71–79. URL: http://dl.acm.org/citation.cfm?id=1631836.1631845.
[26] L. Feng, M. Jansche, M. Huenerfauth, N. Elhadad, A comparison of features for automatic readability assessment, in: Proceedings of the 23rd International Conference on Computational Linguistics: Posters, COLING '10, Association for Computational Linguistics,

Stroudsburg, PA, USA, 2010, pp. 276–284. URL: http://dl.acm.org/citation.cfm?id=1944566.1944598.
[27] S. E. Petersen, M. Ostendorf, A machine learning approach to reading level assessment, Comput. Speech Lang. 23 (2009) 89–106. URL: http://dx.doi.org/10.1016/j.csl.2008.04.003. doi:10.1016/j.csl.2008.04.003.
[28] A. J. Stenner, I. Horablin, D. R. Smith, M. Smith, The Lexile Framework. Durham, NC: Metametrics (1988).
[29] J. S. Chall, E. Dale, Readability revisited: The new Dale–Chall readability, MA: Brookline Books, Cambridge, 1995.
[30] E. Fry, A readability formula for short passages, Journal of Reading 8 (1990) 594–597. 33.
[31] J. Tavernier, P. Bellot, Combining relevance and readability for INEX 2011 question–answering track (2011) 185–195.
[32] A. Mutton, M. Dras, S. Wan, R. Dale, Gleu: Automatic evaluation of sentence-level fluency, ACL–07 (2007) 344–351.
[33] S. Wan, R. Dale, M. Dras, Searching for grammaticality: Propagating dependencies in the viterbi algorithm, Proc. of the Tenth European Workshop on Natural Language Generation (2005).
[34] S. Zwarts, M. Dras, Choosing the right translation: A syntactically informed classification approach, Proc. of the 22nd International Conference on Computational Linguistics (2008) 1153–1160.
[35] J. Chae, A. Nenkova, Predicting the fluency of text with shallow structural features: case studies of machine translation and human–written text, Proc. of the 12th Conference of the European Chapter of the ACL (2009) 139–147.
[36] A. Stenner, I. Horabin, D. R. Smith, M. Smith, The lexile framework, Durham, NC: MetaMetrics (1988).
[37] P. Bellot, A. Doucet, S. Geva, S. Gurajada, J. Kamps, G. Kazai, M. Koolen, A. Mishra, V. Moriceau, J. Mothe, M. Preminger, E. SanJuan, R. Schenkel, X. Tannier, M. Theobald, M. Trappett, Q. Wang, Overview of INEX, in: Information Access Evaluation. Multilinguality, Multimodality, and Visualization - 4th International Conference of the CLEF Initiative, CLEF 2013, Valencia, Spain, September 23-26, 2013. Proceedings, 2013, pp. 269–281.
[38] R. Barzilay, N. Elhadad, K. R. McKeown, Inferring strategies for sentence ordering in multidocument news summarization, Journal of Artificial Intelligence Research (2002) 35–55. 17.
[39] L. Ermakova, J. Mothe, A. Firsov, A Metric for Sentence Ordering Assessment Based on Topic-Comment Structure, in: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '17, Association for Computing Machinery, New York, NY, USA, 2017, pp. 1061–1064. URL: https://doi.org/10.1145/3077136.3080720. doi:10.1145/3077136.3080720, event-place: Shinjuku, Tokyo, Japan.
[40] W. Coster, D. Kauchak, Simple english wikipedia: a new text simplification task, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2011, pp. 665–669.
[41] J. Ladyman, J. Lambert, K. Wiesner, What is a complex system?, European Journal for Philosophy of Science 3 (2013) 33–67. URL: https://doi.org/10.1007/s13194-012-0056-8.

doi:10.1007/s13194-012-0056-8.
[42] P. M. McCarthy, R. H. Guess, D. S. McNamara, The components of paraphrase evaluations, Behavior Research Methods 41 (2009) 682–690. URL: https://doi.org/10.3758/BRM.41.3.682. doi:10.3758/BRM.41.3.682.
[43] D. Cram, B. Daille, Terminology Extraction with Term Variant Detection, in: Proceedings of ACL-2016 System Demonstrations, Association for Computational Linguistics, Berlin, Germany, 2016, pp. 13–18. URL: https://www.aclweb.org/anthology/P16-4003. doi:10.18653/v1/P16-4003.
[44] S. O. Søe, Algorithmic detection of misinformation and disinformation: Gricean perspectives, Journal of Documentation 74 (2018) 309–332. URL: https://doi.org/10.1108/JD-05-2017-0075. doi:10.1108/JD-05-2017-0075, publisher: Emerald Publishing Limited.
[45] S. Wubben, A. van den Bosch, E. Krahmer, Sentence simplification by monolingual machine translation, in: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2012, pp. 1015–1024.
[46] W. Xu, C. Napoles, E. Pavlick, Q. Chen, C. Callison-Burch, Optimizing statistical machine translation for text simplification, Transactions of the Association for Computational Linguistics 4 (2016) 401–415.
[47] E. Sulem, O. Abend, A. Rappoport, Semantic structural evaluation for text simplification, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), 2018, pp. 685–696.
[48] S. Rao, J. Tetreault, Dear sir or madam, may i introduce the gyafc dataset: Corpus, benchmarks and metrics for formality style transfer, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), 2018, pp. 129–140.
[49] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318.
[50] N. Grabar, R. Cardon, CLEAR-Simple Corpus for Medical French, 2018. URL: https://halshs.archives-ouvertes.fr/halshs-01968355.
[51] R. Cardon, N. Grabar, Détection automatique de phrases parallèles dans un corpus biomédical comparable technique/simplifié, in: TALN 2019, Toulouse, France, 2019. URL: https://hal.archives-ouvertes.fr/hal-02430446.
[52] A. Koptient, N. Grabar, Typologie de transformations dans la simplification de textes, in: Congrès mondial de la linguistique française, Montpellier, France, 2020. URL: https://hal.archives-ouvertes.fr/hal-03095235.