



HAL
open science

Overview of the CLEF 2022 SimpleText Task 2: Complexity Spotting in Scientific Abstracts

Liana Ermakova, Irina Ovchinnikova, Jaap Kamps, Diana Nurbakova, Sílvia Araújo, Radia Hannachi

► To cite this version:

Liana Ermakova, Irina Ovchinnikova, Jaap Kamps, Diana Nurbakova, Sílvia Araújo, et al.. Overview of the CLEF 2022 SimpleText Task 2: Complexity Spotting in Scientific Abstracts. Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022, Sep 2022, Bologna, Italy. hal-03828451

HAL Id: hal-03828451

<https://hal.science/hal-03828451v1>

Submitted on 25 Oct 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

Overview of the CLEF 2022 SimpleText Task 2: Complexity Spotting in Scientific Abstracts

Liana Ermakova¹, Irina Ovchinnikov², Jaap Kamps³, Diana Nurbakova⁴,
Sílvia Araújo⁵ and Radia Hannachi⁶

¹Université de Bretagne Occidentale, HCTI, Brest, France

²ManPower Language Solution, Israel

³University of Amsterdam, Amsterdam, The Netherlands

⁴University of Lyon, INSA Lyon, CNRS, LIRIS, UMR5205, F-69621 Villeurbanne, France

⁵Universidade do Minho, CEHUM, 4710-057 Braga, Portugal

⁶Université de Bretagne Sud, HCTI, 56321 Lorient, France

Abstract

This paper provides an overview of the *Task 2: What is unclear?* of the Automatic Simplification of Scientific Texts (SimpleText) lab, run as part of CLEF 2022. The main aim of the SimpleText lab is to promote a more open scientific information access via automatic text simplification. *Task 2* focuses on complexity spotting within scientific texts (passage). Thus, the goal is to detect the terms/concepts that require specific background knowledge for understanding of the passage and to assess their complexity for non-experts. Overall, four runs from four different teams have been submitted to this task. In this paper, we describe the data collection, the task setup, and the evaluation procedure. We also give a brief overview of the participating approaches.

Keywords

automatic text simplification, terminology, background knowledge, scientific article, science popularization, contextualization, term difficulty

1. Introduction

Nowadays, scientific literature has become more available to every citizen thanks to digitalisation. However, an important barrier preventing citizens to access the objective scientific knowledge from the original sources remains present. One of the key issues here is a high complexity of scientific texts to non-experts due to the lack of required background knowledge, including the comprehension of terminology. Even for native speakers it is hard to understand the terminology beyond their area of expertise. Nevertheless, a basic set of terms the general public acquired thanks to secondary and college education allows them to comprehend popular science publications. *Comprehension of the term* presupposes grasping of the concept it refers to

CLEF 2022: Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy


✉ liana.ermakova@univ-brest.fr (L. Ermakova)

🌐 <https://simpletext-project.com/> (L. Ermakova)

🆔 0000-0002-7598-7474 (L. Ermakova); 0000-0003-1726-3360 (I. Ovchinnikov); 0000-0002-6614-0087 (J. Kamps); 0000-0002-6620-7771 (D. Nurbakova); 0000-0003-4321-4511 (S. Araújo)



© 2022 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)

without any definition. To understand the concept, we need to involve it in a structured system in our semantic memory that can require more knowledge than we had learned.

To help readers to stay up-to-date with scientific advances, text simplification can be used. To facilitate the reading, the traditional methods try to eliminate complex concepts and constructions [1]. However, it is not always possible, especially in the case of scientific literature. Thus, readers of a popular science publication lean on their experience of processing new information and recognize a case when they need definition or clarification of an unfamiliar term since they do not understand its concept.

To alleviate the lack of background knowledge that can prevent a proper comprehension [2], we argue that a simplification method should provide information, essential to understanding of complex scientific concepts. This is one of the objectives of CLEF 2022 SimpleText lab. Despite some recent efforts that have been done in automatic text simplification (e.g. [3]), improving scientific text comprehensibility and its adaptation to different audiences in an automatic manner remains an open challenge.

The CLEF 2022 SimpleText track¹ is an open forum for researchers and practitioners working on the automatic generation of simplified summaries of scientific texts. It is a new evaluation lab that follows up the CLEF 2021 SimpleText Workshop [4]. The track provides data and benchmarks for discussing the challenges of automatic text simplification proposing the following interconnected tasks:

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

Task 2: What is unclear? Given a passage and a query, rank terms/concepts that are required to be explained for understanding this passage (definitions, context, applications,..).

Task 3: Rewrite this! Given a query, simplify passages from scientific abstracts.

This paper focuses on the second task of complexity spotting. We refer for details of the other tasks to the overview papers of Task 1 [5] and Task 3 [6], or the Track overview paper [7].

In the CLEF 2022 edition of SimpleText, a total of 62 teams registered for the SimpleText track. A total of 40 users downloaded data from the server. A total of 9 distinct teams submitted 24 runs, of which 10 runs were updated. The details of statistics on runs submitted for shared tasks are presented in Table 1. As it can be seen, four teams participated in Task 2.

The rest of this paper is structured in the following way. Section 2 presents a brief overview of related works, including other evaluation initiatives, related tasks and related approaches. We provide a detailed description of the task complexity spotting itself, submitted runs, and the evaluation protocol in Section 3. In Section 4, we discuss the results of the official submissions. We end with Section 5 discussing the results and findings, and lessons for the future.

2. Related work

According to the Cambridge Dictionary [16], a *term* is “a word or expression used in relation to a particular subject, often to describe something official or technical”. Almost the same

¹<https://simpletext-project.com>

Table 1
CLEF 2022 SimpleText official run submission statistics

Team	Task 1	Task 2	Task 3	Total runs
aaac		1 (1 updated)		1
CLARA-HD [8]			1	1
CYUT Team2 [9]	1		1	2
HULAT-UC3M [10]			10 (4 updated)	10
LEA_T5 [11]		1	1	2
NLP@IISERB [12]	3 (3 updated)			3
PortLinguE [13]			1 (1 updated)	1
SimpleScientificText [14]		1 (1 updated)		1
UAMS [15]	2	1		3
<i>Total runs</i>	6	4	14	24

definition of terms is given by Kaguera and Marshman [17] describing them as “lexical items that represent concepts of a domain”. Thus, terms form the core vocabulary of a specific and specialised domain.

2.1. Term Complexity

Term perception can be rather ambiguous and subjective [18], especially when it comes to assess term complexity. Indeed, the discrepancy between basic competence of a reader and professional competence of an author of a scientific article derives *the subjective complexity of terminology*. *The objective complexity of terminology* is derived by peculiar characteristics of terminological systems. In this Section, we clarify the objective complexity of terminology caused by complexity of research areas, research traditions and socio-cultural diversity.

Terminology belongs to professional and scientific discourse, where there exist so called *languages for special purpose*. Belonging to the language for special purposes, terminological systems do not share peculiarities of the general lexicon [19]. A terminological system tends to avoid synonyms and polysemy, but has to provide a term for each concept within a system of concepts of the domain. According to the General Theory of Terminology, which is based on the work of Eugen Wüster (see description in [20]), terminological systems support *univocity* (unambiguous match of the term to its concept). This general approach is still relevant in technical communication where professionals (technical writers, translators, etc.) use term banks, e.g. Eurodicautom², Termium³, LEXIS⁴ [21], Normaterm⁵ [22], and the Grand dictionnaire terminologique⁶ (formerly the Banque de terminologie du Québec). In academia, this approach is mostly applied to terminological systems in Science and Computer Science; however, it is not

²A database for terminology and translations created and used by the European Commission, replaced in 2007 by Interactive Terminology for Europe (IATE) <https://iate.europa.eu/>.

³A linguistic and terminology database owned by the Translation Bureau of Public Services and Procurement Canada, <https://www.btb.termiumplus.gc.ca/>

⁴A German term bank used by technical translators.

⁵French term bank covering science and technology fields and developed by AFNOR.

⁶A term bank created by the Quebec Board of the French Language, <https://gdt.oqlf.gouv.qc.ca/>

relevant for Cognitive Science (e.g., Neuroscience) and Humanities.

Complexity of a terminological system is a **derivative of scientific complexity**. The complexity of a scientific area depends on peculiar attributes and conditions [23]. The most basic peculiarities are the *numerosity* of counting entities and their *interaction*: high diversity of disordered interaction among multiple entities represents a complex research area. To refer to the entities, their interactions and degrees of disorder, the research area needs complex terminology. Ladyman et al. [24] offered to determine complexity of a research area according to five qualitative conditions: numerosity of elements, numerosity of interactions, disorder, openness, feedback. Considering terminological systems, *numerosity of elements* and *numerosity of interactions* in a complex research area require a rich and clear structured system of terms, preferably taxonomy. Transparency of the terminological system structure facilitates the research, analysis and description of disordered systems and non-equilibrium states of the systems. Effect of numerosity of elements and their interactions on the complexity of the terminological system of the research area is obvious through comparison of different areas that attract interest of wide readership: Neuroscience and Computer Science [25].

The complexity of terminology is associated with a **formal representation (signifier) of a term**. Putting aside borrowings, we would like to mention symbols and abbreviations (acronyms, backronyms, syllabic abbreviations, clipping etc.). Symbols and abbreviations belong to a set of peculiarities of a language for special purpose. Symbolic language of science involves symbols and abbreviations as means to optimize content transferring, to standardize naming of numerous elements, frequent interaction among them, and standard procedures of data processing. Languages for special purpose in Natural Science and Mathematical Sciences (including Computer Science) contain complicated systems of symbols. Meanwhile, symbols and abbreviations are in use in all research areas disregarding their complexity. Nevertheless, readers of popularized publications expect explanations of the symbols and abbreviations.

Another cause of the terminological complexity is **research traditions**. Neuroscience and computer science represent the new research areas. Nevertheless, humans became curious about the brain and how to treat its damage thousands years ago; the brain has attracted researchers' attention since the very first steps in practical medicine. The neuroscientific terminology reflects rich traditions of the brain study in the history of science: Latin (e.g. *cerebellum* 'little brain') and Greek (e.g. *diencephalon* 'interbrain') borrowings, eponyms (*Broca's area*), metaphors (e.g. *hemispheres*), etc. Diversity of the traditions provides neuroscience with parallel terms, which refer to the same concept (e.g., names of the disease: [26]). Understanding the neuroscientific terminology requires knowledge of the science development.

Computer science has begun to develop its traditions mostly in the middle of the XX century; therefore, it lacks Latin and Greek terminology as well as numerous eponyms. As compared to neuroscience, the terminology in computer science seems less complicated and more transparent for nonprofessionals; moreover, an average reader of popularized science understands many terms since he / she employs computers in the everyday routine. Readership of popular science publications is probably familiar with the basic terminology of this area, while the neuroscientific terminology requires definitions and clarifications.

The complexity of terminology is often caused by **socio-cultural diversity** of readership of popular science publications. The diversity is revealed in comprehension of basic terminology of Science and Humanities that is affected by programs of secondary and college education.

The programs provide people with grounds and backdrops for comprehending current news of popular science. Since content of the programs varies in different institutions and countries, readers have differences in their background and terminological lexicon especially in Humanities.

While popularizing science, journalists **substitute complex terms** by basic ones or clarify the underlying concept, which is denoted by the complex term. Enhancing the popular science text readability, popularization may bring in damaging its comprehensibility. Both ways to avoid the complex terminology may lead to misinformation or distortion of the content. The term substitution may distort the content since semantic relations in terminological systems are not similar to those in the general lexicon of the language. It is presupposed that a network of connections within a terminological system does not support synonyms and maintains a transparent one-to-one relationship between the term and the concept it referred to. A list of the potential substitutions usually includes a widespread name of the concept if any exists in the general lexicon (e.g. *sea cow* instead of *manatee*), hypernym (e.g. *herbivore marine mammal* for *manatee*) and co-hyponyms of the complex term with additional explanation since co-hyponyms denote a different object (quality, action, etc.) within the same category. Meanwhile, common-sense concepts are not equal to scientific concepts in the complex research areas; therefore, appealing to the common sense requires clarifications. Thus, term substitutions do not enhance structure of the popular scientific text. Probably, the best way to clarify the term is to illustrate its concept [27].

Speaking about automatic systems of generating a popular review of scientific publications, we need to choose the way for term recognition and extraction. In order to substitute or clarify any unfamiliar term we need to recognize it in scientific discourse and then provide readers with references, definitions or illustrations.

Summarizing our consideration of complexity of terminology, we note that the selection of a way to facilitate perception of terms in popular scientific publications depends on complexity of the research area, richness of the research tradition of the area, and cultural diversity.

2.2. Automatic Terminology Extraction

Automatic Term Extraction (ATE) or *Automatic Terminology Extraction* is an automated process of detecting terms in a corpus of specialised texts. It has been a relevant NLP task since 1980s and remains challenging from several perspectives, such as data collection (creation of manually annotated domain-specific corpora), extraction algorithms (definition of term length, minimum term frequency, term POS-pattern), evaluation (usually limited to the use of precision metric as the information about all terms in a text is often missing) [18].

The ATE methods are traditionally classified in three groups:

- *Linguistic methods*: these methods are based on linguistic properties such as POS-patterns or other morpho-syntactic patterns (e.g. [28, 29]).
- *Statistical methods*: these methods are based on statistical properties (various weightings have been proposed, e.g. frequency, mutual information, log-likelihood ratio, etc.) and usually analyse n -grams measuring termhood or unithood [30].
- *Hybrid methods*: these methods are combinations of the previous two (e.g. [31]). Usually, the initial selection is performed based on linguistic properties which is followed by the

ranking procedure on the basis of statistical measures [18]. Hybrid approaches have been shown to outperform linguistic or statistical methods [32].

As stated in [18], one of the difficulties is to well define the cut-off threshold for term candidates.

Recent advances in Machine Learning techniques, including Deep Learning models, have made the taxonomy of ATE methodology more complex and diverse [33]. Numerous methods have been proposed (e.g. [34, 35]).

Lately, large transformer models such as Jurassic-1 [36], Google's T5 [37], BERT [38], or GPT-3 [39] have been shown to be successful on several NLP tasks, outperforming other state-of-the-art models. They make use of subword tokenizers, such as Byte-Pair Encoding (BPE) [40] and WordPiece [41]. For instance, BPE that uses the idea of word segmentation into subword units is exploited in GPT-2 [42] and Roberta [43]. A similar subword tokenization algorithm WordPiece is used in BERT [38], DistilBERT [44], and Electra [45]. Despite a comparative shallowness of these models, they have been shown to be quite effective for the related use case of languages with large vocabularies and many rare words [46, 40]. Therefore, their use might be promising for terminology extraction.

In the context of term extraction from scientific texts with the final goal of text simplification, it is also important to consider named entities. Named entities are objects, abstract or physical, such as a person, location, organization, product, etc., that can be denoted with a proper name. They can also designate certain natural terms like biological species, substances [47]. For a recent survey of existing deep learning techniques for Named Entity Recognition (NER) task, refer to [48].

2.3. Related Evaluation Initiatives

This section presents a brief overview of related evaluation initiatives, related tasks and related approaches.

CLEF SimpleText track was first accepted in 2020 (see [49] for the overview of the first edition of CLEF SimpleText workshop). However, there have been other initiatives addressing the related topics on scholarly document processing at NLP conference.

The lack of background knowledge can become a barrier to reading comprehension and there is a knowledge threshold allowing reading comprehension [2]. Scientific text simplification presupposes the facilitation of readers' understanding of complex content by establishing links to basic lexicon while traditional methods of text simplification try to eliminate complex concepts and constructions [1]. SimpleText is not limited to a "Split and Rephrase" task [50] but also aims to provide a sufficient context to a scientific text. Entity linking could mitigate the background knowledge problem, by providing definitions, illustrations, examples, and related entities, but the existing entity linking datasets are focused on people, places, and organisation [51], while a non-expert reader of a scientific article needs assistance with new concepts and methods. INEX/CLEF'11-14 Tweet Contextualization [52] and CLEF'16-17 Cultural Microblog Contextualization [53] tracks aim to provide lacking background knowledge to a tweet. Besides completely different nature of tweets and popular science, this use case differs from the text simplification as this lack of background knowledge is due to the tweet length.

In contrast to the Background Linking task at TREC’20 News Track [54], SimpleText focuses on (1) scientific text; (2) selection of notions to be explained; (3) helpfulness of the provided information rather than its relevance.

Probably, the closest evaluation campaign to SimpleText’s task 2 is *TermEval 2020: Shared Task on Automatic Term Extraction Using Annotated Corpora for Term Extraction Research (ACTER) Dataset* [18]. One of the challenges related to term extraction methodology is stated to be the definition of the degree of specialisation or domain-specification required for a lexical item to be considered a term. This aspect which is difficult to quantify is partially tackled under “*term difficulty*” goal of the *task 2* of the CLEF SimpleText lab. TermEval was set up as a binary task: term or not. In contrast to that, SimpleText aims at detecting a term and identifying its difficulty level.

Datasets Simple Wikipedia based datasets could be useful to train AI models but (1) they are not scientific publications; (2) there is no direct correspondence between Wikipedia and Simple Wikipedia articles [55]. Another dataset was introduced at TAC 2014 Biomedical Summarization Track [56] with a goal to retrieve important aspects of a paper from the perspective of the community. In TermEval task [18], the organisers proposed ACTER, a manually annotated domain-specific corpora covering 3 languages (English, French, and Dutch) and four domains (corruption, dressage (equitation), heart failure, and wind energy). The annotators labelled around 50k tokens for each language and domain. The tokens were judged according to their degree of domain-specificity and lexicon-specificity. Three term labels were used: Specific Terms (i.e. domain- and lexicon-specific), Common Terms (domain-specific, not lexicon-specific), and Out-of-Domain (OOD) Terms (not domain-specific, lexicon-specific). In SimpleText, we focus on term difficulty which is in line with lexicon-specificity of TermEval task (in particular, when using 3-point scale), without assessing domain-specificity.

In contrast to that, we evaluate simplification in terms of lexical and syntax complexity combining with error analysis. As we demonstrated previously, scientific information is often distorted accidentally due to misunderstanding of terminology, omission of essential details, insertion of erroneous background etc. [55]. Information distortion analysis is close to scientific claim verification [57, 58] but fact checking is limited to search for relevant evidence and decide whether it supports the claim. Another close work is [59], where the TF-IDF cosine similarity between documents is computed on (1) a collection of abstracts of scientific papers from the Citation Network Dataset V1 AMINER [60] and (2) a set of articles from Huffington Post. However, this approach is not robust to lexical changes, which are crucial for text simplification. To the best of our knowledge, no other automatic nor semi-automatic method for information distortion analysis exists.

3. CLEF 2022 SimpleText Task 2 Test Collection

In this section, we discuss the second task about complexity spotting in an extracted sentence from a scientific abstract, addressing the task:

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

The goal of this task is to decide which terms (up to 5) 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). For each passage, participants should provide a ranked list of difficult terms with corresponding scores on the scale 1-3 (3 to be the most difficult terms, while the meaning of terms scored 1 can be derived or guessed) and on the scale 1-5 (5 to be the most difficult terms). Passages (sentences) are considered to be independent, i.e. difficult term repetition was allowed.

3.1. Train Data

For this task, data is two-fold: *Medicine* and *Computer Science*, as these two domains are the most popular on forums like ELI5 [25, 61]. As in 2021, for *Computer Science*, we use scientific abstracts from the Citation Network Dataset: DBLP+Citation, ACM Citation network (12th version)⁷ [49]. A master student in Technical Writing and Translation manually annotated each sentence by extracting difficult terms and attributing difficulty scores on a scale of 1-3 (3 to be the most difficult terms, while the meaning of terms scored 1 can be derived or guessed) and on a scale of 1-5 (5 to be the most difficult terms).

In 2022, we introduced new data based on Google Scholar and PubMed articles on muscle hypertrophy and health annotated by a master student in Technical Writing and Translation, specializing in these domains. The selected abstracts included the objectives of the study, the results and sometimes the methodology. The abstracts including only the topic of the study were excluded because of the lack of information. To avoid the curse of knowledge, another master student in Technical Writing and Translation not familiar with the domain was solicited for complexity spotting.

We provided 453 annotated examples in total.

3.2. Test Data

To construct the test data, we retrieved 116,763 sentences from the DBLP abstracts according to the queries from Task 1. We then manually evaluated 592 distinct sentences for 11 queries. For the query *Digital assistant* we took the first 1,000 sentences retrieved by ElasticSearch. We pool terms submitted by all participants for all these queries, representing a number of 4,167 distinct pairs *sentence-term* in total. We ensured that for each evaluated source sentence the pool contained the results of all participants. Statistics of the number of evaluated sentences per query for Task 2 are given in Table 2.

3.3. Input and Output Formats

The input for the train and the test data was provided in JSON and CSV formats with the following fields:

snt_id a unique passage (sentence) identifier.

source_snt passage text.

⁷<https://www.aminer.org/citation>

Table 2

SimpleText Task 2: Statistics of the number of evaluated sentences per query

Query	# Sentences	# Sentence-term pairs
1 <i>guessing attack</i>	60	389
2 <i>end to end encryption</i>	55	390
3 <i>imbalanced data</i>	55	381
4 <i>distributed attack</i>	54	385
5 <i>genetic algorithm</i>	51	374
6 <i>quantum computing</i>	51	385
7 <i>qbit</i>	50	363
8 <i>side-channel attack</i>	49	340
9 <i>traffic optimization</i>	47	344
10 <i>quantum applications</i>	42	320
11 <i>cyber-security</i>	35	244
12 <i>conspiracy theories</i>	23	180
13 <i>crowdsourcing</i>	15	104
14 <i>digital assistant</i>	5	32

doc_id a unique source document identifier.

query_id a query ID.

query_text difficult terms should be extracted from sentences with regard to this query.

Input example (JSON format):

```
{ "snt_id": "G06.2_2548923997_3", "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.", "doc_id": 2548923997, "query_id": "G06.2", "query_text": "self driving" }
```

Participants had to submit a list of terms to be contextualized in a JSON format or a tabulated file TSV (for manual runs) with the following fields:

run_id Run ID starting with (team_id)_(task_id)_(name).

manual Whether the run is manual {0, 1}.

snt_id a unique passage (sentence) identifier from the input file.

term Term or other phrase to be explained.

term_rank_snt term difficulty rank within the given sentence.

score_5 term difficulty score on the scale from 1 to 5 (5 to be the most difficult terms).

score_3 term difficulty score on the scale from 1 to 3 (3 to be the most difficult terms).

Output example (JSON format):

```

{"run_id": "NP_task_2_run1", "manual": 1, "snt_id": "G06.2_2548923997_3", "term": "black
↪ hole attack", "term_rank_snt": 1, "score_5": 5, "score_3": 3},
{"run_id": "NP_task_2_run1", "manual": 1, "snt_id": "G06.2_2548923997_3", "term": "grey
↪ hole attack", "term_rank_snt": 2, "score_5": 5, "score_3": 3},
{"run_id": "NP_task_2_run1", "manual": 1, "snt_id": "G06.2_2548923997_3", "term": "Sybil
↪ attack", "term_rank_snt": 3, "score_5": 5, "score_3": 3},
{"run_id": "NP_task_2_run1", "manual": 1, "snt_id": "G06.2_2548923997_3",
↪ "term": "wormhole attack", "term_rank_snt": 4, "score_5": 5, "score_3": 3},
{"run_id": "NP_task_2_run1", "manual": 1, "snt_id": "G06.2_2548923997_3", "term": "Denial
↪ of service attack", "term_rank_snt": 5, "score_5": 4, "score_3": 3}

```

3.4. Evaluation metrics

We evaluated terms according to:

- correctness of term limits;
- term difficulty score on the scale 1-3;
- term difficulty score on the scale 1-5.

For both scales of term difficulty, we used a converted scale 1-7. This scale 1-7 was chosen following the psycho-linguistic research of the perception and evaluation of lexical meanings performed by Osgood and his colleagues [62], in contrast to the psychometric Likert scale (1-5, Strongly disagree/Disagree/Neither agree nor disagree/Agree/Strongly agree), commonly used in the research that employs questionnaires [63]. In the classical version of the semantic differential technique, the scale shows the variety of the human perception of semantic nuances from negative (-3) to positive (+3) polarity where 0 marks the “norm” [62]. The scale 1-7 matches the Osgood’s scale and seems more suitable to evaluate concepts and features avoiding associations with negative / positive assessment. Since the 1970s, the scale has been employed in various studies as an evaluation tool for qualitative features.

Table 3 provides examples of the used term difficulty scale. We separate the examples of abbreviations from non-abbreviated phrases / words.

We added 0 for terms that should not be explained at all and we converted the original scale 1-7 as presented in Table 5.

Table 6 provides some examples of the annotation for Task 2. *TERM* refers to the terms retrieved by participants, *Correct limits* is a binary category showing whether the retrieved terms is well limited, *Corrected* is an eventual correction of retrieved term limits, *Difficulty* is a term difficulty score in scale 1-7.

4. SimpleText Task 2 Results

In this section we discuss the results for the official submissions to the Task 2.

4.1. Participant Approaches

A total of 4 teams submitted runs, of which 2 runs were updated.

Table 3

Examples of the term difficulty scale used for evaluation. Difficult terms are highlighted with the green color

Grade	Non-abbreviated (ordinary) term	Abbreviation
7	<p>“The qubit–qutrit pair acts as a closed system and one <i>external qubit</i> serve as the environment for the pair.”</p>	<p><i>XCSFHP</i> in “We compared <i>XCSFHP</i> to XCSF on several problems.”</p> <p>“The effect of alphabet cardinality and the selection pressure on the scalability of the real-coded ECGA (<i>rECGA</i>) method is investigated.”</p> <p>“We here study the protection of quantum Fisher information (<i>QFI</i>) of the phase parameter in entangled-atom states within the framework of independently dissipative environments and driven individually by classical fields.”</p>
6	<p>“This paper bring forward based on immune genetic algorithm to solve <i>man on board automated storage and retrieval system</i> optimized problem, immune genetic algorithm remains the characteristic which is not ...”</p> <p>“<i>Tile coding</i> is a well-known function approximator that has been successfully applied to many reinforcement learning tasks.”</p> <p>“<i>Quantum circuits</i> of many qubits are challenging to implement making designs with low qubit cost desirable.”</p>	<p>“<i>XCS</i> with computed prediction, namely XCSF, extends XCS by replacing the classifier prediction with a parametrized prediction function.”</p> <p>“Side-channel attack (<i>SCA</i>) is a very efficient cryptanalysis technology to attack cryptographic devices.”</p>
5	<p>“Experiment simulation result express: the result of <i>immune genetic algorithm</i> is better than traditional genetic algorithm in the circumstance of the same clusters and the same evolution generation.”</p> <p>“The results show that the population size required by rECGA-to successfully solve a class of <i>additively- separable problems</i> -scales sub-quadratically with problem size and the number of function evaluations scales sub-cubically with problem size.”</p>	<p>“This paper presents a simple real-coded estimation of distribution algorithm (EDA) design using x-ary extended compact genetic algorithm (<i>XECGA</i>) and discretization methods.”</p>
4	<p>“Specifically, the real-valued decision variables are mapped to discrete symbols of user-specified cardinality using <i>discretization methods</i>.”</p> <p>“Immune genetic algorithm can shorten storage or retrieval distance in application, and enhance storage or <i>retrieval efficiency</i> .”</p> <p>“The effect of alphabet cardinality and the selection pressure on the <i>scalability</i> of the real-coded ECGA (rECGA) method is investigated.”</p> <p>“<i>Deep learning</i> has become increasingly popular in both academic and industrial areas in the past years.”</p>	<p>“This paper presents a simple real-coded estimation of distribution algorithm (<i>EDA</i>) design using x-ary extended compact genetic algorithm (XECGA) and discretization methods.”</p>

Table 4

Examples of the term difficulty scale used for evaluation: grades 0-3. Difficult terms are highlighted with the green color

Grade	Non-abbreviated (ordinary) term	Abbreviation
3	<p>“The XECGA is then used to build the probabilistic model and to sample a new population based on the <i>probabilistic model</i>.”</p> <p><i>scale sub-quadratically</i> in “The results show that the population size required by rECGA-to successfully solve a class of additively- separable problems-<i>scales sub-quadratically</i> with problem size and the number of function evaluations scales sub-cubically with problem size.”</p> <p>“<i>Molecular transistors</i> can play a very important role in the design and fabrication of complex logic functions inside chips.”</p>	<p>“We evaluate each measure’s performance by <i>AUC</i> which is usually used for evaluation of imbalanced data classification.”</p> <p>“This theoretical analysis is confirmed by the experimental results: using several sampling methods to rebalance the imbalanced data sets, it is found that the performances of <i>LDA</i> on balanced data sets are superior to those of LDA on imbalanced data sets.”</p>
2	<p>“Experiment simulation result express: the result of immune genetic algorithm is better than traditional genetic algorithm in the circumstance of the same <i>clusters</i> and the same evolution generation.”</p> <p>“Specifically, the real-valued <i>decision variables</i> are mapped to discrete symbols of user-specified cardinality using discretization methods.”</p>	<p><i>NIST</i> (The National Institute of Standards and Technology) in “Recently <i>NIST</i> has published the second draft document of recommendation for the entropy sources used for random bit generation.”</p>
1	<p>“video labeling game is a <i>crowdsourcing</i> tool to collect user-generated metadata for video clips.”</p> <p>“On the other hand, a 3dimensional (3D) map, which is one of major themes in machine vision research, has been utilized as a simulation tool in city and <i>landscape planning</i> , and other engineering fields.”</p>	<p><i>2D</i> (2-dimensional), <i>3D</i> (3-dimensional) <i>maps</i> as in “The <i>3D maps</i> will give more intuitive information compared to conventional 2-dimensional (<i>2D</i>) ones.”</p>
0	<p>“This <i>device</i> has two work modes: ”native” and ”remote”.”</p> <p>“Immune genetic algorithm can <i>shorten</i> storage or retrieval distance in application, and enhance storage or retrieval efficiency.”</p> <p>“The proposed rECGA is <i>simple</i> , making it amenable for further empirical and theoretical analysis.”</p>	<p><i>et al.</i> (from latin “<i>et alii</i>” meaning “and others”) in “However, Nam <i>et al.</i> pointed out...”</p>

Team **UAm**s from the University of Amsterdam [15] performed the experiments using IDF-based term weighting allowing to locate the most rare terms. Then the obtained rarity measure was balanced with the relevance or centrality of the terms to the given passage.

Team **SimpleScientificText** from Wuhan University [14] used a pipeline of term recognition and complexity spotting, formulating the latter as classification task. The term recognition

Table 5
SimpleText Task 2: Scale conversion rules

Term difficulty scale	0	1	2	3	4	5	6	7
7 point scale	0	1	2	3	4	5	6	7
⇒ 5 point scale	0	1	2	3	4	5	6	7
7 point scale	0	1	2	3	4	5	6	7
⇒ 3 point scale	0	1	2	3	4	5	6	7

Table 6
SimpleText Task 2: Examples of the annotation

Sentence	Term	Limits		Diff- culty
		OK	Corrected	
<i>This device has two work modes: 'native' and 'remote'.</i>	remote	YES		1
<i>This device has two work modes: 'native' and 'remote'.</i>	work modes	YES		0
<i>This device has two work modes: 'native' and 'remote'.</i>	modes native	NO	work modes	0
<i>This device has two work modes: 'native' and 'remote'.</i>	device work	NO	device	0
<i>This device has two work modes: 'native' and 'remote'.</i>	native remote	NO	native	1

was performed in two main steps: term extraction using KeyBERT⁸ followed by filtering based on the similarity of extracted terms with the query calculated with PhraseSimilarity⁹. The model of the evaluation of complexity is built upon three groups of features (lexical, syntactic and semantic) and assembles various state-of-the-art classification models using a soft voting strategy.

Team **LEA_T5** [11] from the University of Western Brittany (UBO) used T5¹⁰ model [64] via the SimpleT5 library¹¹ as the core of their approach. The Google T5 (Text-To-Text Transfer Transformer) model is based on the transfer learning with a unified text-to-text transformer [64].

Team **aaac** has not provided any detail about their run.

4.2. Results

The results are given in Tables 7 and 8. In both tables, we present results for correctly attributed scores regardless the correctness of term limits (*Score_3* and *Score_5*) and the number of correctly limited terms with correctly attributed scores (+ *Limits*). Table 7 provides the results on all sentences we evaluated. However, to have comparable results for partial runs we also report scores on a subset 167 common sentences in Table 8, although we were constrained to exclude the run *lea_t5* due to a very low number of evaluated sentences.

⁸<https://github.com/MaartenGr/KeyBERT>

⁹<https://github.com/franplk/PhraseSimilarity>

¹⁰<https://github.com/google-research/text-to-text-transfer-transformer>

¹¹<https://github.com/Shivanandroy/simpleT5>

Table 7

SimpleText Task 2: Results for the official runs

	Total	Evaluated		Score_3		Score_5	
			+Limits		+Limits		+Limits
aaac	581,285	2,951	1,388	702	318	415	175
SimpleScientificText	63,027	298	262	48	44	47	42
UAms	263,022	1,315	1,175	105	69	60	49
lea_t5	23,331	5	4	0	0	0	0

Table 8

SimpleText Task 2: Results on a subset of 167 common sentences

	Total	Evaluated		Score_3		Score_5	
			+Limits		+Limits		+Limits
aaac	581,285	833	414	200	104	127	67
UAms	263,022	574	514	46	28	25	21
SimpleScientificText	63,027	208	188	33	32	32	29

5. Conclusion and future work

We overviewed Task 2 of the CLEF 2022 SimpleText track that aims at identifying and ranking difficult terms within scientific texts. We evaluated term difficulty with regard to the queries from Task 1. For Task 2, we created a corpus of sentences extracted from the abstracts of scientific publications, with manual annotations of term complexity.

For next year, we will extend Task 2 to provide a context to difficult terms and we will work on automatic metrics based on the insights we obtained this year. In particular, for Task 2, participants will be asked to provide context for difficult terms. This context should provide a definition and take into account ordinary readers' needs to associate their particular problems with the opportunities that science provides them to solve the problems [25]. This year, the HULAT-UC3M [10] team submitted runs which combine tasks 2 and 3 which demonstrates strong interconnection of the tasks as often the terminology cannot be removed nor simplified but it needs to be explained to a reader.

Further details about the lab can be found at the SimpleText website: <http://simpletext-project.com>. Please join us and help to make scientific results understandable!

Acknowledgments

We like to acknowledge the support of the Lab Chairs of CLEF 2022, Allan Hanbury and Martin Potthast, for their help and patience. Special thanks to the University Translation Office of the Université de Bretagne Occidentale, and to Nicolas Poinssu and Ludivine Grégoire for their major impact in the train data construction and Léa Talec-Bernard and Julien Boccou for their help in evaluation of participants' runs. We thank Josiane Mothe for reviewing papers. We also thank Alain Kerhervé, and the MaDICS (<https://www.madics.fr/ateliers/simpletext/> research group.

References

- [1] M. Maddela, W. Xu, A Word-Complexity Lexicon and A Neural Readability Ranking Model for Lexical Simplification, in: Proc. of EMNLP 2018, ACL, Brussels, Belgium, 2018, pp. 3749–3760. URL: <https://www.aclweb.org/anthology/D18-1410>.
- [2] 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>.
- [3] M. Maddela, F. Alva-Manchego, W. Xu, Controllable Text Simplification with Explicit Paraphrasing (2021). URL: <http://arxiv.org/abs/2010.11004>.
- [4] L. Ermakova, P. Bellot, P. Braslavski, J. Kamps, J. Mothe, D. Nurbakova, I. Ovchinnikova, E. Sanjuan, Text Simplification for Scientific Information Access: CLEF 2021 SimpleText Workshop, in: *Advances in Information Retrieval - 43rd European Conference on IR Research, ECIR 2021, Lucca, Italy, March 28 – April 1, 2021, Proc., Lucca, Italy, 2021*.
- [5] E. SanJuan, S. Huet, J. Kamps, L. Ermakova, Overview of the CLEF 2022 SimpleText Task 1: Passage selection for a simplified summary, in: [65], 2022.
- [6] L. Ermakova, I. Ovchinnikova, J. Kamps, D. Nurbakova, S. Araújo, R. Hannachi, Overview of the CLEF 2022 SimpleText Task 3: Query biased simplification of scientific texts, in: [65], 2022.
- [7] L. Ermakova, E. SanJuan, J. Kamps, S. Huet, I. Ovchinnikova, D. Nurbakova, S. Araújo, R. Hannachi, É. Mathurin, P. Bellot, Overview of the CLEF 2022 SimpleText Lab: Automatic simplification of scientific texts, in: A. Barrón-Cedeño, G. D. S. Martino, M. D. Esposti, F. Sebastiani, C. Macdonald, G. Pasi, A. Hanbury, M. Potthast, G. Faggioli, N. Ferro (Eds.), *CLEF'22: Proceedings of the Thirteenth International Conference of the CLEF Association, Lecture Notes in Computer Science, Springer, 2022*.
- [8] A. Menta, A. Garcia-Serrano, Controllable Sentence Simplification Using Transfer Learning, in: *Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022, CEUR Workshop Proceedings, CEUR-WS.org, Bologna, Italy, 2022*.
- [9] S.-H. Wu, H.-Y. Huang, CYUT Team2 SimpleText Shared Task Report in CLEF-2022, in: *Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022, CEUR Workshop Proceedings, CEUR-WS.org, Bologna, Italy, 2022*.
- [10] A. Rubio, P. Martínez, HULAT-UC3M at SimpleText@CLEF-2022: Scientific text simplification using BART, in: *Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022, CEUR Workshop Proceedings, CEUR-WS.org, Bologna, Italy, 2022*.
- [11] T.-B. Talec-Bernard, Is Using an AI to Simplify a Scientific Text Really Worth It?, in: *Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022, CEUR Workshop Proceedings, CEUR-WS.org, Bologna, Italy, 2022*.
- [12] S. Saha, D. Roy, B. Y. Goud, C. S. Reddy, T. Basu, NLP-IISERB@Simpletext2022: To Explore the Performance of BM25 and Transformer Based Frameworks for Automatic Simplification of Scientific Texts, in: *Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022, CEUR Workshop Proceedings, CEUR-WS.org, Bologna, Italy, 2022*.
- [13] J. Monteiro, M. Aguiar, S. Araújo, Using a Pre-trained SimpleT5 Model for Text Simplification in a limited Corpus, in: *Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022, CEUR Workshop Proceedings, CEUR-WS.org, Bologna, Italy, 2022*.
- [14] H. Jianfei, M. Jin, Assembly Models for SimpleText Task 2: Results from Wuhan University Research Group, in: *Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation*

- Forum, Bologna, Italy, September 5th - to - 8th, 2022, CEUR Workshop Proceedings, CEUR-WS.org, Bologna, Italy, 2022.
- [15] F. Mostert, A. Sampatsing, M. Spronk, J. Kamps, University of Amsterdam at the CLEF 2022 SimpleText Track, in: Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022, CEUR Workshop Proceedings, CEUR-WS.org, Bologna, Italy, 2022.
- [16] term, ???? URL: <https://dictionary.cambridge.org/dictionary/english/term>.
- [17] K. Kageura, E. Marshman, Terminology extraction and management, in: M. O'Hagan (Ed.), *The Routledge Handbook of Translation and Technology*, 1 ed., Routledge, Abingdon, Oxon ; New York, NY : Routledge, 2020. |, 2019, pp. 61–77. URL: <https://www.taylorfrancis.com/books/9781315311241/chapters/10.4324/9781315311258-4>. doi:10.4324/9781315311258-4.
- [18] A. Rigouts Terryn, V. Hoste, P. Drouin, E. Lefever, TermEval 2020: Shared Task on Automatic Term Extraction Using the Annotated Corpora for Term Extraction Research (ACTER) Dataset, in: Proceedings of the 6th International Workshop on Computational Terminology, European Language Resources Association, Marseille, France, 2020, pp. 85–94. URL: <https://aclanthology.org/2020.computerm-1.12>.
- [19] B.-L. Gunnarsson, Language for Special Purposes, in: G. R. Tucker, D. Corson (Eds.), *Encyclopedia of Language and Education*, Springer Netherlands, Dordrecht, 1997, pp. 105–117. URL: http://link.springer.com/10.1007/978-94-011-4419-3_11. doi:10.1007/978-94-011-4419-3_11.
- [20] M. Trojar, Wüster's View of Terminology, *Slovenski jezik / Slovene Linguistic Studies* 11 (2017). URL: <https://ojs.zrc-sazu.si/sjls/article/view/7344>.
- [21] E. Hoffmann, *The LEXIS termbank*, in: Proceedings of Translating and the Computer 9: Potential and practice, Aslib, London, UK, 1987. URL: <https://aclanthology.org/1987.tc-1.14>.
- [22] C. Hermetet-Filez, Des activités de normalisation... à l'élaboration d'un dictionnaire, *Cahiers de l'APLIUT* 9 (1990) 36–39. URL: https://www.persee.fr/doc/apliu_0248-9430_1990_num_9_3_2106. doi:10.3406/apliu.1990.2106.
- [23] K. Wiesner, J. Ladyman, *Measuring complexity* (2019). URL: <https://arxiv.org/abs/1909.13243>. doi:10.48550/ARXIV.1909.13243.
- [24] J. Ladyman, K. Wiesner, *What is a complex system?*, Yale University Press, 2020. URL: <https://yalebooks.yale.edu/book/9780300251104/what-complex-system/>.
- [25] I. Ovchinnikova, D. Nurbakova, L. Ermakova, What science-related topics need to be popularized? A comparative study, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), *Proc. of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum*, Bucharest, Romania, September 21st - to - 24th, 2021, volume 2936 of *CEUR Workshop Proceedings*, 2021, pp. 2242–2255. URL: <http://ceur-ws.org/Vol-2936/paper-203.pdf>.
- [26] B. J. Good, M.-J. Vecchio Good, *The Semantics of Medical Discourse*, in: R. D. Whitley, E. Mendelsohn, Y. Elkana (Eds.), *Sciences and Cultures*, volume 5, Springer Netherlands, Dordrecht, 1981, pp. 177–212. URL: http://link.springer.com/10.1007/978-94-009-8429-5_6. doi:10.1007/978-94-009-8429-5_6.
- [27] A. M. Silletti, *The Role of Illustrations in Popularizing Medical Discourse*, *Linguae & - Rivista di lingue e culture moderne* (2015) 65–81. URL: <http://www.ledonline.it/index.php/linguae/article/view/839>. doi:10.7358/ling-2015-002-sill.
- [28] D. Bourigault, Surface grammatical analysis for the extraction of terminological noun phrases, in: Proceedings of the 14th conference on Computational linguistics -, volume 3, Association for Computational Linguistics, Nantes, France, 1992, p. 977. URL: <http://portal.acm.org/citation.cfm?doid=992383.992415>. doi:10.3115/992383.992415.
- [29] D. A. Evans, R. G. Lefferts, CLARIT-TREC experiments, in: Proceedings of the second conference on Text retrieval conference, TREC-2, Pergamon Press, Inc., USA, 1995, pp. 385–395.
- [30] K. Kageura, B. Umino, Methods of automatic term recognition: A review, *Terminology. International*

Journal of Theoretical and Applied Issues in Specialized Communication 3 (1996) 259–289. URL: <http://www.jbe-platform.com/content/journals/10.1075/term.3.2.03kag>. doi:10.1075/term.3.2.03kag.

- [31] V. Kosa, D. Chaves-Fraga, H. Dobrovolskyi, V. Ermolayev, Optimized Term Extraction Method Based on Computing Merged Partial C-Values, in: V. Ermolayev, F. Mallet, V. Yakovyna, H. C. Mayr, A. Spivakovsky (Eds.), *Information and Communication Technologies in Education, Research, and Industrial Applications*, volume 1175, Springer International Publishing, Cham, 2020, pp. 24–49. URL: http://link.springer.com/10.1007/978-3-030-39459-2_2. doi:10.1007/978-3-030-39459-2_2.
- [32] J. Valaski, S. Reinehr, A. Malucelli, Approaches and Strategies to Extract Relevant Terms: How Are They Being Applied?, in: *Proceedings of The 2015 World Congress in Computer Science, Computer Engineering, and Applied Computing (WorldComp'15)*, Monte Carlo Resort, Las Vegas, USA, 2015, pp. 478–484. URL: <http://worldcomp-proceedings.com/proc/p2015/ICA2668.pdf>.
- [33] Y. Gao, Y. Yuan, Feature-Less End-to-End Nested Term Extraction, in: J. Tang, M.-Y. Kan, D. Zhao, S. Li, H. Zan (Eds.), *Natural Language Processing and Chinese Computing*, volume 11839, Springer International Publishing, Cham, 2019, pp. 607–616. URL: http://link.springer.com/10.1007/978-3-030-32236-6_55. doi:10.1007/978-3-030-32236-6_55.
- [34] M. Kucza, J. Niehues, T. Zenkel, A. Waibel, S. Stüker, Term Extraction via Neural Sequence Labeling a Comparative Evaluation of Strategies Using Recurrent Neural Networks, in: *Interspeech 2018*, ISCA, 2018, pp. 2072–2076. URL: https://www.isca-speech.org/archive/interspeech_2018/kucza18_interspeech.html. doi:10.21437/Interspeech.2018-2017.
- [35] S. Shah, S. S. S. Reddy, Similarity Driven Unsupervised Learning for Materials Science Terminology Extraction, *Computación y Sistemas* 23 (2019). URL: <https://www.cys.cic.ipn.mx/ojs/index.php/CyS/article/view/3266>. doi:10.13053/cys-23-3-3266.
- [36] O. Lieber, O. Sharir, B. Lentz, Y. Shoham, Jurassic-1: Technical Details and Evaluation (2021) 9.
- [37] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, C. Raffel, mT5: A massively multilingual pre-trained text-to-text transformer, in: *Proc. of the 2021 Conference of the North American Chapter of the ACL: Human Language Technologies*, ACL, Online, 2021, pp. 483–498. URL: <https://aclanthology.org/2021.naacl-main.41>.
- [38] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, 2017. URL: <http://arxiv.org/abs/1706.03762>.
- [39] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language Models are Few-Shot Learners (2020). URL: <http://arxiv.org/abs/2005.14165>.
- [40] R. Sennrich, B. Haddow, A. Birch, Neural machine translation of rare words with subword units, in: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 1, Association for Computational Linguistics, 2016, pp. 1715–1725. URL: <http://aclweb.org/anthology/P16-1162>. doi:10.18653/v1/P16-1162.
- [41] M. Johnson, M. Schuster, Q. V. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. Viégas, M. Wattenberg, G. Corrado, et al., Google’s multilingual neural machine translation system: Enabling zero-shot translation, *Transactions of the Association for Computational Linguistics* 5 (2017) 339–351.
- [42] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language models are unsupervised multitask learners, *OpenAI blog* 1 (2019) 9.
- [43] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, *arXiv preprint arXiv:1907.11692* (2019).
- [44] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, *arXiv preprint arXiv:1910.01108* (2019).
- [45] K. Clark, M.-T. Luong, Q. V. Le, C. D. Manning, Electra: Pre-training text encoders as discriminators

- rather than generators, arXiv preprint arXiv:2003.10555 (2020).
- [46] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching Word Vectors with Subword Information, volume 5, 2017, pp. 135–146. URL: https://doi.org/10.1162/tacl_a_00051. doi:10.1162/tacl_a_00051.
 - [47] D. Nadeau, S. Sekine, A survey of named entity recognition and classification, *Linguisticae Investigationes* 30 (2007) 3–26.
 - [48] J. Li, A. Sun, J. Han, C. Li, A Survey on Deep Learning for Named Entity Recognition, *IEEE Transactions on Knowledge and Data Engineering* 34 (2022) 50–70. URL: <https://ieeexplore.ieee.org/document/9039685/>. doi:10.1109/TKDE.2020.2981314.
 - [49] L. Ermakova, P. Bellot, P. Braslavski, J. Kamps, J. Mothe, D. Nurbakova, I. Ovchinnikova, E. SanJuan, Overview of SimpleText 2021 - CLEF Workshop on Text Simplification for Scientific Information Access, in: K. S. Candan, B. Ionescu, L. Goeuriot, B. Larsen, H. Müller, A. Joly, M. Maistro, F. Piroi, G. Faggioli, N. Ferro (Eds.), *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, Lecture Notes in Computer Science, Springer International Publishing, Cham, 2021, pp. 432–449.
 - [50] S. Narayan, C. Gardent, S. B. Cohen, A. Shimorina, Split and Rephrase, in: *Proc. of EMNLP 2017, ACL*, Copenhagen, Denmark, 2017, pp. 606–616. URL: <https://www.aclweb.org/anthology/D17-1064>.
 - [51] 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: *Proc. of EMNLP 2011*, 2011, pp. 782–792.
 - [52] P. Bellot, V. Moriceau, J. Mothe, E. SanJuan, X. Tannier, INEX tweet contextualization task: Evaluation, results and lesson learned, *Inf. Process. Manage.* 52 (2016) 801–819. URL: <https://doi.org/10.1016/j.ipm.2016.03.002>.
 - [53] L. Ermakova, L. Goeuriot, J. Mothe, P. Mulhem, J.-Y. Nie, E. SanJuan, CLEF 2017 Microblog Cultural Contextualization Lab Overview, in: *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 8th International Conference of the CLEF Association, CLEF 2017*, Dublin, Ireland, September 11-14, 2017, *Proc.*, 2017, pp. 304–314. URL: https://doi.org/10.1007/978-3-319-65813-1_27.
 - [54] A. Anand Deshmukh, U. Sethi, IR-BERT: Leveraging BERT for Semantic Search in Background Linking for News Articles 2007 (2020). URL: <http://adsabs.harvard.edu/abs/2020arXiv200712603A>.
 - [55] L. N. Ermakova, D. Nurbakova, I. Ovchinnikova, Covid or not Covid? Topic Shift in Information Cascades on Twitter, in: A. f. C. Linguistics (Ed.), *3rd International Workshop on Rumours and Deception in Social Media (RDSM) Collocated with COLING 2020, Proc. of the 3rd International Workshop on Rumours and Deception in Social Media (RDSM)*, Barcelona (on line), Spain, 2020, pp. 32–37. URL: <https://hal.archives-ouvertes.fr/hal-03066857>.
 - [56] Text Analysis Conference (TAC) 2014 Biomedical Summarization Track, 2014. URL: <https://tac.nist.gov/2014/BiomedSumm/>.
 - [57] D. Wadden, S. Lin, K. Lo, L. L. Wang, M. van Zuylen, A. Cohan, H. Hajishirzi, Fact or Fiction: Verifying Scientific Claims (2020). URL: <http://arxiv.org/abs/2004.14974>.
 - [58] P. Nakov, D. Corney, M. Hasanain, F. Alam, T. Elsayed, A. Barrón-Cedeño, P. Papotti, S. Shaar, G. D. S. Martino, Automated Fact-Checking for Assisting Human Fact-Checkers (2021). URL: <http://arxiv.org/abs/2103.07769>.
 - [59] R. Pradeep, X. Ma, R. Nogueira, J. Lin, Scientific Claim Verification with VERT5ERINI (2020). URL: <http://arxiv.org/abs/2010.11930>.
 - [60] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, Z. Su, ArnetMiner: extraction and mining of academic social networks, in: *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 08*, ACM Press, Las Vegas, Nevada, USA, 2008, p. 990. URL: <http://dl.acm.org/citation.cfm?doid=1401890.1402008>.
 - [61] L. Ermakova, P. Bellot, J. Kamps, D. Nurbakova, I. Ovchinnikova, E. SanJuan, E. Mathurin, S. Araújo, R. Hannachi, S. Huet, N. Poinssu, Automatic Simplification of Scientific Texts: SimpleText Lab at CLEF-2022, in: M. Hagen, S. Verberne, C. Macdonald, C. Seifert, K. Balog, K. Nørsvåg, V. Setty (Eds.),

- Advances in Information Retrieval, volume 13186, Springer International Publishing, Cham, 2022, pp. 364–373.
- [62] C. E. Osgood, Semantic Differential Technique in the Comparative Study of Cultures¹, *American Anthropologist* 66 (1964) 171–200. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1525/aa.1964.66.3.02a00880>.
 - [63] R. Likert, A technique for the measurement of attitudes, *Archives of Psychology* 22 140 (1932) 55–55.
 - [64] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, *Journal of Machine Learning Research* 21 (2020) 1–67. URL: <http://jmlr.org/papers/v21/20-074.html>.
 - [65] G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), Proc. of the Working Notes of CLEF 2022: Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2022.