Overview of the CLEF 2024 SimpleText Task 2: Identify and Explain Difficult Concepts

Notebook for the SimpleText Lab at CLEF 2024

Giorgio Maria Di Nunzio[1], Federica Vezzani[1], Vanessa Bonato[1], Hosein Azarbonyad[2], Jaap Kamps[3] and Liana Ermakova[4]

[1]University of Padova, Padova, Italy
[2]Elsevier, The Netherlands
[3]University of Amsterdam, Amsterdam, The Netherlands
[4]Université de Bretagne Occidentale, HCTI, Brest, France

Abstract

In this paper, we present an overview of the *"Task 2: Complexity Spotting, Identifying and explaining difficult concepts"* within the context of the Automatic Simplification of Scientific Texts (SimpleText) lab, run as part of CLEF 2024. The primary objective of the SimpleText lab is to advance the accessibility of scientific information by facilitating automatic text simplification, thereby promoting a more inclusive approach to scientific knowledge dissemination. *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 the passage, assess their complexity for non-experts, and provide explanations for these detected difficult concepts. A total of 39 submissions were received for this task, originating from 12 distinct teams. In this paper, we describe the data collection process, task configuration, and evaluation methodology employed. Additionally, we provide a brief summary of the various approaches adopted by the participating teams.

Keywords

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

1. Introduction

Despite digitalization making scientific literature more accessible to the public, a major barrier persists: the high complexity of scientific texts. Non-experts struggle to understand these texts due to a lack of background knowledge and specialized terminology. Even native speakers find it challenging to understand terms outside their expertise. Although those with basic education can somewhat understand popular science publications, scientific articles are mostly ignored by the citizens. Understanding terminology is crucial for comprehending scientific information. *Comprehension of the term* implies grasping the concept it represents without the need for an explicit definition. Definitions provide clear explanations of scientific terms, making complex ideas more understandable. Providing accurate definitions and background knowledge can reduce the risk of misinterpreting scientific information and help readers connect new information with what they already know, facilitating better integration and retention of new concepts.

While traditional simplification methods focus on removing complex terms and structures to improve readability [1], providing term definitions and background knowledge could help to make scientific texts more accessible, comprehensible, and meaningful to readers, enabling them to engage with the complex scientific information more effectively. Moreover, Scientific concepts often require precise terminology to avoid ambiguity and ensure clear communication among experts. These terms have

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France
giorgiomaria.dinunzio@unipd.it (G. Di Nunzio); federica.vezzani@unipd.it (F. Vezzani); vanessa.bonato@unipd.it (V. Bonato); liana.ermakova@univ-brest.fr (L. Ermakova)
https://simpletext-project.com/ (L. Ermakova)
0000-0001-9709-6392 (G. Di Nunzio); 0000-0001-9709-6392 (F. Vezzani); 0009-0002-9918-282X (V. Bonato); 0000-0002-6614-0087 (J. Kamps); 0000-0002-7598-7474 (L. Ermakova)

Table 1
CLEF 2024 Simpletext Task 2 official run submission statistics

Task	AIIR Lab	AMATU	Arampatzis	Elsevier	L3S	LIA	PiTheory	Sharigans	SINAI	SONAR	AB/DPV	Dajana/Katya	Frane/Andrea	Petra/Regina	Ruby	Tomislav/Rowan	UAmsterdam	UBO	UniPD	UZH Pandas	Total
2.1	3		5					1	3		1	1	1	1	1	2	1	1	3		24
2.2	3		5					1	3		1		1					1	3		18
2.3			2														2				4

specific meanings that cannot be easily replaced by simpler words without losing accuracy. Besides, many scientific ideas are inherently complex and cannot be adequately described with simple language. Readers can recognize when they need definitions or clarifications for unfamiliar terms, reflecting their awareness of comprehension gaps. This awareness highlights their perception of difficulty with unfamiliar terminology. Thus, we argue that a text simplification method should provide essential information to help readers understand complex scientific concepts instead of deleting difficult terms. This objective is one of the crucial points of the CLEF 2024 SimpleText lab.

The CLEF 2024 SimpleText track[1] is an evaluation lab that follows up on the CLEF 2021 SimpleText Workshop [2], and CLEF 2022-2023 SimpleText Track [3, 4].

The track offers valuable data and benchmarks to facilitate discussions on the challenges associated with automatic text simplification. The CLEF 2024 SimpleText track is based on four interrelated tasks:

1. Task 1 on *Content Selection*: retrieve passages to include in a simplified summary.
2. Task 2 on *Complexity Spotting*: identify and explain diffficult concepts.
3. Task 3 on *Text Simplification*: simplify scientific text.
4. Task 4 on *SOTA?*: track the state-of-the-art in scholarly publications.

This paper focuses on the second task of complexity spotting. The goal of this task is to detect difficult terms and provide contextual explanations for them. Identifying and effectively explaining difficult terms is crucial for promoting accessibility and comprehension of scientific texts. Please refer for details of the other tasks to the overview papers of Task 1 [5] and Task 3 [6], Task 4 [7], as well as the Track overview paper [8].

A total of 45 teams registered for our SimpleText track at CLEF 2024. A total of 20 teams submitted 207 runs in total for the Track, of which 13 teams submitted a total of 46 runs for Task 2. The statistics for the Task 2 runs submitted are presented in Table 1. However, some runs had problems that we could not resolve. We do not detail them in the paper as well as the 0-scored runs.

The rest of this paper is structured in the following way. A comprehensive description of the Task 2 is presented in Section 2. Following that, Section 3 provides an overview of the dataset used, including its composition, size, and relevant characteristics. In Section 4, the paper discusses the evaluation metrics employed to assess the performance of the participants' runs. Section 5 delves into the details of the systems and approaches employed by the participants. In Section 6, we discuss the results of the official submissions. We end with Section 7 discussing the results and findings, and lessons for the future.

2. Task description

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

[1] https://simpletext-project.com

- Task2.1: To predict what the terms are in a passage of a document and the difficulty of the concepts they designate (easy/medium/difficult).
- Task 2.2: To generate a definition and an explanation for each difficult term.
- Task 2.3: To retrieve the provided definitions of the difficult terms in "correct" order.

In Task 2.1, for each passage of a document, participants should provide a list of terms with corresponding scores (easy/medium/difficult) of the concepts they designate. Passages (sentences) are considered to be independent, that is term repetition is allowed (the same term can be detected in different sentences, even in the same document). Detected terms, their spans, and their difficulty will be evaluated. Both qualitative (manual review by terminologists) and quantitative metrics (recall and precision of the extracted terms) will be used to evaluate participants' results.

In Task 2.2, for each term that refers to a difficult concept (those that have been evaluated with the highest level of difficulty), participants should provide the definition and explanation which will be evaluated both from a qualitative point of view (manual review by terminologists) and from a quantitative point of view (overlapping text measures, for example, BLEU score [9]).

In Task 2.3, participants should rank the set of definitions provided for the difficult terms in a way that the "best" definitions are ranked higher in the list of definitions. In particular, for each term there will be one manual definition (considered the best one) and two automatically generated good definitions that should be placed at the top of the list of retrieved definitions.. Quantitative metrics (for example, P@1, P@3, rank correlation measures) will be used to evaluate participants' results.

In general, we asked participants who wanted to run experiments on Task 2.2 to accomplish Task 2.1 first. On the other hand, Task 2.1 and Task 2.3 can be performed independently.

3. Data

The corpus of Task 2 is based on the sentences in high-ranked abstracts to the requests of Task 1. and collected in 2023 [4]. A total of 175 documents and 1,077 sentences were used to generate the training and test data. In particular, we had 115 documents and 576 sentences for building the training set and 60 documents and 501 sentences for building the test set.

We provide a dataset for training the systems before the evaluation phase and a test set for the evaluation phase. In particular, the dataset comprises the following files:

- The documents and their sentences.
- Terms manually extracted from each sentence and their relative difficulty.
- Definitions and the explanations provided by the experts for the difficult terms.
- Definitions automatically generated by a large language model.

For the training set, we engaged 21 experts to manually annotate each document, identifying the terms in each sentence, assessing their difficulty, and providing definitions and explanations for each difficult term. This effort resulted in the generation of 1,609 terms and 899 definitions and explanations. To further analyze the consistency among experts, we deliberately assigned the same documents to multiple experts in some instances.

Additionally, for each term accompanied by a definition, we created two "good" definitions and two "bad" definitions. This was done to develop a set of definitions for ranking in Task 2.3, leading to a total of 2,356 sentences, evenly split between good and bad definitions.

Beyond this initial training set, we introduced an additional set of files produced by an external expert who reviewed the annotations of the 21 experts. This secondary set, referred to as the validation set, included the expert's additions of missing terms, definitions, or both. This review added 677 terms, 960 definitions, and 3,732 generated definitions (equally divided between good and bad) to the training data.

For the test set, we asked the external expert to annotate the remaining 60 documents. A total of 1,440 terms were extracted and 424 definitions were written from the 501 sentences of the test set. An additional 3,816 definitions (equally distributed between good and bad definitions) were also added.

3.1. Annotation Process

A first round of annotation was performed by a number of experts while a second round of validation was performed on the same set of documents by an external expert in order to look for additional (maybe missing) terms and definitions.

The process of annotation of the dataset consisted of two main phases:

1. the extraction of candidate terms from scientific abstracts, and
2. the construction of a collection of definitions of the concepts designated by candidate terms.

Concerning the first phase, we considered that terminology is a "set of designations and concepts belonging to one domain or subject" (ISO 1087: 2019 [10]). The term, therefore, is a "designation that represents a general concept by linguistic means". By reading the abstracts, we identified the subjects or domains of knowledge that each abstract respectively deals with. Some examples of subjects or domains are the medical domain, drone technology and autonomous vehicle technology. Taking subjects and domains into consideration, we identified and extracted candidate terms in the texts of the abstracts. We refer to extracted terms as candidate terms due to the absence of validation of the result of term extraction on the part of experts.

The second phase concerned the construction of a collection of definitions of the concepts designated at the linguistic level by candidate terms. This phase involved two different stages: 1) the retrieval of definitions of the concepts, and 2) the transformation – where necessary – of source definitions into intensional definitions. In line with the ISO 1087: 2019, we adopt the view according to which an intensional definition is a "definition that conveys the intension of a concept by stating the immediate generic concept and the delimiting characteristic(s)".

Specifically, we retrieved definitions in different types of sources: general language dictionaries, specialized dictionaries, websites, papers and quotations included in websites or papers. In particular, the consultation of full-text articles from which abstracts were taken proved to be a useful method to retrieve definitions for specialized concepts. In many cases, the provided definition is a direct quotation of a definition contained in a source. In other cases, we adopted different approaches to the formulation of definitions. For example, we reformulated the source definitions or we embedded in a single definition information contained in more than one source.

In the first round of the annotation, the first set of experts were also asked to write explanations as a more natural and less structured way to clarify and make more intelligible a concept.

3.2. Input format

The training, validation, and test data are provided in TSV formats.

3.2.1. Documents and sentences

The dataset containing the documents and sentences will be stored in a file with the following format:

- doc_id: the identifier of the document
- snt_id: the identifier of sentence
- snt_source: the text of the sentence

For example

doc_id	snt_id	snt_source
1533716782	G01.1_1533716782_5	The users in an initial study [. . .]
2093013061	G10.1_2093013061_1	In this paper, we present an [. . .]
2093013061	G10.1_2093013061_2	From the data of an onboard [. . .]

3.2.2. Terms extracted

The terms manually extracted terms are stored in a file with the following information:

- snt_id: the identifier of each sentence
- term: the term extracted by the user
- difficulty: the difficulty assigned by the expert ([e]asy/[m]edium/[d]ifficult)
- exp_id: the identifier of the expert who annotated that sentence (this column is not present in the validation files)

snt_id	**term**	**difficulty**	**exp_id**
G06.2_2968176166_5	automated	e	1
G06.2_2968176166_5	bayesian	d	1
G01.1_1019677957_1	mobile technology	m	2
G01.1_1019677957_1	mobile emerging carrier	d	2
G01.1_1019677957_1	personal digital assistant	d	2

3.2.3. Definitions and explanations

The definitions and explanations for the difficult terms are stored in a file with the following information:

- snt_id: the identifier of each sentence
- term: the term extracted by the user
- definition: the definition of the term
- explanation: the explanation of the term
- exp_id: the identifier of the expert who annotated that sentence (this column is not present in the validation files)

snt_id	term	definition	explanation	exp_id
G01.1_1549686097_6	computation expenditure	the term "computation expenditure" refers …	This term refers to the resources or costs …	19
G01.2_1448624402_3	descriptive decision theory	Descriptive decision theory is a field of study that ….	Descriptive decision theory is concerned with …	4

3.2.4. Definitions generated

The definitions automatically generated are stored in a file with the following information:

- *snt_id*: the identifier of each annotated sentence
- *term*: the term extracted by the expert
- *definition*: the definition that has been used to generate the positive/negative definitions
- *positive*: two definitions generated automatically that provide a "good" alternative to the "manual" definition. The two definitions are separated by a pipe "|" symbol. +
- *negative*: two definitions generated automatically that provide a "wrong" alternative to the "manual" definition. The two definitions are separated by a pipe "|" symbol.+

3.3. Output format

Results should be provided in a TSV format or JSON format.

snt_id	term	definition	positive	negative
G01.1_1533716782_1	CPU	Principal part of any …	The CPU, or Central Processing Unit, ….|CPU is responsible for …	The CPU temperature….|The performance of the ….
G01.1_1533716782_3	direct migration	A feature which enables....	Direct migration refers to ….|In cloud computing, direct migration….	During the server upgrade...|The direct migration ….

run_id	manual	snt_id	term	difficulty	definition	explanation
unipd_Task2.2_auto	0	G08.1_2889349357_2	cryptojacking	d	my definition of cryptojacking[...]	my explanation of cryptojacking[...]
unipd_Task2.2_auto	0	G08.1_2889349357_2	mine	m		
unipd_Task2.2_auto	0	G01.1_1019677957_3	study	e		
unipd_Task2.2_auto	0	G01.1_1019677957_3	mobile learning	d	my definition of mobile learning[...]	my explanation of mobile learning[...]

Figure 1: *TSV Example for Task 2.1 and Task 2.2*

3.3.1. Task 2.1 and Task 2.2

For Task 2.1 and Task 2.2, we have a single output file that is similar to the input files but it must contain two additional fields :

1. *run_id*: Run ID starting with <team_id>_<task_id>_<method_used>, e.g., *UBO_Task2.2_TFIDF*
2. *manual*: whether the run is manual {0 = no, 1 = yes}
3. *snt_id*: a unique passage (sentence) identifier from the input file
4. *term*: extracted term
5. *difficulty*: difficulty score of the retrieved term e/m/d (easy/medium/difficult)
6. *definition (optional)*: definition of the term (only for difficult terms, and only for Task 2.2)
7. *explanation (optional)*: explanation of the term (only for difficult terms, and only for Task 2.2)

A tabular example (TSV) of the output for Task 2.1 and Task 2.2 is shown in Figure 1.
A JSON example of the same output is shown in Figure 2.

3.3.2. Task 2.3

For Task 2.3, the output file will be more similar to a TREC file. It must contain the following fields:

1. *run_id*: Run ID starting with <team_id>_<task_id>_<method_used>, e.g., *UBO_Task**2.3**_TFIDF*
2. *manual:* whether the run is manual {0 = no, 1 = yes}
3. *snt_id*: a unique passage (sentence) identifier from the input file of the test set
4. *term*: term for which definitions must be ranked
5. *def_id*: a unique identifier of the definition to be ranked
6. *rank:* an integer to specify the rank of this definition (1 highest rank, 2 second highest rank, …)

A tabular example (TSV) of the output for Task 2.3 is shown in Figure 3.
A JSON example of the same output is shown in Figure 4.

4. Evaluation metrics

In order to be as consistent as possible with the previous editions of SimpleText [4], we evaluated

- Task 2.1, difficult concept spotting, in terms of recall and precision;
- Task 2.2, generation of definitions, in terms of BLEU score;
- Task 2.3, the ranking of definitions, with precision@1 and precision@5.

```
[
  {
    "run_id": "unipd_Task2.2_auto",
    "manual": 0,
    "snt_id": "G08.1_2889349357_2",
    "term": "cryptojacking",
    "difficulty": "d",
    "definition": "my definition of cryptojacking […]",
    "explanation": "my explanation of cryptojacking […]"
  },
  {
    "run_id": "unipd_Task2.2_auto",
    "manual": 0,
    "snt_id": "G08.1_2889349357_2",
    "term": "mine",
    "difficulty": "m"
  },
  {
    "run_id": "unipd_Task2.2_auto",
    "manual": 0,
    "snt_id": "G01.1_1019677957_3",
    "term": "study",
    "difficulty": "e"
  },
  {
    "run_id": "unipd_Task2.2_auto",
    "manual": 0,
    "snt_id": "G01.1_1019677957_3",
    "term": "mobile learning",
    "difficulty": "d",
    "definition": "my definition of mobile learning […]",
    "explanation": "my explanation of mobile learning […]"
  }
]
```


Figure 2: *JSON Example for Task 2.1 and Task 2.2*

run_id	manual	snt_id	term	def_id	rank
unipd_Task2.2_auto	0	st123	term 1	def10	1
unipd_Task2.2_auto	0	st456	term 1	def51	2
unipd_Task2.2_auto	0	st098	term 1	def02	3
unipd_Task2.2_auto	0	st876	term 1	def05	4

Figure 3: TSV Example for Task 2.3

In particular, for Task 2.1, we wanted to evaluate both the recall and precision of all the terms, regardless of the difficulty of the concept, and of the difficult concepts only.

For Task 2.2, for the BLEU score we tried different ranges of values of the parameter n for the overlapping n-grams.[2]

For Task 2.3, given the minimal number of runs submitted for Task 3 and the low significance of those results, no analysis will be presented in this paper.

In the future, a qualitative analysis will also be performed in order to study the problems of term identification and the generation of definitions. In addition, we will manually evaluate the provided explanations in terms of their usefulness with regard to a query as well as their complexity for a general

[2] https://cran.r-project.org/web/packages/sacRebleu/vignettes/sacReBLEU.html

```
[
  {
    "run_id": "unipd_Task2.2_auto",
    "manual": 0,
    "snt_id": "st123",
    "term": "term 1",
    "def_id": "def10",
    "rank": 1
  },
  {
    "run_id": "unipd_Task2.2_auto",
    "manual": 0,
    "snt_id": "st456",
    "term": "term 1",
    "def_id": "def51",
    "rank": 2
  },
  {
    "run_id": "unipd_Task2.2_auto",
    "manual": 0,
    "snt_id": "st098",
    "term": "term 1",
    "def_id": "def02",
    "rank": 3
  },
  {
    "run_id": "unipd_Task2.2_auto",
    "manual": 0,
    "snt_id": "st876",
    "term": "term 1",
    "def_id": "def05",
    "rank": 4
  }
]
```


Figure 4: *JSON Example for Task 2.3*

audience. The provided explanations can have different forms, e.g., abbreviation deciphering, examples, use cases, etc.

5. Participant's Approaches

In this section, we describe the main approaches of each participant who submitted at least one run of a model to be evaluated on the test set.[3]

AB&DPV [11] submitted one run, employing natural language processing techniques to identify difficult terms within passages. They generated definitions for these terms or retrieved them from sources like Wikipedia. However, they did not submit runs on the test set.

AIIRLab [12] submitted three runs, utilizing LLaMA3 and Mistral language models. Their approach included prompt engineering and reinforcement learning with human feedback to enhance the quality of outputs generated by the LLaMA model.

Dajana&Kathy [13] submitted one run, using the LLAMA-2 13B model. No further information about their approach was provided.

Frane&Andrea [14] submitted one run. No additional details about their methodology were given.

[3]Some participant decided to submit the experimental result on the training set which will be useful for future post-hoc analyses.

Sharingans [15] submitted one run for, fine-tuning the GPT-3.5 turbo model for selecting difficult terms and generating definitions and explanations. They employed prompt-engineering techniques to create specific prompts that guided the model in producing accurate and contextually relevant definitions.

SINAI [16] submitted three runs, applying learning cues without prior examples to the GPT-4-Turbo model. They used the OpenAI API in Python to interact with the model, facilitating the integration of GPT-4-Turbo into their workflow.

team1_Petra_and_Regina [17] submitted one run, combining named entity recognition (NER) techniques with rule-based approaches to identify and extract entities such as proteins, genes, and chemical compounds. They utilized spaCy for NER and developed custom rules for entity extraction.

Tomislav&Rowan [18] submitted two runs. They created prompts for the LLAMA-2 13B model to extract three scientific terms from each source sentence and then prompted the model to return a difficulty rating. Definitions for the difficult terms were retrieved from Wikipedia.

UAms [19] submitted three runs. They employed an idf-based term weighting to identify the rarest terms for Task 2.1. For Task 2.3, they developed a method to rank definitions or explanations for given sentence-term pairs by examining the textual similarity of the provided sentences.

UBO [20] submitted one run, using the Small Language Model, Phi3 mini, without fine-tuning, employing a one-shot prompt approach.

UNIPD [21] submitted three runs, focusing on identifying and explaining difficult content using Large Language Models (LLMs) to enhance text simplification. They iteratively experimented with various prompting strategies to optimize model performance for this task.

6. Results

In this Section, we present the results on the test set for Task 2.1 and Task 2.2. At present time, the results for Task 2.3 are still ongoing (with only two runs by one participant) and will be made available in the future.

6.1. Test Results

The results on the test set are summarized in Table **??**. For each run, we report:

- the recall of all the terms, independently from the level of difficulty;
- the precision of all the terms, independently from the level of difficulty;
- the F1 score of all the terms, independently from the level of difficulty;
- the recall of the difficult terms;
- the precision of the difficult terms;
- the F1 score of the difficult terms;
- the BLEU score computed for bigrams (ngrams from n = 1 to n = 4).[4]

For recall, precision, and F1 we report both the "overall" scores computed by summing up all the data for all the retrieved and predicted terms for all the sentences, and the "average" scores computed by averaging the recall and precision of the predicted terms per sentence.

In Table 2, we report the results for all the terms; in Table 3, we show the results for the difficult terms only; in Table 4, we present the results of the BLEU scores.

Table 2
Results for CLEF 2024 SimpleText Task 2 for all terms.

runid	recall overall	precision overall	f1 overall	recall average	precision average	f1 average
AIIRLab_Task2.2_LLaMA	0.301	0.525	0.383	0.297	0.693	0.383
AIIRLab_Task2.2_LLaMAFT	0.008	0.989	0.016	0.005	0.217	0.016
AIIRLab_Task2.2_Mistral	0.469	0.559	0.510	0.436	0.494	0.510
Dajana&Kathy_SimpleText_Task2.2_LLAMA2_13B_CHAT	0.007	0.585	0.015	0.011	0.591	0.015
FRANE_AND_ANDREA_SimpleText_Task2.2_LLAMA2_13B_CHAT	0.005	0.645	0.010	0.007	0.458	0.010
team1_Petra_and_Regina_Task2_ST	0.003	0.500	0.005	0.004	0.500	0.005
Sharingans_Task2.2_GPT	0.485	0.428	0.455	0.524	0.650	0.455
SINAI_task_2_PRM_ZS_TASK2_V1	0.090	0.739	0.160	0.086	0.671	0.160
SINAI_task_2_PRM_ZS_TASK2_V2	0.167	0.672	0.268	0.161	0.499	0.268
SINAI_task_2_PRM_ZS_TASK2_V3	0.101	0.790	0.180	0.103	0.442	0.180
Tomislav&Rowan_Task2.2_LLAMA2_13B_CHAT_1	0.005	0.613	0.010	0.008	0.519	0.010
Tomislav&Rowan_Task2.2_LLAMA2_13B_CHAT	0.004	0.333	0.009	0.004	0.667	0.009
UAms_Task2-1_RareIDF	0.082	0.959	0.152	0.093	0.326	0.152
UboNLP_Task2.1_phi3-oneshot	0.582	0.527	0.553	0.555	0.563	0.553
unipd_t21t22_chatgpt	0.116	0.562	0.192	0.138	0.703	0.192
unipd_t21t22_chatgpt_mod1	0.227	0.398	0.289	0.234	0.764	0.289
unipd_t21t22_chatgpt_mod2	0.331	0.338	0.334	0.311	0.780	0.334

Table 3
Results for CLEF 2024 SimpleText Task 2 only for difficult terms.

runid	recall overall (difficult)	precision overall (difficult)	f1 overall (difficult)	recall average (difficult)	precision average (difficult)	f1 average (difficult)
AIIRLab_Task2.2_LLaMA	0.299	0.681	0.415	0.307	0.950	0.465
AIIRLab_Task2.2_LLaMAFT	0.006	1.000	0.012	0.007	1.000	0.014
AIIRLab_Task2.2_Mistral	0.212	0.485	0.295	0.199	0.892	0.326
Dajana&Kathy_SimpleText_Task2.2_LLAMA2_13B_CHAT	0.000	0.000	0.000	0.000	0.989	0.000
FRANE_AND_ANDREA_SimpleText_Task2.2_LLAMA2_13B_CHAT	0.006	0.364	0.012	0.010	0.981	0.020
team1_Petra_and_Regina_Task2_ST	0.000	0.000	0.000	0.000	0.995	0.000
Sharingans_Task2.2_GPT	0.565	0.587	0.576	0.583	0.854	0.693
SINAI_task_2_PRM_ZS_TASK2_V1	0.105	0.538	0.176	0.092	0.935	0.167
SINAI_task_2_PRM_ZS_TASK2_V2	0.149	0.806	0.251	0.134	0.978	0.236
SINAI_task_2_PRM_ZS_TASK2_V3	0.053	0.857	0.101	0.047	0.995	0.090
Tomislav&Rowan_Task2.2_LLAMA2_13B_CHAT_1	0.000	0.000	0.000	0.000	1.000	0.000
Tomislav&Rowan_Task2.2_LLAMA2_13B_CHAT	0.000	0.000	0.000	0.000	1.000	0.000
UAms_Task2-1_RareIDF	0.025	0.091	0.040	0.034	0.780	0.066
UboNLP_Task2.1_phi3-oneshot	0.351	0.387	0.368	0.332	0.737	0.457
unipd_t21t22_chatgpt	0.077	0.612	0.137	0.087	0.979	0.160
unipd_t21t22_chatgpt_mod1	0.226	0.591	0.327	0.234	0.979	0.378
unipd_t21t22_chatgpt_mod2	0.385	0.682	0.492	0.324	0.986	0.488

In, Figure 5 and Figure 6 the precision and recall results for the overall and averaged measures. We

[4] https://cran.r-project.org/web/packages/sacRebleu/vignettes/sacReBLEU.html

Table 4
Results for CLEF 2024 SimpleText Task 2 for the quality of definitions.

runid	BLEU (n1) average	BLEU (n2) average	BLEU (n3) average	BLEU (n4) average
AIIRLab_Task2.2_LLaMA	0.286	0.150	0.047	0.018
AIIRLab_Task2.2_LLaMAFT	0.240	0.117	0.000	0.000
AIIRLab_Task2.2_Mistral	0.259	0.133	0.041	0.014
Dajana&Kathy_SimpleText_Task2.2_LLAMA2_13B_CHAT	0.000	0.000	0.000	0.000
FRANE_AND_ANDREA_SimpleText_Task2.2_LLAMA2_13B_CHAT	0.000	0.000	0.000	0.000
team1_Petra_and_Regina_Task2_ST	0.000	0.000	0.000	0.000
Sharingans_Task2.2_GPT	0.227	0.106	0.031	0.016
SINAI_task_2_PRM_ZS_TASK2_V1	0.252	0.157	0.082	0.060
SINAI_task_2_PRM_ZS_TASK2_V2	0.276	0.159	0.067	0.049
SINAI_task_2_PRM_ZS_TASK2_V3	0.216	0.112	0.039	0.025
Tomislav&Rowan_Task2.2_LLAMA2_13B_CHAT_1	0.000	0.000	0.000	0.000
Tomislav&Rowan_Task2.2_LLAMA2_13B_CHAT	0.000	0.000	0.000	0.000
UAms_Task2-1_RareIDF	0.000	0.000	0.000	0.000
UboNLP_Task2.1_phi3-oneshot	0.001	0.000	0.000	0.000
unipd_t21t22_chatgpt	0.309	0.185	0.089	0.049
unipd_t21t22_chatgpt_mod1	0.311	0.181	0.082	0.045
unipd_t21t22_chatgpt_mod2	0.030	0.007	0.003	0.000

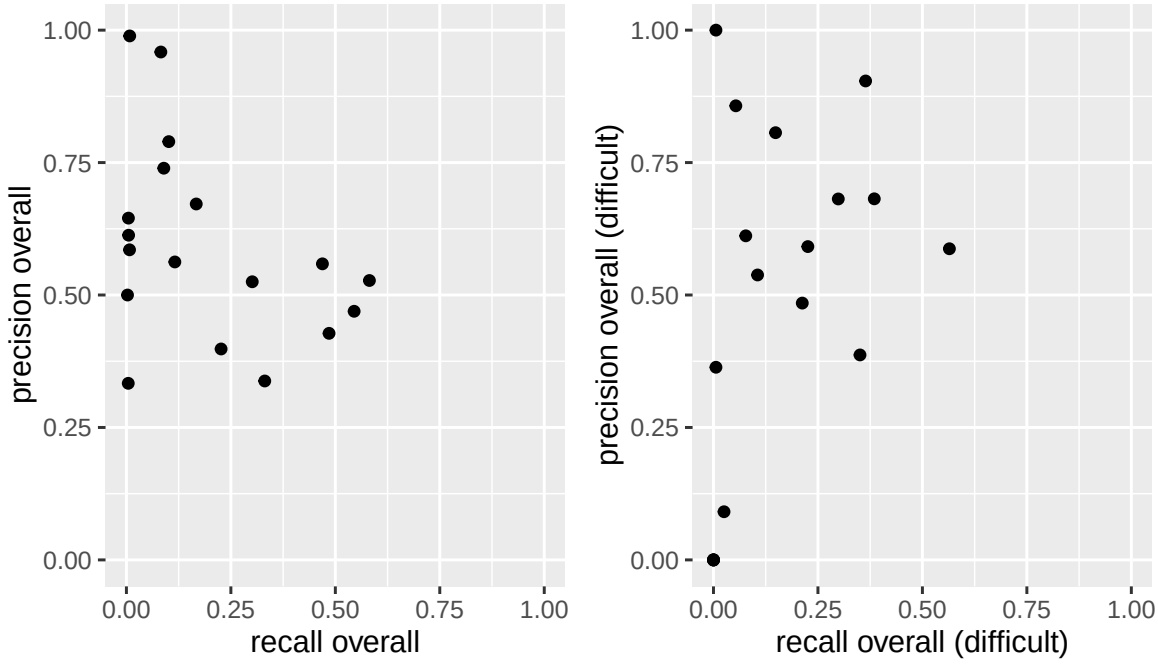

Figure 5: Summary of Task 2 Recall-Precision "overall" results for all terms and difficult terms.

also show the trend of the main results of the BLEU scores (for n = 1 and n = 2) compared to the average f1 score for difficult terms in Figure 7.

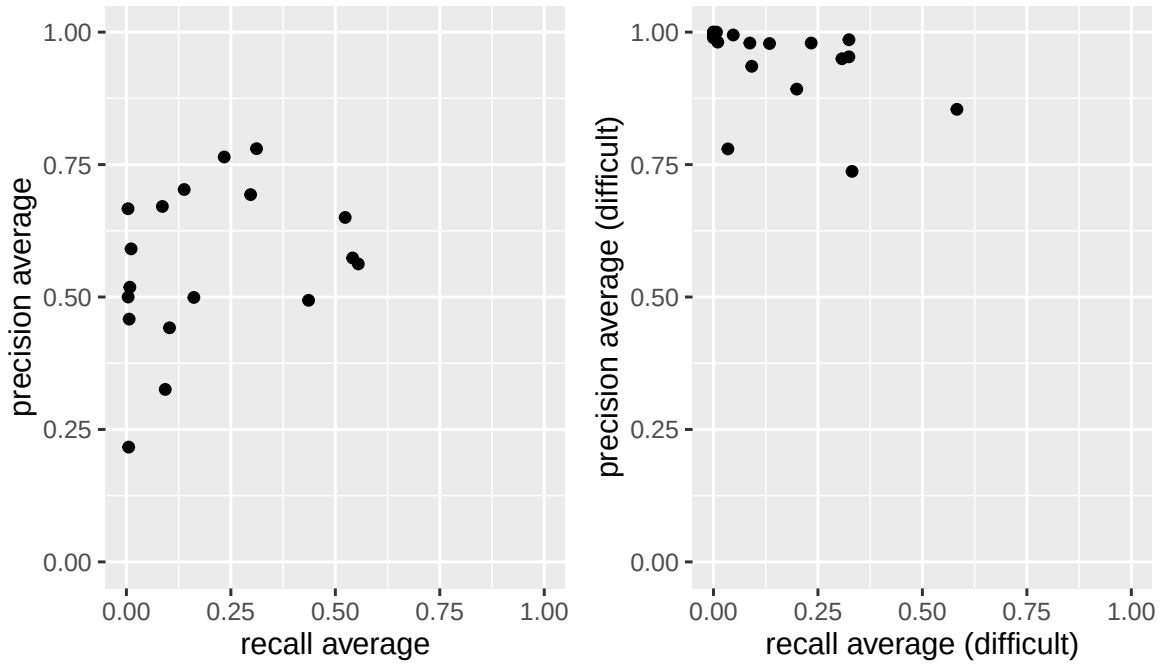

Figure 6: Summary of Task 2 Recall-Precision "average" results for all terms and difficult terms.

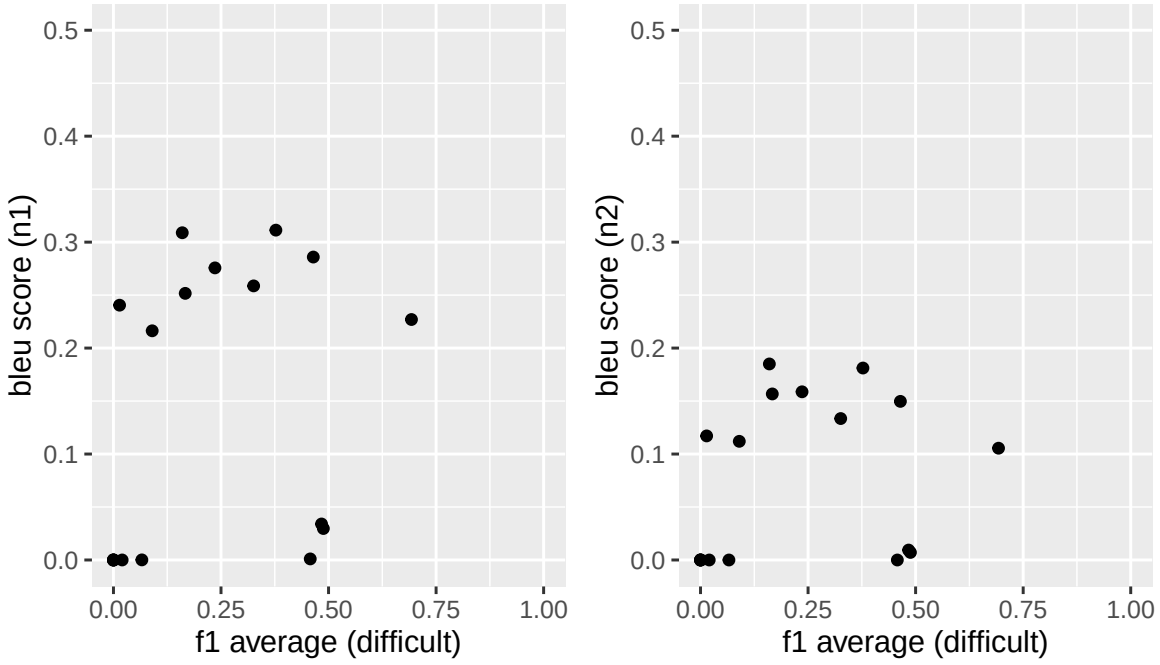

Figure 7: Summary of Task 2 BLEU scores results for n = 1 and n = 2.

6.2. Quantitative Analysis

The results shown in the previous section reveal that the use of large language models for the extraction of terms, the assessment of the difficulty of these terms, and the generation of the definitions to explain the difficult concepts are at an initial stage that will open new perspective in the Automatic Term Extraction panorama. In particular, compared to the recent results and surveys (see [22]), the values of

recall and precision are sufficiently good but suboptimal when compared to the state-of-the-art models (of course, we need to take into consideration that this is the first time participants dealt with this new dataset).

Our main findings are the following. First, the runs submitted by the participants to this task are quite stable in terms of recall-precision performances when dealing with all the terms or the difficult ones. Independently from the difficulty of terms, the models proposed by the participants can achieve precision higher than .50 across a range of recall values. The best runs can achieve an average recall in the range of 0.3 - 0.5 while obtaining a precision between 0.4 and 0.7. It is interesting to see that the best average precision (0.7604) for all the terms is achieved by the experiment that performed a manual cleaning and intervention of the output of ChatGPT. If we focus only on the difficult terms, the recall in general decreases while the precision increases, which means that a smaller amount of difficult terms are found but the system is very precise in detecting them. Second, the BLEU score of the generated definitions is also relatively stable ranging from 0.2 to 0.3 for n = 1 and and from 0.1 to 0.2 for n = 2 for any recall values. Third, the best performing runs are usually those that has some analysis of the optimal prompting or a manual interaction with the model. This is inline with the latest research studies on this issue [23].

6.3. Qualitative Analysis

In this section, we present a qualitative analysis in which we comparatively evaluate the definitions of concepts provided by an expert in terminology and the definitions proposed by different participants. In the following analysis, the first definition is the one provided by the terminologist and the second one is the definition provided by the participants. For each of the definitions submitted by the participants of task 2.2, the BLEU score is computed. This score is calculated by computing the degree of similarity that is established between the concept definition provided by and the definition formulated by the participants. For each analyzed participant, we analyze form a qualitative perspective the definitions that respectively obtained the highest and the loweest BLEU score computed with n = 2.

For what concerns the concept designated by the term “MEC server”, we analyze the following two definitions. The definition provided by the participant is the one that obtained the highest BLEU score which amounts to 0.3223. The term is included in the sentence identified by the code G06.2_2895666646_7.

- Server of Multi-access edge computing, which is a cloud service running at the edge of a network and performing specific tasks that would otherwise be processed in centralized core or cloud infrastructures.
- A Multi-Access Edge Computer (MEC) server is a type of server technology located at the edge of a communication network to reduce the latency and increase the speed of data delivery.

From a qualitative viewpoint, it is possible to observe that both definitions include the indication of the extended form of the acronym “MEC” and refer to the location of the server with respect to a network. However, some terms that are present in the expert’s definition do not match the terms included in the participant’s definition, as in the case of “cloud service”, “task”, “centralized core infrastructure” and “cloud infrastructure”.

The definition provided by the same participant that obtained the lowest BLEU score (0.0247) is the definition of the concept designated by the term “NCI”. The sentence in which the term is present is identified by the code G08.2_1607424157_4.

- Measure of collective behaviour based on financial news on the Web, which captures the average mutual similarity between the documents and entities in the financial corpus.
- NCI stands for Named Complexity Index. As it can be observed, the text provided by the participant aims at explaining the abbreviated form “NCI” by indicating the extended version of the acronym rather than defining the concept designated by the term itself. Indeed, the text does not express the intension (that is the main characteristics) of the concept and, as a consequence, it cannot be considered as a terminological definition.

Another participant, proposed the definition of the concept designated by the term “Windows 2000 machine” obtaining the highest BLEU score, amounting to 0.5939. The term is contained in the sentence whose identifier is G01.1_1522515958_7.

- Computer system that uses the operating system called Windows 2000.
- A computer that uses the Windows 2000 operating system.

The major difference between the two definitions lies in the indication of the generic concept, which is respectively identified as “computer system” by the terminologist, and as “computer” by the participant. In this case, “computer system” refers to the entire assembly of hardware and software (including the operating system) is the most accurate choice rather than "computer" which usually refers to the hardware only.

The definition provided by the same participant of the concept designated by the term “Mixed Logit model” obtained the lowest BLEU score (0.0023). The term is included in the sentence identified by the code G06.2_2982382045_4.

- Statistical model, that is fully general, for examining discrete choices, which allows for random taste variation across choosers, unrestricted substitution patterns across choices, and correlation in unobserved factors over time.
- Model used in statistics. In the definition provided by the participant, an absence of delimiting characteristics of the concepts can be observed. As a matter of fact, the definition would not allow to distinguish the Mixed Logit model from other models used in the domain of statistics. In this sense, the text could be better classified as a generic explanation rather than a terminological definition.

The same procedure is applied to the definitions provided by another participant.

The definition presented by the participant that reached the highest BLEU score (0.5298) refers to the concept designated by the term “GTAV”. The sentence in which the term is included is identified by the code G06.2_2890116921_6.

- Action-adventure game developed by Rockstar North and published by Rockstar Games.
- Grand Theft Auto V, an action-adventure video game developed by Rockstar North.

As it can be observed, the definition presented by the participant includes elements that are not present in the definition provided by the expert, that are: 1) the extended form of the term, and 2) the specification concerning the fact that Grand Theft Auto V is a video game. The definition provided by the terminologist, however, also indicates the publisher of the game. In the definition provided by the participant, the generic concept is not placed as the first element of the text, thus not following the structure attributed to intensional definitions. The definition included by the participant that reached the lowest BLEU score (0.03556) is related to the concept designated by the term “Autoware”. The term can be individuated in the sentence identified by the code G06.2_2931522054_8.

- Software stack platform for self-driving that is ROS-based, composed of an abundant set of self-driving modules, such as sensing, localization, detection, planning, and actuation, and libraries that render it possible to operate and simulate autonomous vehicle.
- An open-source software stack designed for self-driving vehicles. The difference between the two definitions consists in the richness of information that characterizes the definition of the expert, with respect to the amount of information provided in the second one. Moreover, the second definition does not follow the structure of intensional definitions.

We additionally perform the analysis of the definitions inserted by another participant. The definition provided by the participant of the concept designated by the term “Denial-of-Service” obtained the highest BLEU score (0.8678). The term is found in the sentence whose identifier is G06.2_2548923997_6.

- Cyber-attack in which the perpetrator seeks to make a machine or network resource unavailable to its intended users by temporarily or indefinitely disrupting services of a host connected to a network.
- A cyber-attack where the perpetrator seeks to make a machine or network resource unavailable to its intended users by temporarily or indefinitely disrupting services of a host connected to the Internet.

In this circumstance, it is possible to observe that the two definitions present a high level of similarity. The major difference lies in the respective usage of the term "network" and "Internet" to designate the connection used by the host in this type of cyber-attack. The participant also provided a definition of the concept designated by the term "Autoware", that presented the lowest BLEU score (0.0231) with respect to the expert's definition. The term can be found in the sentence identified as G06.2_2931522054_7.

- Software stack platform for self-driving that is ROS-based, composed of an abundant set of self-driving modules, such as sensing, localization, detection, planning, and actuation, and libraries that render it possible to operate and simulate autonomous vehicle.
- An open-source software platform designed to enable autonomy for various types of vehicles.

In the first definition, a greater amount of information is given. Another difference that can be observed is that the second definition. Finally, we evaluate the performance in terms of BLEU score obtained by two definitions inserted by another participant. We begin with the analysis of the definition that obtained the highest BLEU score (0.5621), related to the concept designated by the term "Windows 2000 machine". The term is comprised in a sentence whose identifier is G01.1_1522515958_7.

- Computer system that uses the operating system called Windows 2000.
- Computer system running the Windows 2000 operating system.

Both definitions are characterized by the presence of the same generic concept, linguistically designated by the term "computer system". The differences between the two definitions are: 1) the use of different verbs, and 2) the diverse structural arrangement of information. The same participant also provided a definition of the concept designated by the term "Dr Who CSR engine". This definition obtained the lowest BLEU score, amounting to 0.0240. The term is included in a sentence identified by the code G01.1_1522515958_8.

- Engine for continuous speech recognition developed in the context of the Microsoft research project Dr Who, which uses a unified language model that takes advantage of rule-based and data-driven approach.
- Speech recognition engine used in MiPad. Also in this case, it can be observed that the amount of information provided in the context of the two definitions does not match.

To conclude, we propose a qualitative analysis concerning the results stemming from the task of term extraction performed by an expert in terminology and by different participants. In particular, we focus on the detection of lexical units classified as terms relevant for the domain by participants, that are however excluded from the list of terms extracted by the expert. From the observation of the sentence identified by the code G01.1_130055196_1, we noticed that a participant considered "PDA (Personal Digital Assistant)" as a term. This string of characters, however, does not correspond to a term. As a matter of fact, it is possible to identify two different terms designating the same concept: 1) "PDA", and 2) "Personal Digital Assistant". In particular, "PDA" is the acronym for "Personal Digital Assistant". Moreover, with specific reference to the sentence whose identifier is G01.1_135571562_4, the participant also selected "with a barcode reader" as a term. The correct term, however, is "barcode reader". Another string of characters that does not constitute a term is "regarded as important", contained in the sentence coded as G01.1_135571562_7. This is due to the fact that "regarded as important" does not designate a concept in specialized domains of knowledge. The lexical units "regarded" and "important" were also extracted as terms by another participant, in the context of the sentence coded as G01.1_135571562_7.

Nevertheless, even when considered as two different lexical units, they do not constitute designations of concepts in specialized fields of knowledge. Moreover, a participant considered cardinal numbers as constituting elements of terms, such as in “eight sets” and “five repetitions”, both contained in the sentence coded as M1_13_1. These strings of characters should not be extracted as terms, due to the consideration that a correct term extraction would result in the extraction of the terms “set” and “repetition”. Furthermore, the string of characters “clinicianu0027s personal assistant”, included in the sentence identified by the code G01.1_1462481249_3, was also detected as a term by a participant. However, in this case two different terms can be individuated: 1) “clinician” and 2) “personal assistant”. The string of characters “u0027” should not be considered as a constituting element of a term, as it represents the Unicode number for the Unicode character “’”. The term “closest facilities”, contained in the sentence whose identifier is G01.1_1000902583_3, was also regarded as a term by a participant. However, “closest” constitutes a superlative adjective. Considering that, the term that should be extracted is “facility”.

7. Conclusion and future work

The results of Task 2 of the CLEF 2024 SimpleText challenge have demonstrated the potential and limitations of current natural language processing (NLP) models in identifying and defining difficult complex within scientific texts, and ranking available definitions for those concepts. The task was divided into three subtasks: identifying terms and their difficulty (Task 2.1), generating definitions and explanations for difficult terms (Task 2.2), and ranking definitions (Task 2.3). The diversity of approaches taken by the participants showcased various strategies and methodologies in tackling these problems.

7.1. Summary of Findings

In Task 2.1, precision and recall metrics highlighted that while some systems could accurately identify terms, there was a general challenge in consistently predicting the difficulty level. Several approaches, using models like LLaMA and Mistral, showed promising results in term identification, but the precision for difficult terms varied significantly.

Task 2.2 focused on generating definitions, where the BLEU score was used to evaluate the overlap between generated and reference definitions. Here, the performance varied, with some models generating coherent definitions, while others struggled with accuracy and relevance. Some fine-tuned LLM model achieved notable BLEU scores, indicating effective use of reinforcement learning and prompt engineering.

For Task 2.3, which required ranking definitions, there was limited participation, and the results are still ongoing. The preliminary findings suggest that ranking automatically generated definitions in the correct order remains a significant challenge, and further analysis is needed to draw concrete conclusions.

7.2. Future Work

The task results point to several directions for future research and development. Future work should focus on improving the precision and recall of term identification, particularly for difficult terms. This may involve integrating more sophisticated context-aware models and leveraging domain-specific knowledge bases. The quality of generated definitions needs enhancement. Research into more advanced language models, fine-tuning techniques, and hybrid approaches combining rule-based and machine learning methods could yield better results. The current evaluation metrics provide a good starting point, but future tasks could benefit from more nuanced and context-sensitive metrics that better capture the quality and relevance of generated definitions and explanations. Incorporating human feedback iteratively in the model training process can significantly improve the quality of outputs.

In conclusion, the CLEF 2024 SimpleText challenge has provided valuable insights into the capabilities of current NLP models in understanding and processing complex scientific texts. Continued research and collaboration within the community will be essential in addressing the identified challenges and advancing the state of the art in this field.

Acknowledgments

This work is partially supported by the HEREDITARY Project, as part of the European Union's Horizon Europe research and innovation programme under grant agreement No GA 101137074.

This research was funded, in whole or in part, by the French National Research Agency (ANR) under the project ANR-22-CE23-0019-01.

We would like to thank Jaap Kamps, Valentin Laimé, Radia Hannachi, Silvia Araújo, Pierre De Loor, Olga Popova, Diana Nurbakova, Quentin Dubreuil, and all the other colleagues and participants who helped run this track.

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] 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 - 12th International Conference of the CLEF Association, CLEF 2021, Virtual Event, September 21-24, 2021, Proceedings, volume 12880 of *Lecture Notes in Computer Science*, Springer, 2021, pp. 432–449. URL: https://doi.org/10.1007/978-3-030-85251-1_27. doi:10.1007/978-3-030-85251-1_27.

[3] 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.), Experimental IR Meets Multilinguality, Multimodality, and Interaction - 13th International Conference of the CLEF Association, CLEF 2022, Bologna, Italy, September 5-8, 2022, Proceedings, volume 13390 of *Lecture Notes in Computer Science*, Springer, 2022, pp. 470–494. URL: https://doi.org/10.1007/978-3-031-13643-6_28. doi:10.1007/978-3-031-13643-6_28.

[4] L. Ermakova, E. SanJuan, S. Huet, H. Azarbonyad, O. Augereau, J. Kamps, Overview of the CLEF 2023 simpletext lab: Automatic simplification of scientific texts, in: A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, A. Giachanou, D. Li, M. Aliannejadi, M. Vlachos, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction - 14th International Conference of the CLEF Association, CLEF 2023, Thessaloniki, Greece, September 18-21, 2023, Proceedings, volume 14163 of *Lecture Notes in Computer Science*, Springer, 2023, pp. 482–506. URL: https://doi.org/10.1007/978-3-031-42448-9_30. doi:10.1007/978-3-031-42448-9_30.

[5] E. SanJuan, S. Huet, J. Kamps, L. Ermakova, Overview of the CLEF 2024 SimpleText task 1: Retrieve passages to include in a simplified summary, in: [24], 2024.

[6] L. Ermakova, V. Laimé, H. McCombie, J. Kamps, Overview of the CLEF 2024 SimpleText task 3: Simplify scientific text, in: [24], 2024.

[7] J. D'Souza, et al., Overview of the CLEF 2024 SimpleText task 4: Track the state-of-the-art in scholarly publications, in: [24], 2024.

[8] L. Ermakova, E. SanJuan, S. Huet, H. Azarbonyad, G. M. Di Nunzio, F. Vezzani, J. D'Souza, J. Kamps, Overview of the CLEF 2024 SimpleText track: Improving access to scientific texts for everyone, in: L. Goeuriot, G. Q. Philippe Mulhem, D. Schwab, L. Soulier, G. M. D. Nunzio, P. Galuščáková, A. G. S. de Herrera, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024), Lecture Notes in Computer Science, Springer, 2024.

[9] 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 on Association for Computational Linguistics, ACL '02, Association for Computational Linguistics, USA, 2002, pp. 311–318. URL: https://doi.org/10.3115/1073083.1073135. doi:10.3115/1073083.1073135.

[10] ISO1087:2019, Terminology work and terminology science — Vocabulary, Standard, International Organization for Standardization, Geneva, CH, 2019.
[11] D. P. Varadi, A. Bartulović, SimpleText 2024: Scientific Text Made Simpler Through the Use of AI, in: [24], 2024.
[12] N. Largey, R. Maarefdoust, S. Durgin, B. Mansouri, AIIR Lab Systems for CLEF 2024 SimpleText: Large Language Models for Text Simplification, in: [24], 2024.
[13] K. Seng, D. Simunovic, Simplify It Like It's Hot: Making Complex Texts Easy to Digest, in: [24], 2024.
[14] A. Zečević, F. Doljanin, Simplify It Like It's Hot: Making Complex Texts Easy to Digest, in: [24], 2024.
[15] S. M. Ali, H. Sajid, O. Aijaz, O. Waheed, F. Alvi, A. Samad, Improving Scientific Text Comprehension: A Multi-Task Approach with GPT-3.5 Turbo and Neural Ranking, in: [24], 2024.
[16] J. A. Ortiz-Zambrano, C. Espin-Riofrio, A. Montejo-Ráez, SINAI Participation in SimpleText Task 2 at CLEF 2024: Zero-shot Prompting on GPT-4-Turbo for Lexical Complexity Prediction, in: [24], 2024.
[17] R. Elagina, P. Vučić, AI Contributions to Simplifying Scientific Discourse in SimpleText 2024, in: [24], 2024.
[18] R. Mann, T. Mikulandric, CLEF 2024 SimpleText Tasks 1-3: Use of LLaMA-2 for text simplification, in: [24], 2024.
[19] J. Bakker, G. Yüksel, J. Kamps, University of Amsterdam at the CLEF 2024 SimpleText Track, in: [24], 2024.
[20] B. Vendeville, L. Ermakova, P. De Loor, UBO NLP report on the SimpleText track at CLEF 2024, in: [24], 2024.
[21] G. M. Di Nunzio, F. Vezzani, E. Gallina, UNIPD@SimpleText2024: A Semi-Manual Approach on Prompting ChatGPT for Extracting Terms and Write Terminological Definitions, in: [24], 2024.
[22] G. Di Nunzio, S. Marchesin, G. Silvello, A systematic review of Automatic Term Extraction: What happened in 2022?, Digital Scholarship in the Humanities 38 (2023) i41–i47. URL: https://doi.org/10.1093/llc/fqad030. doi:10.1093/llc/fqad030.
[23] Z. Lin, How to write effective prompts for large language models, Nature Human Behaviour 8 (2024) 611–615. URL: https://www.nature.com/articles/s41562-024-01847-2. doi:10.1038/s41562-024-01847-2, publisher: Nature Publishing Group.
[24] G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024: Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS.org, 2024.