Overview of the CLEF 2022 SimpleText Task 1: Passage Selection for a Simplified Summary

Eric SanJuan[1], Stéphane Huet[1], Jaap Kamps[2] and Liana Ermakova[3]

[1]Avignon Université, LIA, France
[2]University of Amsterdam, Amsterdam, The Netherlands
[3]Université de Bretagne Occidentale, HCTI, France

Abstract

This paper presents an overview of the CLEF 2022 SimpleText track's Task 1, asking systems to retrieve scientific abstracts in response to a query prompted by a popular science article. We discuss the details of the task set-up: First, the SimpleText Corpus with over 4 million academic papers and abstracts. Second, the Topics based on 40 popular science articles in the news and the 114 Queries prompted by them. Third, the Formats of requests and results, the Evaluation labels and Evaluation measures used. Fourth, the Results of the runs submitted by our participants.

Keywords

information retrieval, popular science, automatic summarization, simplification

1. Introduction

In this paper, we discuss the SimpleText track's first task about content selection (and *avoiding* complexity) from a corpus of scientific abstracts, addressing the task:

> *Select passages to include in a simplified summary, given a query.*

The task aims at finding references in computer science that could be inserted as citations in original press articles of general audience for illustration, fact checking or actualisation. For each of the selected references, more relevant sentences need to be extracted. These passages can be complex and require further simplification to be carried out in tasks 2 and 3. Task 1 focuses on content retrieval.

A total of 62 potential participants registered for the track, and many teams downloaded the data or used the online API to explore the collection. However, building an effective retrieval system proved challenging for many teams. Ultimately, we received a total of six official submissions from three different teams:

- Chaoyang University of Technology (CYUT) [1] submitted a single run;
- Indian Institute of Science Education and Research Bhopal (IISERB) [2] submitted three different runs; and

CLEF 2022: Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy
eric.sanjuan@univ-avignon.fr (E. SanJuan); liana.ermakova@univ-brest.fr (L. Ermakova)
https://simpletext-project.com/ (L. Ermakova)
0000-0002-7598-7474 (L. Ermakova)

- University of Amsterdam (UAms) [3] submitted two runs.

The rest of this paper is structured as follows. Next, in Section 2, we provide some context of the task and discuss related work. In Section 3 we detail the exact setup of the tasks and the resulting SimpleText Task 1 Test Collection, consisting of the corpus, topics and queries, judgments and evaluation measures. In Section 4, we discuss the results of the official submissions. Finally, in Section 5 we end by summarizing and discussing the main findings of the task.

2. Related Work

In this section, we position the task in the broader context of related work.

To deal with a constant growing volume of scientific publications, concise overview, i.e. a summary, is needed. Automatic query-focused summarization may help users to familiarise with recent scientific achievement by presenting salient and query-relevant information from newly-published articles in a condensed manner [4]. Scientific texts are usually simplified by journalists (Nature[1], The Guardian[2], ScienceX[3]), researchers (Papier-Mâché[4], ScienceBites[5]), and internet forums (Explain Like I'm 5[6]). In contrast to the ANR projects CLEAR (medical texts simplification in French) [5, 6, 7, 8, 9, 10, 11, 12] and ALECTOR (Reading Aids to leverage Document Accessibility for Children with Dyslexia) [13, 14, 15, 16, 17], SimpleText tackles questions of information selection and provides background knowledge. While structured abstracts tend to be informative [18], often the information in a summary designed for an expert in the scientific domain is drastically different from that from a popularised version. Moreover, different levels of simplification, details, and explanation can be applied, e.g. Papier-Mâché publishes two levels of simplification: curiosity and advanced. Popular science articles are generally much shorter than scientific publications. Thus, summarization is a step to text simplification as it reduces the amount of information to be processed. Passage selection is a crucial but understudied task in document simplification [19], especially regarding the target audience [19], as existing works mainly focus on word/phrase-level [20] or sentence-level simplification [21].

Automatic query-biased summarization can simplify access to primary scientific documents; the resulting concise text is expected to highlight the most important parts of the document and thus reduces the reader's efforts. As the information in a summary designed for a scientist from a specific field should be different from that adapted for the general public, the main challenge is to choose which information from primary scientific sources should be included in a simplified text in order to remove barriers that non-expert users experience reading/accessing scientific information. Despite recent significant progress in the domains of information retrieval (IR) and natural language processing (NLP), the problem of constructing a consistent overview has

[1] https://www.nature.com/news
[2] https://www.theguardian.com/science
[3] https://sciencex.com/
[4] https://papiermachesciences.org/
[5] https://science-bits.com/
[6] https://www.reddit.com/r/explainlikeimfive/

not been solved yet [22]. Notice that this information extracted from primary scientific sources still needs to be further simplified and contextualised in order to be accessible for non-experts

3. SimpleText Task 1 Test Collection

This section provides an overview of the resulting test collection, detailing the corpus, the topics and queries, the exact input and output format used, as well as the exact relevance judgements and used evaluation measures.

3.1. Corpus

As in 2021, we use the Citation Network Dataset: DBLP+Citation, ACM Citation network (12th version) [23][7] as source of scientific documents that can be used as reference passages [24]. It contains: $4,894,083$ bibliographic references published before 2020, $4,232,520$ abstracts in English, $3,058,315$ authors with their affiliations, and $45,565,790$ ACM citations. From this corpus can be extracted textual content together with authorship. Although we manually preselected abstracts for topics, participants also have access to an ElasticSearch index; this index is adequate to passage retrieval using BM25.

The shared datasets provide: document abstract content for LDA (Latent Dirichlet Allocation) or Word Embedding (WE); document authors for coauthoring analysis; citation relationship between documents for co-citation analysis; citations by author for author impact factor analysis. These extra datasets are intended to be used to select passages by authors who are experts on the topic (highly cited by the community).

3.2. Topics

Topics are a selection of 40 press articles: 20 from *The Guardian*, a major international newspaper for a general audience with a tech section, and 20 from *Tech Xplore*[8], a Web site taking part in the Science X Network to provide a comprehensive coverage of engineering and technology advances. Each article was selected in the computer science field to be in accordance with the provided corpus. URLs to original articles, the title and textual content of each topic are provided to participants. Articles were also enriched with queries manually extracted from their content to provide an indication of the essential technical concepts covered. It has been manually checked that each query allows participants to retrieve from the corpus at least 5 relevant passages that could be inserted as citations in the press article. The use of these queries were optional.

Table 1 shows examples of the used Topics and Queries. The topic is represented by the title of the news article, and in addition the full text of the article was separately provided to use for the Task.

[7] https://www.aminer.cn/citation
[8] https://techxplore.com/

Table 1
CLEF 2022 SimpleText Task 1: Examples of Topics and Queries

Topic ID	Query ID	Title or Query
G12		*Patient data from GP surgeries sold to US companies*
	G12.1	`patient data`
G13		*Baffled by digital marketing? Find your way out of the maze*
	G13.1	`digital marketing`
	G13.2	`advertising`

Table 2
CLEF 2022 SimpleText Task 1 on content selection: example of output

Run	M/A	Topic	Query	Doc	Passage
ST1_task1_1	0	G01	G01.1	1564531496	A CDA is a mobile user device, similar to a Personal Digital Assistant (PDA). It supports the citizen when dealing with public authorities and proves his rights - if desired, even without revealing his identity.
ST1_task1_1	0	G01	G01.1	3000234933	People are becoming increasingly comfortable using Digital Assistants (DAs) to interact with services or connected objects
ST1_task1_1	0	G01	G01.2	1448624402	As extensive experimental research has shown individuals suffer from diverse biases in decision-making.

3.3. Formats

Results had to be provided in a JSON format (with a ".json" extension), and the following fields were required:

run_id Run ID starting with team ID, followed by "task1" and run name

manual Whether the run is manual {0,1}

topic_id Topic ID

query_id Query ID used to retrieve the document (if one of the queries provided for the topic was used; 0 otherwise)

doc_id ID of the retrieved document (to be extracted from the JSON output)

passage Text of the selected passage

Manual runs could also opt for a TREC style tabulated format (with a ".csv" extension), including headers with column names as the first line.

An example of the output is shown in Table 2. For each topic, the maximum number of distinct DBLP references (_id json field) was 100 and the total length of passages was not to exceed 1,000 tokens.

Table 3
Relevance score distribution per subset of topics

Topics	**0**	**1**	**2**	**3**	**4**	**5**	total
Guardian	100	83	43	66	57	36	376
Tech	0	8	15	61	9	6	99
Total	100	91	58	127	66	42	475

3.4. Evaluation Metrics

All passages retrieved from DBLP by participants are expected to have some overlap (lexical or semantic) with the article content.

To build a pooled test collection, we first extracted all the article IDs ranked by the number of participants who used the article to select passages. From this extraction, we only kept articles chosen by at least two participants and gave a relevance score on a scale of 0 to 5:

0 for irrelevant articles;
1 for marginally relevant articles;
2 when the abstract is relevant with the query;
3 when the abstract and keywords are relevant with the query;
4 when the abstract and keywords are relevant with the query and the topic (title of the original article);
5 when the abstract and keywords are relevant with the query and the extended topic (content of the original article).

In order to speed up the judgment process, for this edition we only evaluated relevance at the article level, and not at the sentence level. The abstract was considered as relevant as soon it has a sentence useful to explain the title or the original article.

A total of 475 documents have been assessed. Table 3 shows the score distributions.

For the documents returned by two runs, we had a high number of 1 and 2 scores for the Guardian topics. As regards the Tech Xplore topics, which have more technical queries since they deal with more technical and specific areas, queries were less ambivalent and more in keeping with the content of DBLP corpus. This has resulted in usually higher relevant scores, with many articles retrieved by two participants having a score of 3. Globally, whether the query comes from the Guardian or Tech Xplore, human evaluators found abstracts, among the articles retrieved by the participants from DBLP, that really explain the article or have matters which should have been addressed in the original article. Passages were often issued from publications that are more related to cognitive or information sciences than to technical fields, which shows that the DBLP corpus has expanded beyond computer science.

For example, among the 376 documents manually assessed for the Guardian topics, 92 appear to be highly relevant to expand the Guardian article (score above 4) and 36 reach a score of 5. Only three of them have been returned by all participants, all about CRISPR databases:

- “Anti-CRISPRdb: a comprehensive online resource for anti-CRISPR proteins” by the School of Life Science and Technology, University of Electronic Science and Technology of China

Table 4
CLEF 2022 SimpleText Task 1: Evaluation (graded measures)

Team	#Queries	Avg #Doc.	NDCG		
			5	10	20
CYUT	114	4.9	0.5866	0.5636	0.5536
UAMS	114	95.5	0.3531	0.3776	0.4073
UAMS-MF*	69	2.7	0.3494	0.3328	0.3270
NLP@IISERB 1	30	92.5	0.0605	0.0680	0.0819
NLP@IISERB 2	114	100	0.0503	0.0640	0.0815
NLP@IISERB 3	114	100	0.0467	0.0522	0.0722

* *Manual run.*

- "CRISPRcompar: a website to compare clustered regularly interspaced short palindromic repeats" by Paris-Sud University.
- "Application of Arrayed CRISPR/cas9 Screen and its Data Analysis: a Systematic Review" by Faculty of Health Sciences, University of Macau.

These three documents contain abstracts that do not require simplification and could have been quoted in the Guardian paper. However, none of these documents are highly cited by other DBLP documents. They appear to be marginal in the field of computer science.

4. SimpleText Task 1 Results

In this section we discuss the results for the official submissions to the Task 1.

After the "workshop" format of CLEF 2021 [24], this is the first year of SimpleText running as a track or CLEF lab. As expected in a first year, there is a steep learning curve and the attrition rate is high. Of the 62 teams who signed up for the track, many downloaded the data or explored the corpus with the provided Elastic Search API, but ultimately few teams managed to complete a run submission before the deadline.

A total of 3 teams submitted 6 runs: CYUT [1] submitted 1 run; IISERB [2] submitted 3 runs; and UAms [3] submitted 2 runs. A total of 4 automatic runs extracted 100 documents or abstracts per subquery, the CYUT automatic run extracted 5 sentences per subquery, and the UAms manual run extracted passages for a selection of subqueries. For evaluation, we consider here the reduced pool of documents returned by at least two runs; there are 72 queries with judgments, with a mean of 6.7 and a median of 4 judged documents per query.

Table 4 shows the number of queries with at least one returned document (*#Queries*), the average number of returned documents with a score ≥ 1 (*Avg #Docs*), and the evaluation against the graded relevance judgments (*NDCG*). We view NDCG@5 as the main metric for the official ranking on this task. These values show that the automatic run made by CYUT clearly outperforms the other runs in terms of selecting the abstracts with a high relevance.

Table 5 provides additional measures using a Boolean quantisation of relevance, at three different relevance thresholds. In the top part, we use the standard threshold of at least 1

Table 5
CLEF 2022 SimpleText Task 1: Evaluation (Boolean measures)

Team	Rel.	MRR	Precision			MAP
			5	10	20	
CYUT	1+	0.7111	0.5000	0.2500	0.1250	0.4987
UAMS	1+	0.5003	0.2917	0.1931	0.1333	0.3706
UAMS-MF*	1+	0.5289	0.2750	0.1375	0.0687	0.2813
NLP@IISERB 1	1+	0.1036	0.0667	0.0611	0.0556	0.0741
NLP@IISERB 2	1+	0.1118	0.0528	0.0514	0.0444	0.0548
NLP@IISERB 3	1+	0.1103	0.0556	0.0472	0.0444	0.0489
CYUT	2+	0.6139	0.4111	0.2056	0.1028	0.4499
UAMS	2+	0.4404	0.2417	0.1528	0.1028	0.3192
UAMS-MF*	2+	0.4537	0.2417	0.1208	0.0604	0.2599
NLP@IISERB 1	2+	0.0990	0.0528	0.0472	0.0437	0.0673
NLP@IISERB 2	2+	0.1031	0.0444	0.0431	0.0368	0.0539
NLP@IISERB 3	2+	0.1041	0.0528	0.0417	0.0354	0.0508
CYUT	4+	0.2697	0.1167	0.0583	0.0292	0.1967
UAMS	4+	0.2048	0.0778	0.0458	0.0333	0.1580
UAMS-MF*	4+	0.2118	0.0889	0.0444	0.0222	0.1504
NLP@IISERB 1	4+	0.0483	0.0222	0.0222	0.0229	0.0300
NLP@IISERB 2	4+	0.0453	0.0139	0.0167	0.0174	0.0302
NLP@IISERB 3	4+	0.0604	0.0222	0.0181	0.0188	0.0304

* *Manual run.*

("marginal relevant"), and observe high early precision scores, and solid average precision scores for the top runs. In the middle part, we restrict the evaluation to at least 2 ("relevant") leading unavoidably to lower numbers, and relatively better performance of the manual run. In the bottom part, we only restrict the evaluation to abstracts relevant to the context article motivating the query, and observe that the top runs are still able to achieve a reasonable mean reciprocal ranks of the first retrieved highly relevant article, and even solid mean average precision over this small set of these most desirable abstracts.

We led further analysis on two subsets of topics with a higher level of assessment. From the 31 topics with at least one assessment in the pooled test collection, we first kept only the 16 topics with at least 10 evaluated documents (G01, G03, G04, G05, G07, G08, G09, G10, G15, G17, G19, T02, T04, T05, T10, T16). The upper part of Table 6 exhibits NDCGs measured on this first subset, with the same ranking of runs and values similar to the ones observed on the whole data collection. Then, we kept only the 6 topics with at least 20 assessed documents (G03, G04, G05, G07, G08, G17). The lower part of Table 6 shows that on these easiest topics, all the systems were able to retrieve relevant documents, which results in a dramatic increase of NDCG values for NLP@IISERB runs, with the first run approaching the result quality of the CYUT run.

Table 6
Runs score in task 1 on topics with ≥ 10 or ≥ 20 assessments

Team	#Topics	#Queries	NDCG		
			5	**10**	**20**
CYUT	16	45	0.5855	0.5478	0.5344
UAMS	16	45	0.3463	0.3733	0.4058
UAMS-MF*	15	33	0.3738	0.3473	0.3376
NLP@IISERB 1	10	20	0.0780	0.0906	0.1101
NLP@IISERB 2	16	45	0.0605	0.0732	0.0864
NLP@IISERB 3	16	45	0.0651	0.0634	0.0796
CYUT	6	15	0.4735	0.3730	0.3317
UAMS	6	15	0.2211	0.2326	0.2742
UAMS-MF*	6	10	0.2668	0.1990	0.1692
NLP@IISERB 1	6	15	0.2364	0.2755	0.3025
NLP@IISERB 2	6	15	0.1035	0.2755	0.1726
NLP@IISERB 3	6	15	0.1644	0.2755	0.1759

* *Manual run.*

5. Conclusion

This paper presented an overview of the CLEF 2022 SimpleText Task 1, on retrieving relevant abstracts in response to popular science request.

During CLEF 2022, we developed an impressive test collection, consisting of a huge corpus of over 4 million scientific abstracts, a set of 40 topics in the form of popular science news articles, a set of 114 queries motivated by these articles, and relevance judgments for 72 of these queries and 31 topics.

A large number of teams signed up for the track, and downloaded the data or explored the collection using the elastic search API made available by the organisers. However, building an effective system is challenging and we received six submissions from three teams who managed to create interesting approaches to tackle this important problem.

We presented the evaluation results of the official run submissions, and found that some teams obtain very competitive performance. Using graded measures reflecting the relevance of the retrieved abstracts to the motivating context of the popular science article, the top runs obtain a high NDCG@5 of over 0.5. These results are very promising and we hope and expect that next year, with the extensive data from 2022 available, this leads to the further development of novel retrieval models and approaches for this important problem.

Recall that the main overall goal of the track, and CLEF in general, is to support and promote research by building corpora and test collections. We can look back with gratitude on the construction of the CLEF 2022 SimpleText Task 1 test collection, and particularly thank the help of active participants by submitting runs that created the pool of abstracts leading to the ultimate relevance judgments. We hope and expect that the resulting test collection will be used and reused by many other researchers in IR, NLP, and AI, who together can make a difference in solving the important societal problem addressed by the track.

Acknowledgments

We thank all the participants who expressed interest in the task by registering for the track, and even more those participants who downloaded all the data, and even *more those who managed to submit runs in time for evaluation.*

References

[1] S.-H. Wu, H.-Y. Huang, CYUT Team2 SimpleText Shared Task Report in CLEF-2022, in: [25], 2022.

[2] 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: [25], 2022.

[3] F. Mostert, A. Sampatsing, M. Spronk, D. Rau, J. Kamps, University of Amsterdam at the CLEF 2022 SimpleText Track, in: [25], 2022.

[4] A. Sarker, Y.-C. Yang, M. A. Al-Garadi, A. Abbas, A Light-Weight Text Summarization System for Fast Access to Medical Evidence, Frontiers in Digital Health 2 (2020). URL: https://www.frontiersin.org/article/10.3389/fdgth.2020.585559.

[5] N. Grabar, E. Farce, L. Sparrow, Study of readability of health documents with eye-tracking approaches, in: 1st Workshop on Automatic Text Adaptation (ATA), 2018.

[6] N. Grabar, T. Hamon, A large rated lexicon with French medical words, in: LREC (Language Resources and Evaluation Conference), 2016.

[7] N. Grabar, R. Cardon, CLEAR-Simple Corpus for Medical French, 2018. URL: https://halshs.archives-ouvertes.fr/halshs-01968355.

[8] R. Cardon, N. Grabar, French Biomedical Text Simplification: When Small and Precise Helps, in: Proceedings of the 28th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 710–716. URL: https://www.aclweb.org/anthology/2020.coling-main.62. doi:10.18653/v1/2020.coling-main.62.

[9] R. Cardon, N. Grabar, Recherche de phrases parallèles à partir de corpus comparables pour la simplification de textes médicaux en français, in: Actes des Ateliers d'INFORSID - Dessinons ensemble le futur des systèmes d'information, 2021, pp. 61–63. URL: http://inforsid.fr/actes/2021/ActesAteliers_INFORSID2021.pdf#page=63.

[10] A. Koptient, N. Grabar, Fine-grained text simplification in French: steps towards a better grammaticality, in: ISHIMR Proceedings of the 18th International Symposium on Health Information Management Research, Kalmar, Sweden, 2020. URL: https://hal.archives-ouvertes.fr/hal-03095247. doi:10.15626/ishimr.2020.xxx.

[11] A. Koptient, N. Grabar, Rated Lexicon for the Simplification of Medical Texts, in: The Fifth International Conference on Informatics and Assistive Technologies for Health-Care, Medical Support and Wellbeing HEALTHINFO 2020, Porto, Portugal, 2020. URL: https://hal.archives-ouvertes.fr/hal-03095275.

[12] A. Koptient, N. Grabar, Typologie de transformations dans la simplification de textes,

in: Congrès mondial de la linguistique française, Montpellier, France, 2020. URL: https://hal.archives-ouvertes.fr/hal-03095235.

[13] N. Gala, A. Tack, L. Javourey-Drevet, T. François, J. C. Ziegler, Alector: A parallel corpus of simplified French texts with alignments of misreadings by poor and dyslexic readers, in: Language Resources and Evaluation for Language Technologies (LREC), 2020.

[14] N. Gala, T. François, C. Fairon, Towards a french lexicon with difficulty measures: Nlp helping to bridge the gap between traditional dictionaries and specialized lexicons, in: eLex-Electronic Lexicography, 2013.

[15] M. Billami, T. François, N. Gala, ReSyf: a French lexicon with ranked synonyms, in: 27th International Conference on Computational Linguistics (COLING 2018), 2018.

[16] T. François, N. Gala, P. Watrin, C. Fairon, FLELex: a graded lexical resource for French foreign learners, in: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), 2014, pp. 3766–3773.

[17] L. Sauvan, N. Stolowy, C. Aguilar, T. François, N. Gala, F. Matonti, E. Castet, A. Calabrese, Text simplification to help individuals with low vision read more fluently, in: Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI), 2020, pp. 27–32.

[18] L. Ermakova, F. Bordignon, N. Turenne, M. Noel, Is the Abstract a Mere Teaser? Evaluating Generosity of Article Abstracts in the Environmental Sciences, Frontiers in Research Metrics and Analytics 3 (2018). URL: https://www.frontiersin.org/articles/10.3389/frma.2018.00016/full. doi:10.3389/frma.2018.00016.

[19] Y. Zhong, C. Jiang, W. Xu, J. J. Li, Discourse Level Factors for Sentence Deletion in Text Simplification, Proceedings of the AAAI Conference on Artificial Intelligence 34 (2020) 9709–9716. URL: https://ojs.aaai.org/index.php/AAAI/article/view/6520. doi:10.1609/aaai.v34i05.6520, number: 05.

[20] M. Maddela, W. Xu, A Word-Complexity Lexicon and A Neural Readability Ranking Model for Lexical Simplification, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 3749–3760. URL: https://www.aclweb.org/anthology/D18-1410. doi:10.18653/v1/D18-1410.

[21] Y. Dong, Z. Li, M. Rezagholizadeh, J. C. K. Cheung, EditNTS: An Neural Programmer-Interpreter Model for Sentence Simplification through Explicit Editing, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 3393–3402. URL: https://www.aclweb.org/anthology/P19-1331. doi:10.18653/v1/P19-1331.

[22] L. Ermakova, J. V. Cossu, J. Mothe, A survey on evaluation of summarization methods, Information Processing & Management 56 (2019) 1794–1814. URL: http://www.sciencedirect.com/science/article/pii/S0306457318306241. doi:10.1016/j.ipm.2019.04.001.

[23] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, Z. Su, Arnetminer: Extraction and mining of academic social networks, in: KDD'08, 2008, pp. 990–998.

[24] 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 Multilin-

guality, Multimodality, and Interaction, Lecture Notes in Computer Science, Springer International Publishing, Cham, 2021, pp. 432–449. doi:10.1007/978-3-030-85251-1_27.
[25] G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), Proceedings of the Working Notes of CLEF 2022: Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS.org, 2022.