UNIPD@SimpleText2024: A Semi-Manual Approach on Prompting ChatGPT for Extracting Terms and Write Terminological Definitions

Notebook for the SimpleText Lab at CLEF 2024

Giorgio Maria Di Nunzio[1], Elena Gallina[2] and Federica Vezzani[2]

[1]Department of Information Engineering, University of Padova, Italy
[2]Department of Linguistic and Literary Studies, University of Padova, Italy

Abstract

In this experimental work, we explore Task 2 of the SimpleText Lab, which aims to enhance text simplification technologies using manually annotated datasets. The objective of this work is to propose a methodology for evaluating the capability of Large Language Models to identify and explain difficult terms through optimal prompting. Additionally, we assess improvements by manually correcting the extracted terms and definitions, aiming to refine and advance the utility of text simplification tools for broader applications.

Keywords

Text Simplification, Automatic Term Extraction, Terminological Definition

1. Introduction

The SimpleText Lab[1] is an initiative that aims to address the challenge of text simplification, which involves modifying complex text to make it easier to read and understand while retaining the original meaning [1]. The primary objectives of this Lab are to improve the accessibility and readability of text for various audiences, including individuals with cognitive disabilities, language learners, and children. In particular, the Lab encourages the following initiatives:

- Creating and refining models that can effectively simplify text by making it more accessible without losing essential information;
- Establishing benchmarks and evaluation metrics to assess the performance of text simplification models;
- Engaging the research community to participate in text simplification tasks, thereby fostering innovation and collaboration.

The SimpletText Lab 2024 edition proposed four tasks: 1) Retrieving passages to include in a simplified summary [2], 2) Identifying and explaining difficult concepts [3], 3) Simplify Scientific Text [4], 4) Tracking the State-of-the-Art in Scholarly Publications [5].

In this experimental work, we focus on Task 2 which aims to study current text simplification technologies by testing them on practical, diverse datasets and to foster advancements that could lead to more effective and widely usable simplification tools. In particular, Task 2 have these specific objectives: i) applying text simplification to real-world texts, which could include news articles, educational materials, or other publicly available documents; ii) ensuring that the simplified text adheres to specific criteria such as readability, preservation of meaning, and grammatical correctness; iii) implementing a framework to evaluate the effectiveness of the simplification process based on parameters like fluency, adequacy, and simplicity.

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France
giorgiomaria.dinunzio@unipd.it (G. Di Nunzio); elena.gallina.6@studenti.unipd.it (E. Gallina); federica.vezzani@unipd.it (F. Vezzani)
0000-0001-9709-6392 (G. Di Nunzio); 0000-0003-2240-6127 (F. Vezzani)

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

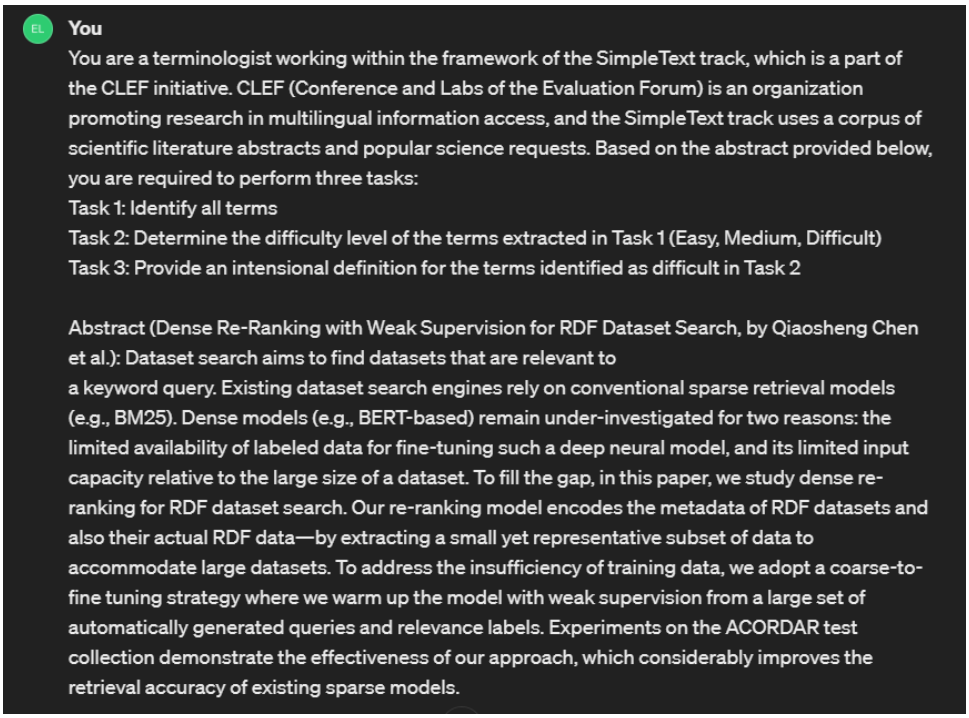

Figure 1: Example of initial prompt to set the experiment.

Our participation to this task has the objective to study the capability of a Large Language Model to extract difficult terms and build terminological definition to explain those terms with the right prompt. In addition, we also want to evaluate the improvement (if any) of the initial results with a manual correction of the extracted terms and the provided definitions.

2. Methodology

Our participation to Task 2 focuses on identifying and explaining difficult content using Large Language Models (LLMs) to enhance text simplification. The methodology involves iterative experimentation with various prompting strategies to optimize the performance of the model in this task. The methodology that we designed with the help of a Master Student in Translation-oriented Terminography followed these steps:

- Initially analyze a diverse set of complex texts to identify common linguistic and contextual difficulties.
- Design and test a series of prompts to guide the LLM not only to detect these difficult sections but also to provide clear and concise explanations or simplifications.
- Refining prompts based on feedback and evaluation metrics like readability, clarity, and fidelity to the original meaning.

3. Experimental Setting

In order to find the most suitable prompt to submit to ChatGPT 3.5 (April 15 2024 is the time the experiment was performed) we followed the procedure presented in the previous section. In particular, we started by analyzing the abstract of the paper [6] and started trying different prompts in to obtain an output that performed tasks relating to terminology extraction, identification of the level of difficulty of each term and the formulation of definitions for those considered difficult.

An example of the first prompt is shown in Figure 1 (initial prompt) and Figure 3 (output).

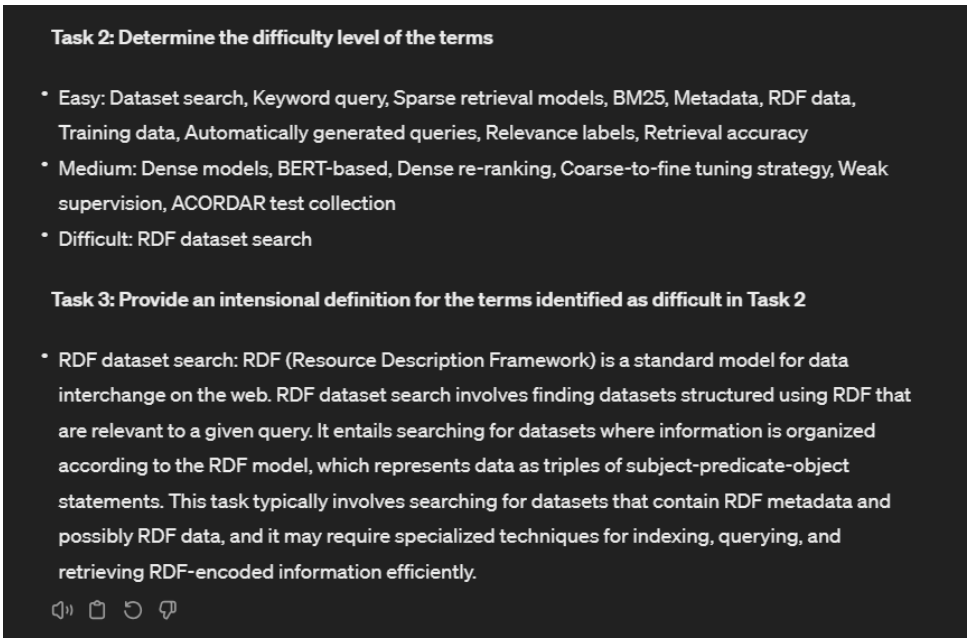

Figure 2: Example of output relative to the initial prompt shown in Figure 1.

A second and third attempt of the prompt was necessary to be more precise in the request: for this reason, we added two brief definitions of "term" and "intensional definition" (according to ISO 1087: 2019, intensional definition "conveys the intension of a concept by stating the immediate generic concept and the delimiting characteristic(s)") were included in the input and explicitly mention the fact that the evaluation of the difficulty of each term should be performed as the user is a general public user.

The output produced by ChatGPT maintained the same terms extracted in the previous attempt (the second one, not shown here in the figures), while adding "coarse-to-fine tuning strategy" to the terms considered difficult to understand. As already seen in the second attempt, the definitions provided contain elements of the intensional definition (superordinate concept and delimiting characteristics) in the first part of the output related to subtask 2.3 (building definitions), while the second part contains a further explanation aimed at deepening the terms analyzed. The results are shown in Figure 3.

After this preliminary analysis to tune the right prompt, we run the same prompt on each abstract of the dataset and collected all the extracted terms, their difficulty, and the intensional definition.

We produced three runs:

- "unipd_t21t22_chatgpt" contains the ChatGPT output without any modification;
- "unipd_t21t22_chatgpt_mod1" contains the output of the original runs contains minus the elements that we do not consider as terms (so the only operation we did was to eliminate elements from the original run);
- "unipd_t21t22_chatgpt_mod2" contains additional manual corrections like:
  - remove partial/not meaningful multi-word terms;
  - for situations like "body mass (BM)", we separate "body mass" and "BM" into two entries;
  - incomplete terms are completed;
  - terms assigned to an incorrect sentence are reassigned to the correct sentence.

In addition, we created a non-official run (that was not submitted to this Task) completely manual named:

- "unipd_t21t22_manual".

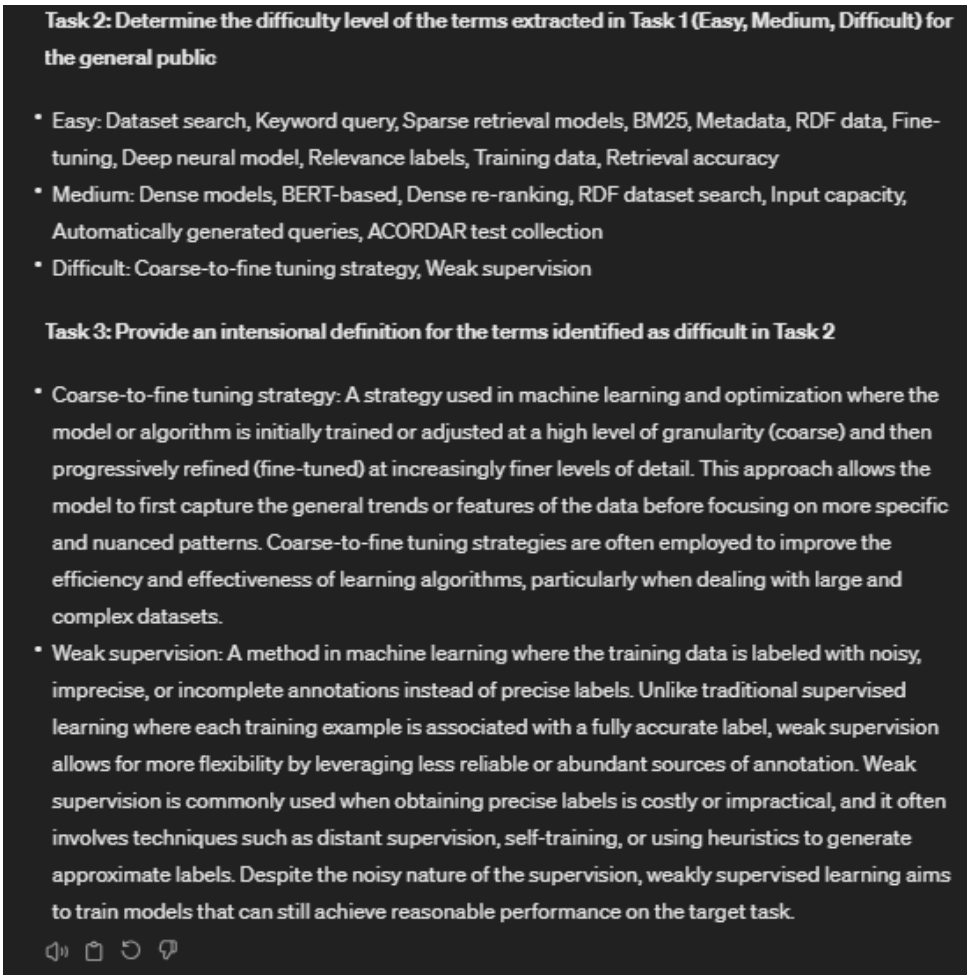

Figure 3: Example of output relative to the improved prompts.

4. Results

In this section, we present a summary of the quantitative results obtained with the three official runs plus the one additional manual run that was prepared afterwards. For all the runs, we have the following information:

- name of the run;
- recall overall: the proportion of terms (independently from the difficulty) that were found;
- precision overall: the proportion of terms (independently from the difficulty) correctly categorized as terms;
- f1 overall score;
- recall average: the average of the recall of terms computed per sentence;
- precision average: the average of the recall of terms computed per sentence;
- f1 average score;
- recall difficult terms: the proportion of difficult terms that were found;
- precision difficult: the precision of terms that were labeled as difficult;
- f1 difficult overall score;
- recall difficult terms: the proportion of difficult terms that were found;
- precision difficult: the precision of terms that were labeled as difficult;
- f1 difficult average score;
- bleu_nx: the BLEU score computed with ngrams x = 1, 2, 3, 4.

Table 1
Summary of results of Task 2 for all terms.

runid	recall overall	precision overall	f1 overall	recall average	precision average	f1 average
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
unipt_t21t22_manual	0.545	0.469	0.504	0.541	0.574	0.504
median score for all the runs in the task	0.109	0.561	0.186	0.121	0.568	0.186

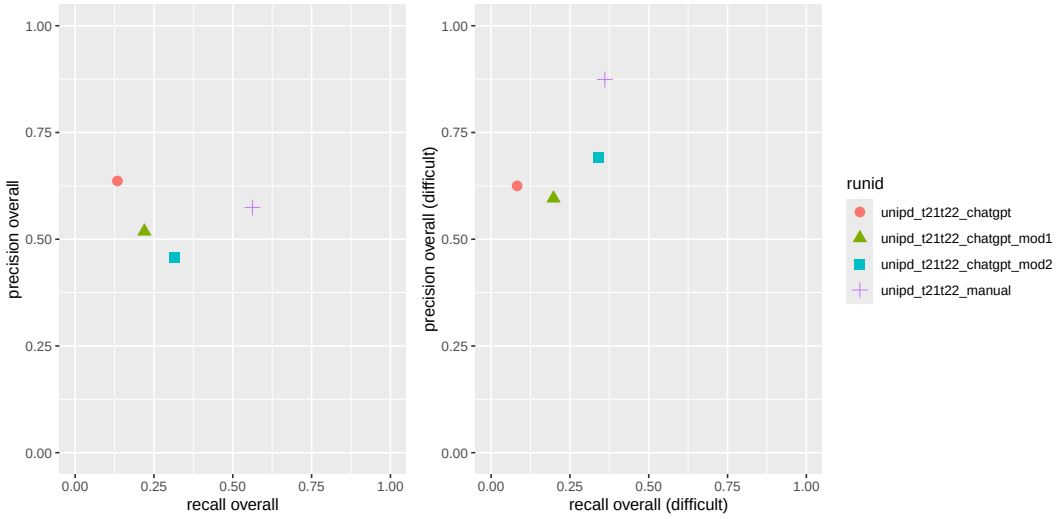

Figure 4: Summary of results of Task 2. Recall-Precision overall scores.

In particular: in Table 1, we show the overall results for the term extraction independently from the difficulty of the term. In Table 2, we show the overall results for the term extraction only for difficult terms. In Table 3, we present the scores of the BLEU measure for the provided definitions. In Figure 4, we show the recall-precision plot of the overall scores and the scores averaged per sentence for all the terms; in Figure 5, the same information for difficult terms only; in Figure 6, we display the BLEU score, for n = 1 and n = 2, in relation to the f1 value for difficult terms.

For almost all the results, we can see that the performance of all the run is much better than the median values for the task for both the extraction of terms and the generation of definitions. Manual interventions on the output of ChatGPT are beneficial, as expected, in particular in creasing the recall maintaining a high precision in the extraction of the terms. The fully manual run, has shown the best recall but a slightly worse performance for what concerns precision across all the terms. On the other hand, the manual correction of definitions has not improved the BLAU score significantly.

5. Final Considerations

In this paper, we described the methodology and the experiments submitted to the SimpleText Lab for Task 2 which is about identifying and explaining difficult concepts. The objective of this work was to analyze the performance of a Large Language Model, specifically ChatGPT 3.5, in extracting terms, evaluate their difficulty, and create intensional definitions to explain the difficult terms. The preliminary

Table 2
Summary of results of Task 2 for the difficult terms.

runid	recall overall (difficult)	precision overall (difficult)	f1 overall (difficult)	recall average (difficult)	precision average (difficult)	f1 average (difficult)
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
unipt_t21t22_manual	0.364	0.904	0.519	0.324	0.953	0.484
median score for all the runs in the task	0.091	0.563	0.157	0.089	0.979	0.163

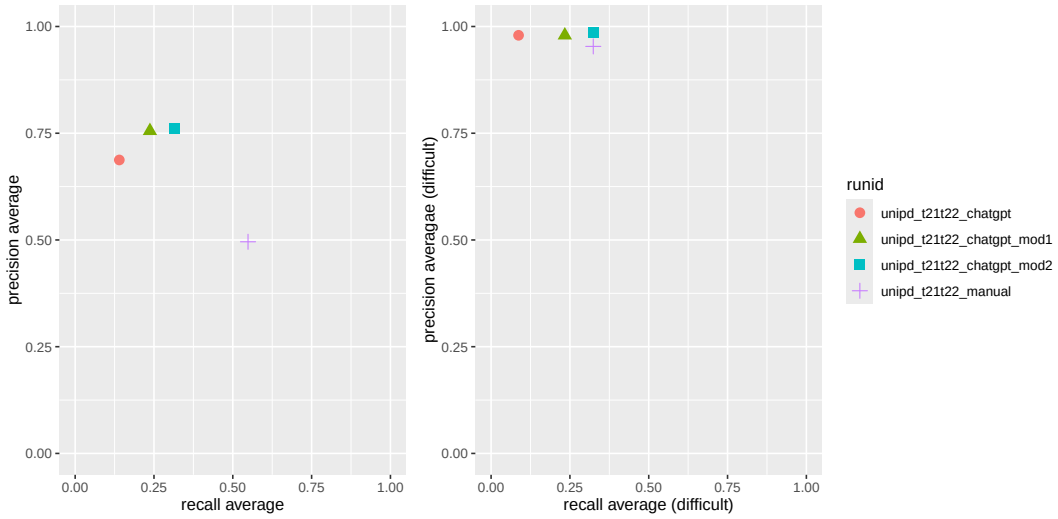

Figure 5: Summary of results of Task 2. Recall-Precision average scores.

results show that the performance of a fully automated approach of this methodology is an initial step but it is far from being useful in terms of proportion of terms found. Additional manual corrections to remove and correct the terms can improve the performance. This suggests that there are some steps that can be formalized in an automatic way (for example, splitting proposed terms that contain both the correct term and the actual acronym) in order to obtain a better system.

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 work is also part of the initiatives carried out by the Center for Studies in Computational Terminology (CENTRICO) of the University of Padua and in the research directions of the Italian Common Language Resources and Technology Infrastructure CLARIN-IT.

Table 3
Summary of results of Task 2. BLEU scores.

runid	bleu n1 average	bleu n2 average	bleu n3 average	bleu n4 average
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	0.294	0.184	0.091	0.052
unipt_t21t22_manual	0.299	0.189	0.095	0.054
median score for all the runs in the task	0.258	0.143	0.045	0.021

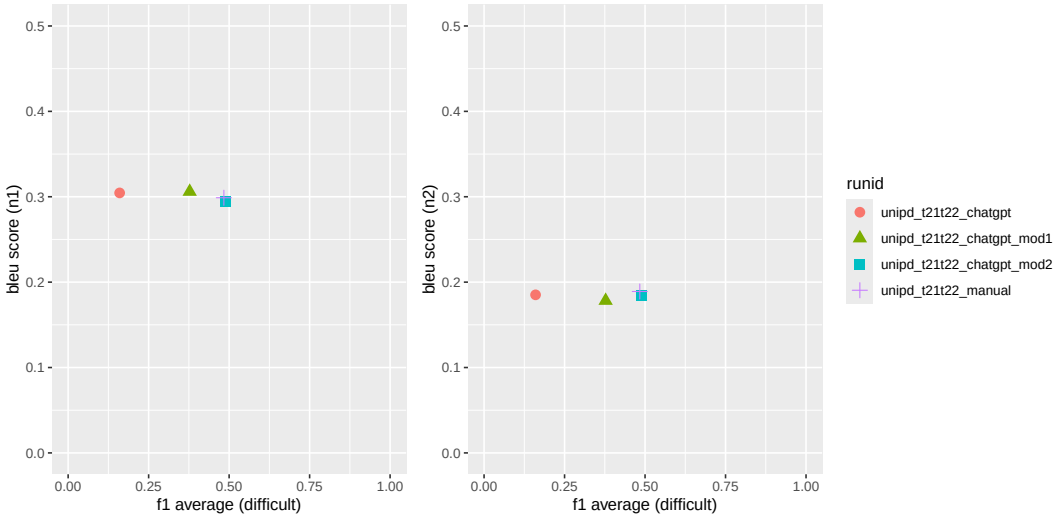

Figure 6: Summary of results of Task 2. BLEU score compared to f1 scores on difficult terms.

References

[1] 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.

[2] E. SanJuan, S. Huet, J. Kamps, L. Ermakova, Overview of the CLEF 2023 simpletext task 1: Passage selection for a simplified summary, in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), Thessaloniki, Greece, September 18th to 21st, 2023, volume 3497 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 2823–2834. URL: https://ceur-ws.org/Vol-3497/paper-238.pdf.

[3] G. M. Di Nunzio, F. Vezzani, V. Bonato, H. Azarbonyad, , J. Kamps, L. Ermakova, Overview of the CLEF 2024 SimpleText task 2: Identify and explain difficult concepts, in: [7], 2024.

[4] L. Ermakova, V. Laimé, H. McCombie, J. Kamps, Overview of the CLEF 2024 SimpleText task 3: Simplify scientific text, in: [7], 2024.

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

[6] Q. Chen, Z. Huang, Z. Zhang, W. Luo, T. Lin, Q. Shi, G. Cheng, Dense Re-Ranking with Weak Supervision for RDF Dataset Search, in: T. R. Payne, V. Presutti, G. Qi, M. Poveda-Villalón, G. Stoilos, L. Hollink, Z. Kaoudi, G. Cheng, J. Li (Eds.), The Semantic Web – ISWC 2023, Springer Nature Switzerland, Cham, 2023, pp. 23–40. doi:10.1007/978-3-031-47240-4_2.

[7] 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.