Augmented Intelligence in Technology-Assisted Review Systems (ALTARS 2022): Evaluation Metrics and Protocols for eDiscovery and Systematic Review Systems

Giorgio Maria Di Nunzio[1][0000−0001−9709−6392], Evangelos Kanoulas[2][0000−0002−8312−0694], and Prasenjit Majumder[3]

[1] Department of Information Engineering
University of Padova, Italy
giorgiomaria.dinunzio@unipd.it
[2] Faculty of Science, Informatics Institute
University of Amsterdam, The Netherlands
E.Kanoulas@uva.nl
[3] DAIICT, Gandhinagar, India
prasenjit_t@isical.ac.in

Abstract. In this workshop, we aim to fathom the effectiveness of Technology-Assisted Review Systems from different viewpoints. In fact, despite the number of evaluation measures at our disposal to assess the effectiveness of a "traditional" retrieval approach, there are additional dimensions of evaluation for these systems. For example, it is true that an effective high-recall system should be able to find the majority of relevant documents using the least number of assessments. However, this kind of evaluation usually discards the resources used to achieve this goal, such as the total time spent on those assessments, or the amount of money spent for the experts judging the documents.

Keywords: Technology-Assisted Review Systems · Augmented Intelligence · Evaluation · Systematic Reviews · eDiscovery

1 Motivations

Augmented Intelligence is "a subsection of AI machine learning developed to enhance human intelligence rather than operate independently of or outright replace it. It is designed to do so by improving human decision-making and, by extension, actions taken in response to improved decisions."[4] In this sense, users are supported, not replaced, in the decision-making process by the filtering capabilities of the Augmented Intelligence solutions, but the final decision will always be taken by the users who are still accountable for their actions.

Given these premises, we focus on High-recall Information Retrieval (IR) systems which tackle challenging tasks that require the finding of (nearly) all

[4] https://digitalreality.ieee.org/publications/what-is-augmented-intelligence

the relevant documents in a collection. Electronic discovery (eDiscovery) and systematic review systems are probably the most important examples of such systems where the search for relevant information with limited resources, such as time and money, is necessary.

In this field, Technology-assisted review (TAR) systems use a kind of human-in-the-loop approach where classification and/or ranking algorithms are continuously trained according to the relevance feedback from expert reviewers, until a substantial number of the relevant documents are identified. This approach, named Continuous Active Learning (CAL), has been shown to be more effective and more efficient than traditional e-discovery and systematic review practices, which typically consists of a mix of keyword search and manual review of the search results.

In order to achieve high recall values, machine-learning methods need large numbers of human relevance assessments which represent the primary cost of such methods. It is therefore necessary to evaluate these systems not only in terms of "batch"/off-line performances, but also in terms of the time spent per assessment, the hourly pay rate for assessors, and the quality of the assessor. For example, by reducing the amount of work by using sentence-level assessments in place of document-level assessments to reduce the time to read the document and the number of judgments needed. In addition, it would be also necessary to include in the validation of the system the feedback of the users by asking direct questions about the information carried in the missing documents instead of just asking about their relevance[2, 5–7].

In the context of High Recall Information Retrieval Systems, we believe that it is necessary to compare 1) the vetting approach that use evaluation collections to optimize systems and carry out pre-hoc evaluation, 2) the validation of the system to measure the actual outcome of the system in real situations.

2 Topics of Interest

In this workshop, we aim to fathom the effectiveness of these systems which is a research challenge itself. In fact, despite the number of evaluation measures at our disposal to assess the effectiveness of a "traditional" retrieval approach, there are additional dimensions of evaluation for TAR systems. For example, it is true that an effective high-recall system should be able to find the majority of relevant documents using the least number of assessments. However, this type of evaluation discards the resources used to achieve this goal, such as the total time spent on those assessments, or the amount of money spent for the experts judging the documents.

The topics of the workshop are:

- Novel evaluation approaches and measures for e-Discovery;
- Novel evaluation approaches and measures for Systematic reviews;
- Reproducibility of experiments with test collections;
- Design and evaluation of interactive high-recall retrieval systems;
- Study of evaluation measures;

– User studies in high-recall retrieval systems;
– Novel evaluation protocols for Continuous Active Learning;
– Evaluation of sampling bias.

3 Organizing Team

Giorgio Maria Di Nunzio is Associate Professor at the Department of Information Engineering of the Universiryt of Padova. He has been the co-organizer of the ongoing Covid-19 Multilingual Information Access Evaluation forum,[5] in particular for the evaluation of high-recall systems and high-precision systems tasks. He will bring to this workshop the perspective of alternative (to the standard) evaluation measures and multilingual challenges.

Evangelos Kanoulas is Full Professor at the Faculty of Science of the Informatics Institute at the University of Amsterdam. He has been the co-organizer CLEF eHealth Lab and of the Technologically Assisted Reviews in Empirical Medicine task.[6] He will bring to the workshop the perspective of the evaluation of the costs in eHealth TAR systems, in particular of the early stopping strategies.

Prasenjit Majumder is Associate Professor at the Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), Gandhinagar and TCG CREST, Kolkata, India. He has been the co-organizer of the Forum for Information Retrieval Evaluation and, in particular, the Artificial Intelligence for Legal Assistance (AILA) task.[7] He will bring to the workshop the perspective of the evaluation of the costs of eDiscovery, in particular of the issues related to legal precedence findings.

All the three organizing committee members have been active participants in the past editions of the TREC , CLEF and FIRE evaluation forum for the Total Recall and Precision Medicine TREC Tasks, TAR in eHealth tasks, and AI for Legal Assistance.[8][9][10] The committee members have strong research record with a total of more than 400 papers in international journals and conferences. They and have been doing research in technology assisted review systems and problems related to document distillation both in the eHealth and eDiscovery domain and made significant contributions in this specific research area [1, 4, 3].

[5] http://eval.covid19-mlia.eu
[6] https://clefehealth.imag.fr
[7] https://sites.google.com/view/aila-2021
[8] https://scholar.google.it/citations?user=Awl_HDoAAAAJ
[9] https://scholar.google.com/citations?user=0HybxV4AAAAJ
[10] https://scholar.google.co.in/citations?user=3xIpiKEAAAAJ

References

1. Di Nunzio, G.: A study on a stopping strategy for systematic reviews based on a distributed effort approach. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction - 11th International Conference of the CLEF Association, CLEF 2020, Thessaloniki, Greece, September 22-25, 2020, Proceedings. pp. 112–123 (2020). https://doi.org/10.1007/978-3-030-58219-7_10, https://doi.org/10.1007/978-3-030-58219-7_10
2. Lewis, D.D., Yang, E., Frieder, O.: Certifying one-phase technology-assisted reviews. In: CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 - 5, 2021. pp. 893–902 (2021). https://doi.org/10.1145/3459637.3482415, https://doi.org/10.1145/3459637.3482415
3. Li, D., Kanoulas, E.: When to stop reviewing in technology-assisted reviews: Sampling from an adaptive distribution to estimate residual relevant documents. ACM Trans. Inf. Syst. **38**(4), 41:1–41:36 (2020). https://doi.org/10.1145/3411755, https://doi.org/10.1145/3411755
4. Mehta, P., Mandl, T., Majumder, P., Gangopadhyay, S.: Report on the FIRE 2020 evaluation initiative. SIGIR Forum **55**(1), 3:1–3:11 (2021). https://doi.org/10.1145/3476415.3476418, https://doi.org/10.1145/3476415.3476418
5. Pickens, J., III, T.C.G.: On the effectiveness of portable models versus human expertise under continuous active learning. In: Joint Proceedings of the Workshops on Automated Semantic Analysis of Information in Legal Text (ASAIL 2021) & AI and Intelligent Assistance for Legal Professionals in the Digital Workplace (LegalAIIA 2021) held online in conjunction with 18th International Conference on Artificial Intelligence and Law (ICAIL 2021), São Paolo, Brazil (held online), June 21 & 25, 2021. pp. 69–76 (2021), http://ceur-ws.org/Vol-2888/paper10.pdf
6. Sneyd, A., Stevenson, M.: Stopping criteria for technology assisted reviews based on counting processes. In: SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021. pp. 2293–2297 (2021). https://doi.org/10.1145/3404835.3463013, https://doi.org/10.1145/3404835.3463013
7. Zhang, H., Cormack, G.V., Grossman, M.R., Smucker, M.D.: Evaluating sentence-level relevance feedback for high-recall information retrieval. Inf. Retr. J. **23**(1), 1–26 (2020). https://doi.org/10.1007/s10791-019-09361-0, https://doi.org/10.1007/s10791-019-09361-0