WORKSHOP REPORT

Report on EVIA 2017: 8th International Workshop on Evaluating Information Access

Nicola Ferro
University of Padua, Italy
ferro@dei.unipd.it

Ian Soboroff
National Institute of Standards and Technology (NIST), USA
ian.soboroff@nist.gov

Abstract

This is a report on the eighth edition of the *International Workshop on Evaluating Information Access* (EVIA 2017), co-located with the 13th NTCIR Conference on the Evaluation of Information Access Technologies (NTCIR13) held in Tokyo, Japan, on December 5, 2017.

1 Motivations and Goals

EVIA 2017, the 8th International Workshop on Evaluating Information Access[1], was co-located with the 13th NTCIR Conference on the Evaluation of Information Access Technologies (NTCIR13) held in Tokyo, Japan, on December 5, 2017.

Information Access technologies provide the interface between human information needs and digital information resources. The reliable evaluation of these technologies has been recognized for decades as central to the advancement of the field. As information retrieval technologies become more pervasive, the forms of retrieval more diverse, and retrieval tools richer, the importance of effective, efficient, and innovative evaluation grows as well.

The goal of the workshop was to investigate how to improve information access evaluation, by bringing in new perspectives which have not been explored or fully addressed yet. Therefore, the workshop solicited the submission of contributions covering new approaches for: test collection formation; evaluation metrics; statistical issues in information retrieval evaluation; user studies and the evaluation of human-computer interaction in information retrieval (HCIR); evaluation methods for multilingual, multimedia, or mobile information access; novel information access tasks and their evaluation; evaluation and assessment using implicit user feedback, crowdsourcing, living labs, or inferential methods; evaluation issues in industrial and enterprise retrieval systems; and, reproducibility issues in information retrieval evaluation.

[1] `http://research.nii.ac.jp/ntcir/evia2017/`

The workshop received 10 submissions out of which 7 were accepted for publication and presentation: 4 full papers and 3 short papers; Section 3 provides a short summary of the presented papers. The proceedings of EVIA 2017 [2] have been published for the first time in the CEUR-WS proceedings series to spread their diffusion and ease their permanent archiving.

EVIA 2017 also featured a remarkable keynote talk, summarized in Section 2, by a leading scientist in information retrieval and multilingual information access – prof. Doug Oard at University of Maryland, USA.

The workshop enjoyed an audience of about 30 participants, who actively participated to the discussions fostered by the paper presentations and the keynote talk.

2 Keynote

Cross-Language Information Retrieval in the MATERIAL Program

Prof. Doug Oard described a research program called MAchine Translation for English Retrieval of Information in Any Language (MATERIAL) that includes a substantial focus on Cross Language Information Retrieval (CLIR). Over four years, this program expects to build new CLIR test collections for ten new languages, in each case with English queries. Novel aspects of these test collections will include (1) domain-limited, sense-specific, and morphology-specific queries, and (2) mixed collections including both text and speech. Two novel aspects of the evaluation design are a focus on set-based rather than ranked retrieval, and the use of a linear utility measure for evaluating result set selection. The MATERIAL program also includes an interactive CLIR evaluation in which assessors use system-generated English summaries in an effort to identify the truly relevant documents in the result set. Prof. Oard started by walking through these evaluation design issues, and then he offered his initial thoughts on the consequences of these evaluation choices for our system designs. Additional information on the MATERIAL program is available at `https://www.iarpa.gov/index.php/research-programs/material`.

3 Paper Presentations

Test Collections and Measures for Evaluating Customer-Helpdesk Dialogues

Zeng et al. [8] addressed the problem of evaluating textual, task-oriented dialogues between the customer and the help-desk, such as those that take the form of online chats. As an initial step towards evaluating automatic help-desk dialogue systems, they have constructed a test collection comprising 3,700 real Customer-Helpdesk multi-turn dialogues by mining Weibo, a major Chinese social media. They have annotated each dialogue with multiple subjective quality annotations and nugget annotations, where a nugget is a minimal sequence of posts by the same utterer that helps towards problem solving. In addition, 10% of the dialogues have been manually translated into English. The test collection DCH-1 is made publicly available for research purposes. They also proposed a simple nugget-based evaluation measure for task-oriented dialogue evaluation, called UCH, and explored its usefulness and limitations.

An Interval-Like Scale Property for IR Evaluation Measures

Ferrante et al. [1] discussed the role played by evaluation measures in IR experimental evaluation and how their properties determine the kind of statistical analyses we can conduct. They have previously shown that it is questionable that IR effectiveness measures are on an interval-scale and that, as a consequence, computing means and variances is not a permissible operation. They further investigated whether it is possible to relax a bit the definition of interval scale, introducing the notion of interval-like scale, and to what extent IR effectiveness measures comply with this relaxed definition.

Evaluating Evaluation Measures with Worst-Case Confidence Interval Widths

Sakai [3] dealt with how IR evaluation measures are often compared in terms of rank correlation between two system rankings, agreement with the users' preferences, the swap method, and discriminative power. While he viewed the agreement with real users as the most important, he proposed to use the Worst-case Confidence interval Width (WCW) curves to supplement it in test-collection environments. WCW is the worst-case width of a confidence interval (CI) for the difference between any two systems, given a topic set size. He argued that WCW curves are more useful than the swap method and discriminative power, since they provide a statistically well-founded overview of the comparison of measures over various topic set sizes, and visualise what levels of differences across measures might be of practical importance. First, he proved that Sakais ANOVA-based topic set size design tool can be used for discussing WCW instead of his CI-based tool that cannot handle large topic set sizes. He then provided some case studies of evaluating evaluation measures using WCW curves based on the ANOVA-based tool, using data from TREC and NTCIR.

Automatic Evaluation of World History Essay Using Chronological and Geographical Measures

Sakamoto et al. [7] proposed a method for measuring chronological and geographical consistency of the world history essays in Japanese university entrance exams. The experimental results show a weak positive correlation between the scores measured by the proposed method and the scores estimated by a human expert in world history.

Towards Automatic Evaluation of Multi-Turn Dialogues: A Task Design that Leverages Inherently Subjective Annotations

Sakai [4] proposed a design of a shared task whose ultimate goal is automatic evaluation of multi-turn, dyadic, textual help-desk dialogues. Thee proposed task takes the form of an offline evaluation, where participating systems are given a dialogue as input, and output at least one of the following: (1) an estimated distribution of the annotators quality ratings for that dialogue; and (2) an estimated distribution of the annotators nugget type labels for each utterance block (i.e., a maximal sequence of consecutive posts by the same utterer) in that dialogue. This shared task should help researchers build automatic help-desk dialogue systems that respond appropriately to inquiries by considering the diverse views of customers. The proposed task has been accepted as part of the NTCIR-14 Short Text Conversation (STC-3) task. While estimated and gold distributions are traditionally compared by means of root mean squared error, Jensen-Shannon divergence and the like, he proposed a pilot measure that considers the order of the probability bins for the dialogue quality subtask, which we call Symmetric Normalised Order-aware Divergence (SNOD).

The Effect of Inter-Assessor Disagreement on IR System Evaluation: A Case Study with Lancers and Students

Sakai [5] reported on a case study on the inter-assessor disagreements in the English NTCIR-13 We Want Web (WWW) collection. For each of our 50 topics, pooled documents were independently judged by three assessors: two "lancers" and one Waseda University student. A lancer is a worker hired through a Japanese part time job matching website, where the hirer is required to rate the quality of the lancer's work upon task completion and therefore the lancer has a reputation to maintain. Nine lancers and five students were hired in total; the hourly pay was the same for all assessors. On the whole, the inter-assessor agreement between two lancers is statistically significantly higher than that between a lancer and a student. We then compared the system rankings and statistical significance test results according to different qrels versions created by changing which assessors to rely on: overall, the outcomes do differ according to the qrels versions, and those that rely on multiple assessors have a higher discriminative power than those that rely on a single assessor. Furthermore, he considered removing topics with relatively low inter-assessor agreements from the original topic set: we thus rank systems using 27 high-agreement topics, after removing 23 low-agreement topics. While the system ranking with the full topic set and that with the high-agreement set are statistically equivalent, the ranking with the high-agreement set and that with the low-agreement set are not. Moreover, the low-agreement set substantially underperforms the full and the high-agreement sets in terms of discriminative power. Hence, from a statistical point of view, his results suggest that a high-agreement topic set is more useful for finding concrete research conclusions than a low-agreement one.

Unanimity-Aware Gain for Highly Subjective Assessments

Sakai [6] dealt with the issue of subjectivity in assessment: in particular, human assessments of items such as social media posts can be highly subjective, in which case it becomes necessary to hire many assessors per item to reflect their diverse views. For example, the value of a tweet for a given purpose may be judged by (say) ten assessors, and their ratings could be summed up to define its gain value for computing a graded-relevance evaluation measure. He proposed a simple variant of this approach, which takes into account the fact that some items receive unanimous ratings while others are more controversial. He generatef simulated ratings based on a real social-media-based IR task data to examine the effect of his unanimity-aware approach on the system ranking and on statistical significance. The results show that incorporating unanimity can affect statistical significance test results even when its impact on the gain value is kept to a minimum. Moreover, since our simulated ratings do not consider the correlation present in the assessors actual ratings, our experiments probably underestimate the effect of introducing unanimity into evaluation. Hence, if researchers accept that unanimous votes should be valued more highly than controversial ones, then our proposed approach may be worth incorporating.

Acknowledgements

The workshop organizers are sincerely grateful to all the program committee members – Hsin-Hsi Chen (National Taiwan University, Taiwan), Charles Clarke (University of Waterloo, Canada), Atsushi Fujii (Tokyo Institute of Technology, Japan), Donna Harman (National Institute of Standards and Technology (NIST), USA), Gareth Jones (Dublin City University, Ireland), Noriko Kando (National Institute of Informatics, Japan), Evangelos Kanoulas (University of Amsterdam, The Netherlands), Aldo Lipani (Vienna University of Technology, Austria), Claudio Lucchese (Ca' Foscari University of Venice and ISTI CNR Pisa, Italy), Maria Maistro (University of Padua, Italy), Stefano Mizzaro (University of Udine, Italy), Alistair Moffat (University of Melbourne, Australia), Shahzad Rajput (National Institute of Standards and Technology (NIST), USA), Tetsuya Sakai (Waseda University, Japan), Mark Sanderson (RMIT University, Australia), Falk Scholer (RMIT University, Australia), Mark Smucker (University of Waterloo, Canada).

Finally, we want to thank all the authors, speakers, and other participants for making EVIA 2017 a success.

References

[1] M. Ferrante, N. Ferro, and S. Pontarollo. An Interval-Like Scale Property for IR Evaluation Measures. In Ferro and Soboroff [2], pages 10–15.

[2] N. Ferro and I. Soboroff, editors. *Proc. 8th International Workshop on Evaluating Information Access (EVIA 2017)*, 2017. CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, `http://ceur-ws.org/Vol-2008/`.

[3] T. Sakai. Evaluating Evaluation Measures with Worst-Case Confidence Interval Widths. In Ferro and Soboroff [2], pages 16–19.

[4] T. Sakai. Towards Automatic Evaluation of Multi-Turn Dialogues: A Task Design that Leverages Inherently Subjective Annotations. In Ferro and Soboroff [2], pages 24–30.

[5] T. Sakai. The Effect of Inter-Assessor Disagreement on IR System Evaluation: A Case Study with Lancers and Students. In Ferro and Soboroff [2], pages 31–38.

[6] T. Sakai. Unanimity-Aware Gain for Highly Subjective Assessments. In Ferro and Soboroff [2], pages 39–42.

[7] K. Sakamoto, H. Shibuki, M. Ishioroshi, A. Fujita, Y. Kano, T. Mitamura, T. Mori, and N. Kando. Automatic Evaluation of World History Essay Using Chronological and Geographical Measures. In Ferro and Soboroff [2], pages 20–23.

[8] Z. Zeng, C. Luo, S. Shang, H. Li, and T. Sakai. Test Collections and Measures for Evaluating Customer-Helpdesk Dialogues. In Ferro and Soboroff [2], pages 1–9.