EvALL: Open Access Evaluation for Information Access Systems

Enrique Amigo, Jorge Carrillo-de-Albornoz, Mario Almagro-Cádiz, Julio Gonzalo, Javier Rodríguez-Vidal and Felisa Verdejo
nlp.uned.es
Universidad Nacional de Educación a Distancia (UNED)
Calle Juan del Rosal 16, Madrid, Spain
{enrique,jcalbornoz,malmagro,julio,jrodriguez,felisa}@lsi.uned.es

ABSTRACT

The EvALL online evaluation service aims to provide a unified evaluation framework for Information Access systems that makes results completely comparable and publicly available for the whole research community. For researchers working on a given test collection, the framework allows to: (i) evaluate results in a way compliant with measurement theory and with state-of-the-art evaluation practices in the field; (ii) quantitatively and qualitatively compare their results with the state of the art; (iii) provide their results as reusable data to the scientific community; (iv) automatically generate evaluation figures and (low-level) interpretation of the results, both as a pdf report and as a latex source. For researchers running a challenge (a comparative evaluation campaign on shared data), the framework helps them to manage, store and evaluate submissions, and to preserve ground truth and system output data for future use by the research community.

EvALL can be tested at http://evall.uned.es.

KEYWORDS

Information Retrieval, Information Access, Evaluation, Evaluation infrastructure, Evaluation Metrics

1 INTRODUCTION

In spite of the strong focus of Information Retrieval (and Information Access in general) on comparative evaluation and replicability, researchers still face many challenges at the time of assessing the quality of their systems with respect to the state of the art. For instance: system descriptions are often not detailed enough and prevent replication of results, datasets are sometimes difficult to obtain or subject to privacy issues, etc. We focus here on a particular set of issues regarding evaluation:

- Finding all relevant state of the art results is costly. Once a suitable test collection is selected (and acquired), locating state of the art results on the test collection takes time and effort. If the test collection was created as part of a shared evaluation campaign, it is usually easy to find descriptions and results of the participating systems, but anything published after the campaign (where algorithms and results are usually optimized) takes time and effort to locate.
- Unavailability of system outputs prevents full comparison with state of the art systems. Usually, system performance is reported in terms of a few evaluation metrics, and therefore comparison with previous systems can only be established in a very limited way. When comparing with state of the art systems, no alternative metrics can be used, and no qualitative or per-test-case analysis can be performed. A detailed comparison between systems would only be possible if system outputs were available, and this is not generally the case.
- Proper choice, use and interpretation of adequate evaluation metrics is not straightforward. Because comparison with state of the art is essential, researchers tend to focus on popular evaluation methodologies and metrics, even if the state of the art on evaluation has moved on. As a result, metrics with preferable formal and empirical properties are often dismissed in favor of legacy metrics that warrant backwards comparability. In Document Retrieval, for instance, the adoption of state-of-the-art metrics is remarkably slow compared with the pace of innovation in the area. In fact, selecting appropriate evaluation metrics as a function of the scenario and task at hand, and understanding what they say about system performance, what they do not say, and how they complement each other, is still challenging for most general problems (such as retrieval, clustering and classification, which are the three main document organization task families). Researchers tend to focus on system development, and spend little time in selecting and understanding evaluation metrics. Again, resorting to the most popular metrics is a safe alternative, at the cost of a suboptimal (and sometimes misleading) interpretation of the experimental results.

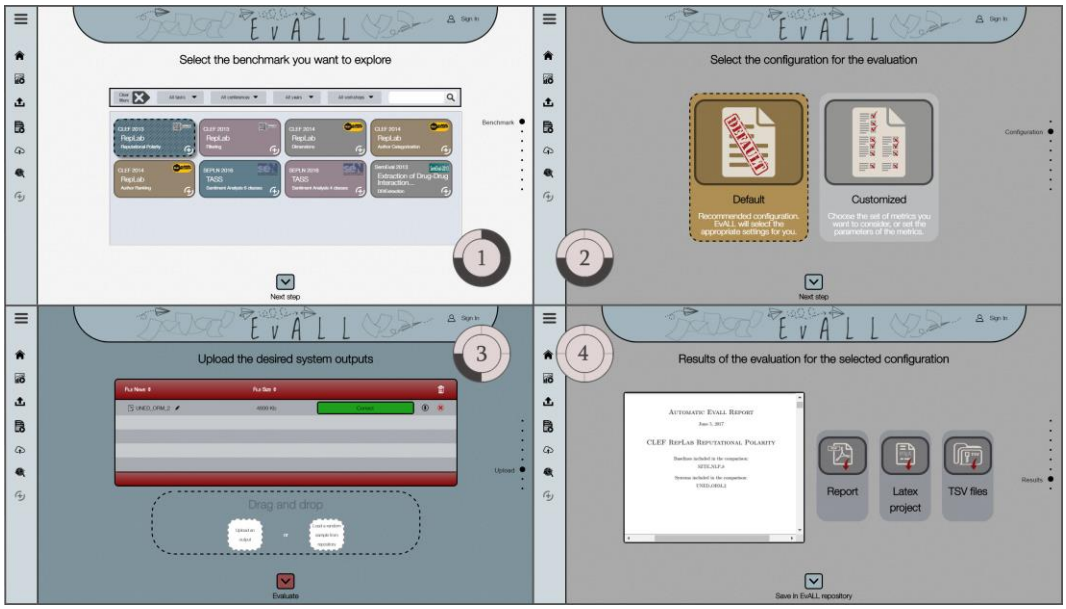

Figure 1: EvALL default workflow using a stored benchmark

- Sharing system outputs is not trivial. Once results are published as scholar articles, there is still no generalized standard procedure to release system outputs together with them, to be used as a reference for future research on the same test collection.

The EvALL system (http:/evall.uned.es) has been designed as an online service for the Information Retrieval and Natural Language Processing research communities that addresses all the above issues, for most of the relevant task families in both fields. In essence, EvALL stores system outputs and gold-standard references, and lets researchers evaluate exhaustively their systems with respect to stored state-of-the-art systems and references, and also upload and publish gold-standard annotations for new collections and new system outputs.

With respect to other online evaluation services in the areas of Information Retrieval and Natural Language Processing, EvALL has two distinctive features: (i) it aspires to be a universal evaluation service for the IR & NLP communities, while current evaluation platforms are focused on specific tasks. To make this feasible and cost-effective, it is focused on the storage of system outputs and gold-standard solutions, leaving aside test collections, systems, executions, workflows, etc.; and (ii) it makes a strong emphasis on assisting users in the choice and interpretation of evaluation metrics, based on a formal analysis and typification of metric properties [2, 3].

2 RELATED WORK

One of the pillars of scientific and technological process is an easily accessible state of the art. In the broad field of Collaborative Science, different initiatives promoting transparency have recently emerged in relation to the preservation of research resources, such as data repositories – such as FigShare (https://figshare.com) or Dryad (http://datadryad.org) – or code repositories such as GitHub (https://github.com) and Zenodo (https://zenodo.org/). Also, software journals, such as Data in Brief, have started to store datasets as publications. Nevertheless, none of these initiatives focus on improving communication between algorithmic methods, and thus replicability can be difficult in spite of total access to software and datasets. Scientific *WorkFlows* [9] and Science Gateways [1] have been proposed to address the excessive complexity of experiments over the last decade.

Focusing on evaluation in the areas of Information Retrieval and Natural Language Processing, evaluation campaigns have usually been the creators and promoters of assessment tools such as *trec eval* [5] for Document Retrieval tasks; and general-purpose Machine-Learning environments such as *Weka* [6] usually integrate evaluation modules for classification tasks. In general, such tools do not have a preservation layer, and do not allow to store and share system outputs and evaluation results with the research community.

On the other hand, there have also been notable efforts to provide online evaluation services, usually for specific tasks. In the context of Information Retrieval, [4] highlighted the lack of consistency reporting progress with respect to the state of the art in IR, and developed an online evaluation service, *EvaluatIR*, which provided a repository for IR system runs and evaluation results to "allow comparison between results submitted by different research groups at different times". More recently, [10] presented RISE (*Reproducible Information Retrieval System Evaluation*), a Web-based service (built on top of a modified version of the Indri toolkit) that implements more than 20 state of the art retrieval functions, and evaluates them over 16 standard TREC collections.

RISE is designed to facilitate the implementation and evaluation of retrieval functions, and the system hosts the data collections instead of shipping the data collections to researchers, which can ensure the privacy of the collections. In this sense, RISE implements the *Evaluation as a Service* (EaaS) philosophy [7], which aims at providing infrastructure support for evaluation such that datasets are stored centrally and accessed (usually via an API) for evaluation purposes by researchers, which circumvents legal restrictions on dataset distribution, facilitates scalability for system developers, and improves system comparability (at the cost of reducing system diversity). In the context of Natural Language Processing, a remarkable example is Gerbil [8], an evaluation framework for entity annotators that stores and archives systems, datasets, evaluation tools and experimental results.

While similar to EvALL with respect to the goals of preserving evaluation data and facilitating and standardizing the evaluation process, the difference with respect all the above services is in scope and depth:

(i) All previous online evaluation services are focused on a specific task or set of related tasks (around entity annotation, for instance, in the case of Gerbil), and they store everything that is needed to carry on meaningful evaluation for such tasks (including centralized datasets and/or code). EvALL, on the other hand, aims to maximize the coverage of Information Access tasks at the most abstract level (starting with classification, ranking and clustering problems) by minimizing what is preserved and stored: system outputs, gold standards, evaluation metrics and procedures, and experimental results.

(ii) EvALL also focuses on documentation and a careful choice and explanation of metrics. Often, users ignore how to correctly interpret metrics and what do they say differently about system behavior. In EvALL, explanatory reports are generated with all the required formal and mathematical background to figure out what evaluation metrics have been applied and how they have been used.

(iii) Finally, EvALL has been designed to have the lowest possible cost of entry; unlike other evaluation facilities, it can be used even with no prior knowledge of evaluation procedures. And registration is only needed if persistent storage of system outputs and/or gold standards is required. As a result of its simplicity, we expect to foster its take-up by the research community, which is one of the major challenges for this type of services. And we also expect EvALL to be particularly useful in master courses in the field as a gentle introduction to evaluation metrics and system performance analysis.

3 EVALL

3.1 System Description

For researchers working on a given benchmark / test collection, the EvALL online evaluation service (http://evall.uned.es) lets users:

(i) evaluate their results in a way compliant with measurement theory and with state-of-the-art evaluation practices in the field. Researchers only have to indicate either the reference collection (if it is already stored in EvALL) or the type of task (otherwise). Depending on the type of task, EvALL selects all suitable measures, checks their preconditions (in the system outputs and the gold standard), and generates the evaluation results. The EvaLL output includes a brief description of the selected measures that summarizes their motivation and properties, indicates their limitations, and provides relevant associated bibliography for further investigation. EvALL exploits the fact that most Information Access tasks belong to a few abstract problems (classification, ranking and clustering being the ones currently handled by the system). The organization and description of measures follows the axiomatic approach introduced in [2, 3].

(ii) quantitatively and qualitatively compare their results with the state-of-the-art. If the gold standard is already stored in EvALL, the system provides a repository of baselines and state-of-the-art systems, so that the new system can be compared in detail (rather than just in overall performance) with state-of-the-art approaches. Note that, if the gold standard is not stored in EvALL yet, users can still evaluate their systems by providing a reference gold standard and specifying the type of task.

(iii) provide their results as reusable data to the scientific community. No logging is necessary to use EvALL; but upon login, researchers can choose to share their system outputs and evaluation results with the community. Also, they can choose to upload a new test collection by providing a description of the task, a gold standard, and suitable system outputs (possibly including baseline approaches). Here, EvALL exploits the fact that system outputs and gold standards (reference annotations) are considerably lighter than datasets and code, and sharing them has significantly less Intellectual Property issues. For researchers running a challenge (a comparative evaluation campaign on shared data), the framework helps them to manage, store and evaluate submissions, and to preserve ground truth and system output data for future use by the research community.

(iv) Automatically generate evaluation figures and (low-level) interpretation of the results. EvALL produces as output a pdf report and its Latex source files, that allow researchers to easily copy and paste tables, descriptions, or results analysis into their publications. In addition, EvALL also generates a set of TSV reports containing all data in a more fine-grained way, which allows researchers to do further experimentation and analysis. EvALL is used via a web interface (http://evall.uned.es) which also includes other features such as statistical significance tests, metric smoothing where appropriate, personalization of evaluation reports, default versus manual measure selection and parameterization, standardization of the input format across tasks, and warnings and statistics about the system outputs.

In summary, with a single click the user obtains LaTeX-formatted information with results in terms of multiple measures, statistical significance tests, and system output data verification,

as well as information about the properties and limitations of the measures. And also, upon login, users can share their system outputs and share new test collections (as represented by a gold standard, baselines, assorted system outputs and specifications for the evaluation, such as, for instance, evaluation campaign guidelines).

Software-wise, EvALL is designed in two independent components: the web service (*EvALL web*), and the evaluation library (*EvALL Toolkit*). The first one is implemented on the *Liferay* framework in combination with a *MySQL* database. The web service manages user requests, stores and interacts with the repositories, and handles users management. It also interacts with the EvALL toolkit via its API to serve user requests. In the EvALL toolkit, all evaluation metrics have been re-implemented by the authors in Java. In order to minimize potential errors, every metric has been implemented twice, and its output has been cross-checked with third party implementations whenever possible (e.g. some ranking measures have been cross-checked with trec eval software).

3.2 Workflows

Once they enter the system, users are requested to select one out of four main actions:

- Evaluate using an existing benchmark. In this case, users provide one or more system outputs and choose the details of the evaluation procedure.
- Evaluate using their own benchmark. In this case, users are guided to define the type of task, and to upload a gold standard in addition to system outputs and baselines. • Publish a new system output. Upon registration, users can add a system output to a dataset that is already stored in EvALL.
- Publish a new benchmark. Upon registration, it is also possible to define and upload materials to include a new benchmark in EvALL and share it with the research community. Note that the test collection itself is not uploaded (which would complicate matters legally from the point of view of distribution); only the gold standard and basic evaluation specifications (type of task, official metrics, etc.) are stored in the system.

Let us seen, for instance, the system workflow when the user selects the first option (evaluate using an existing benchmark). The user is then requested to (a) choose a benchmark by browsing or searching the repository of tasks already included in EvALL, via a faceted search interface which allows to filter results by year, conference and/or keyword; and (b) select a configuration for the evaluation procedure, which can be by default or customized.

If the default option is chosen, users are asked to select or upload their system outputs, and EvALL directly produces the results of the evaluation process: (i) a pdf report that includes latex sources (ii) a tsv file with the evaluation output, and (iii) the result of a consistency check on system outputs, with warnings in case of inconsistent formats. In the default report, EvALL makes its own selection of appropriate metrics and reference systems (which include, for instance, the best stored system as the state-of-the-art reference), and provides a full report with all the theoretical explanations needed to interpret metric results. The default report is verbose in order to be self-contained: even with no previous knowledge of evaluation metrics, the reader has all the information needed to understand and interpret the results. Figure 1 displays screen captures for this default workflow.

If the customized option is chosen, the user is requested to make additional choices:

(1) The system lets the user select metrics by choosing one of three options: (a) official set of metrics (as prescribed in the corresponding evaluation campaign, if there is one); (b) full set of metrics (all appropriate metrics implemented in EvALL), or (c) a fully customized set of metrics. In the last option, the user makes a multiple-selection choice of metrics, and may optionally set metric parameters (such as the α parameter that sets the relative weight between metrics in the F-measure).
(2) The user is then asked to select systems to be included in the comparison. It might select options such as "best stored system" according to the selected metrics, or go to a completely manual selection among stored systems.
(3) Finally, the user is asked to customize the evaluation report, with the possibility of removing the metric descriptions, the latex sources, the tsv file, the output verification, etc.

3.3 EvALL coverage: tasks and measures

One of the main goals of EvALL is becoming a universal evaluation tool for any Information Access problem. We have inspected all tasks proposed in TREC, CLEF and SemEval in 2016, and our evaluation service could be currently used in 47 out of 63 tasks (74%). Thus, potentially, gold standards and system outputs for 74% of all these tasks could be stored in EvALL for evaluation purposes. And, even without storing them, researchers can evaluate their systems by entering their outputs and the gold standard for any of these tasks. Coverage would increase to 84% by incorporating text similarity metrics (such as ROUGE or WER) and value prediction estimators (such as MAE or Pearson correlation).

A key issue for the success of EvALL is how to overcome the cold-start problem. Even without a dataset of tasks, gold standards and system outputs, it still provides a fast and low-cost way of evaluating systems in a wide range of Information Access tasks. But, ultimately, its true success lies in its adoption by the research community, which will lead to a growing database of gold standard and system outputs. Our plan is to start working with shared task organizers to facilitate its uptake by the community.

Acknowledgements: This work has been partially funded by the Spanish Government (grants Vemodalen, TIN2015-71785-R, and Vox-Populi, TIN2013-47090-C3-1-P)

REFERENCES

[1] Robert N Allan. 2009. *Virtual research environments: From portals to science gateways*. Elsevier.

[2] Enrique Amigo, Julio Gonzalo, Javier Artiles, and Felisa Verdejo. 2009. A com-´ parison of extrinsic clustering evaluation metrics based on formal constraints. *Information retrieval* 12, 4 (2009), 461–486.

[3] Enrique Amigo, Julio Gonzalo, and Felisa Verdejo. 2013. A general evaluation´ measure for document organization tasks. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 643–652.

[4] Timothy G. Armstrong, Alistair Moffat, William Webber, and Justin Zobel. 2009. Improvements That Don'T Add Up: Ad-hoc Retrieval Results Since 1998. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM '09)*. ACM, New York, NY, USA, 601–610. DOI: http://dx.doi.org/10.1145/1645953.1646031

[5] C Buckley and others. 2004. The trec _eval evaluation package. (2004).

[6] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. 2009. The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter* 11, 1 (2009), 10–18.

[7] Jimmy Lin and Miles Efron. 2014. Infrastructure Support for Evaluation As a Service. In *Proceedings of the 23rd International Conference on World Wide Web (WWW '14 Companion)*. ACM, New York, NY, USA, 79–82. DOI:http://dx.doi.org/10.1145/2567948.2577014

[8] R. Usbeck, M. Roder, and A. Ngonga. 2015. GERBIL – General Entity Annotator Benchmarking Framework. In *Proc. 25th World Wide Web Conference*. ACM.

[9] WFMC. 1994. *Workflow reference model*. Technical Report. Workflow Management Coalition, Brussels.

[10] Peilin Yang and Hui Fang. 2016. A Reproducibility Study of Information Retrieval Models. In *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval (ICTIR '16)*. ACM, New York, NY, USA, 77–86. DOI: http://dx.doi.org/10.1145/2970398.2970415