Procesamiento del Lenguaje Natural, Revista nº 57, septiembre de 2016, págs. 189-192 recibido 28-03-2016 revisado 10-04-2016 aceptado 23-04-2016

Evall: A Framework for Information Systems Evaluation[*]

Evall: Una Plataforma para la Evaluación de Sistemas de Información

Enrique Amigó, Jorge Carrillo-de-Albornoz, Julio Gonzalo and Felisa Verdejo
Universidad Nacional de Educación a Distancia (UNED)
Calle Juan del Rosal 16, Madrid
{enrique, jcalbornoz, julio, felisa}@lsi.uned.es

Abstract: In this paper, the **Evall** framework for the automatic evaluation of information systems task is presented. With just one click and providing the system outputs of the algorithms, Evall allows researchers to automatically generate a Latex report including the results of their algorithms, statistical significance tests, measures descriptions, and references.
Keywords: Evaluation Framework, Information Systems

Resumen: En este artículo presentamos **Evall**, un framework de evaluación para tareas de investigación en el area de Sistemas de Información. Con un simple click y las salidas de los algoritmos a evaluar, Evall genera un informe automático en Latex con los resultados de todos los sistemas, test de significacia estadística, descripción de las métricas, y referencias.
Palabras clave: Framework de Evaluacion, Sistemas de Información

1 Introduction

In computer science, specially in the area of information systems, a common approach to evaluate the accuracy of proposed methods is by comparing the output generated by the algorithm with a **gold standard** built by human experts. For instance, **Classification** tasks consist of predicting the labels assigned to items by a certain gold standard. **Clustering** tasks aim to group items in the same way than the gold. Finally, the objective of **Ranking** tasks is to order a set of items in correspondence with the gold.

A correct evaluation pursuits, at least, two main objectives: **interpretable**, so it is easy to determine the relative improvements between systems; **standardization and replicability**, so it is possible to replicate the process and to obtain the same results, thus allowing comparison between different systems. Also, the selection of the appropriate evaluation measure determines to a great extent the conclusions of the experimental work. However the evaluation, and specially, the measure selection, is not a trivial issue. First of all, the same problem can be evaluated using several measures, and selecting the best ones for the problem at hand is a challenging work. Second, some measures are complex mathematical formulations that are not easy to understand and to interpret. Third, in many cases there are no standard implementations of the measures, or they are not available, so the measure is implemented several times, which can induce in bugs and errors (specially when comparing results from different implementations). Fourth, the input formats mostly depends of the measure implementation, and usually vary substantially for the same measure. For this reasons, only a small set of measures are commonly used.

The evaluation process plays another very important role in research, since it allows the community to compare their approaches and encourages to overcome them. Up to day, this is a difficult and challenging task, as to evaluate the state of the art of a benchmark (or dataset) used in a workshop or evaluation camping implies several hours searching on Internet for papers published using that benchmark. Also, in many cases the evaluations are performed in different scenarios and under different conditions, so they are not strictly comparable.

Finally, there are other relevant points that highly influence the evaluation process and that are usually complex and become it a tedious process. For example, statistical significance tests are not usually integrated in the measures implementations, or results are not appropriately formatted (an output

[*] This research was supported by the Spanish Ministry of Science and Innovation (VoxPopuli Project, TIN2013-47090-C3-1-P). Mario Almagro and Javier Rodriguez have contritubed to the development.

results, for instance, in Latex would strongly benefit the paper preparation).

2 *What is Evall?*

In this context, Evall aims to help researchers by proposing an easy-to-use **evaluation framework**, **transparent to the user**. Given a set of systems' outputs for a given task (i.e., Classification, Ranking, Clustering, etc.) and a gold standard, Evall produces an informative report in Latex format that includes measure descriptions and explanations, result tables, statistical significance tests, comparisons between systems' outputs, charts, references, etc., as well as a set of CSV files containing a more fine grained description of the results. The five main contributions of Evall are:

(i) Evall allows to evaluate systems outputs by **only indicating the task to be addressed** (i.e., Classification, Ranking, and Clustering). According to the task, Evall selects all available measures, checks their preconditions in the inputs (system outputs and gold standard), and generates the results. The selected measures are described, including a summary of its foundations and properties. Furthermore, indications about its limitations and relevant bibliography are also provided.

(ii) The **replicability of results** and the **comparison** between different systems' outputs is achieved. By using Evall, researches ensure that their evaluations are comparable with others that have used Evall too, and that the evaluations are free of errors.

(iii) Evall produces as output a **Latex report** that allows researchers to easily copy and paste tables, descriptions, or result analysis directly to future papers. Also, Evall produces a set of **CSV reports** containing all data in a more fine grained way, which allows researchers to do more experiments, further analysis, etc.

(iv) Evall is designed to **store benchmarks** (gold standards used in workshops, evaluation campaigns, conferences, or papers) and to evaluate new system outputs with the official measures (among others). This allow a researcher developing a new algorithm or approach to evaluate it just by producing the output in Evall format, avoiding the use of an evaluation library or the implementation of the desired measures. Evall also permits the comparison with all system outputs stored in the repository and addressing the same benchmark, and ensures a strict comparison under the same conditions. Similarly, this allows conference or workshops organizers to forget about evaluation by using Evall, ensuring an appropriate evaluation scenario.

(v) All **system outputs** evaluated against a benchmark using Evall can be **stored in the framework**, so future researchers can compare with them. Researchers can also associate some useful information such a brief description, a reference to a paper, etc.

Apart from these main features, Evall also includes other important features such as significance statistical tests, smoothing process, personalization of reports, manual selection of measures, standardization of the input format across tasks, or warnings and statistics about the system outputs. In summary, with a single click the user obtains edited information in Latex format with his/her results in terms of multiple measures, statistical significance tests, and system output data checking, as well as information about the categories, properties and limitation of the measures.

Up to day, and to the best of our knowledge, there are no standard evaluation frameworks or tools specially designed for this and covering a wide range of information systems tasks. There are some implementations of a small set of measures included in tools developed for another purpose, such as Weka[1], Gate[2] or Open NLP[3]. There are also some specific designed tools for evaluation for concrete problems such as machine translation (IQmt[4]), but there is not a universal and dedicated evaluation framework for information systems tasks.

3 *What can you do in Evall?*

In the design and development of Evall several scenarios have been taken into account:

Browsing scenario: The user is interested in exploring the existing tasks, measures, benchmarks or evaluation results stored in Evall by others researchers. The web interface allows the user to explore, learn and access to all relevant information stored

[1] http://www.cs.waikato.ac.nz/ml/weka/

[2] https://gate.ac.uk/

[3] https://opennlp.apache.org/

[4] http://www.cs.upc.edu/ nlp/IQMT/

in Evall, such as accessing the system outputs rankings associated to the benchmarks.

Evaluating against benchmarks included in Evall: The user is interested in comparing his/her results with the state of the art. Given a previously stored benchmark (i.e. Trec-2013), with only one click, the user can obtain a report comparing his/her own output with the baselines and best approaches stored in Evall for this benchmark. The report includes measures descriptions and suitability, evaluation results and salient aspects extracted from resulting data, as well as information about statistical significance tests. Furthermore, it compares the evaluation results against theoretical baseline approaches such as random systems. The user will need to (i) select an Evall benchmark and (ii) provide the system outputs in the Evall format.

Evaluating against his/her own benchmark: The user has defined his/her own benchmark. In this case the user has to (i) indicate the type of task (i.e. Classification), (ii) provide his/her gold standard in Evall format (iii), and provide the system outputs in Evall format.

Expert scenario: The user understands the nature and suitability of measures. He is interested in executing measures under a wide set of approaches and going beyond the capabilities of Evall standard reports by using the results obtained in the set of CSV files generated. Evall returns a set of CSV files containing the evaluation results for each system, test case and measure, as well as the aggregated results for each system. Evall gives also the possibility of customizing the evaluation results, by selecting measures and parameters.

System output contribution: The user can contribute with new system outputs against a benchmark included in Evall and store the results including some interesting information such as a brief description, a reference to a paper where the approach is better described, etc.

Benchmark contribution: A user, specially workshops and evaluation campaigns organizers, is interested in sharing his/her data under a common evaluation framework such as Evall and generating the results of the competition using Evall, allowing to compare the results achieved by the participants even when the evaluation campaign has been finished.

4 Evall inputs formats

Evall works under a general theoretical framework which categorizes problems in terms of measurement theory. Evall is based on the idea that system outputs or gold standards are measurements. That is, values assigned to items. This information can be captured with tuples containing the test case (i.e. queries in information retrieval), and an item and a value. For instance, the following table (Table 1) represents a relevance measurement for four documents in the context of two queries, (test cases), in a Ranking scenario.

Query_1	d31	0.5
Query_1	d52	0.2
Query_2	d31	0.7
Query_2	d25	0.3

Table 1: Example of Evall input format for ranking tasks

Given that the system output is a ranking (ordinal measurement), the last column can be avoided by considering directly the document order for the same query.

Query_1	d31	-
Query_1	d52	-
Query_2	d31	-
Query_2	d25	-

Table 2: Example of Evall input format for ranking tasks without absolute values

In the case of classification or clustering tasks, the measurement is nominal. Therefore, the relative order of values is not relevant, and the values can be strings as in the following example (Table 3):

Test_case_1	d21	Sport
Test_case_1	d23	World news
Test_case_2	d34	World news
Test_case_2	d43	World news

Table 3: Example of Evall input format for clustering and classification tasks

In the Evall framework, this is the common format for any output in any task: a three column CSV standard format. This format is used both for system outputs and gold standards.

5 Evall coverage: tasks and measures

In terms of tasks, Evall covers most of existing information access evaluation campaigns. For instance, examining the evaluation campaigns Semeval 2013 and 2014 and Clef 2014, we have checked that the Evall tasks cover 30 from 37 tasks or subtasks. The non covered problems include temporal intervals extraction, some text evaluation metrics (i.e. ROUGE), user based evaluation, and evaluation of structures. We expect to add some of them in future versions.

In terms of metrics, in particular the current Evall prototype covers the following sets:

Classification evaluation scenario consists in comparing the labels produced by systems for each item with the value provided by the gold. Both the gold and system outputs are nominal measurements. That is, the absolute difference or ordering relationships between values is not relevant for evaluation purposes. The classification measures provide with Evall includes: **Accuracy**, **Accurate Output**, **Weighted Accuracy**, **Utility** (Cormack and Lynam, 2005), **Lam%** (Hull, 1998), **Macro Average Accuracy**, **Kappa statistic** (Cohen, 1960), **Mutual Information**, **Precision**, **Recall**, **F-measure**, **Reliability** and **Sensitivity** (Amigó, Gonzalo, and Verdejo, 2013).

The **Ranking** evaluation scenario focuses on the priority relationships between items in both the gold and system outputs. That is, the priority relationships in the gold must be reflected in the system output. The measure set in Evall includes binary relevance measure such as **Precision at K**, **R-Precision**, **Mean Reciprocal Rank**, **Mean Average Precision**, and graded relevance measures such as **DCG** (Järvelin and Kekäläinen, 2002), **ERR** or **RBP** (Moffat and Zobel, 2008).

Clustering can be interpreted as the problem of predicting if two items belong or not to the same group. That is, predicting equality relationships between items. This corresponds with the fact of checking the relationship equivalence between two nominal measurements (the system output and the gold standard). The measures available in Evall are grouped into five categories: Set matching (**Purity** and **Inverse Purity**, **F-measure**), Entropy Based (**Class and Cluster entropy** (Steinbach, Karypis, and Kumar, 2000; Ghosh, 2003), **Mutual Information** (Xu, Liu, and Gong, 2003), **Counting Pairs** (Rand, Jaccar and F&M statistics (Halkidi, Batistakis, and Vazirgiannis, 2001; Meila, 2003)), **Editing Distance** and **Bcubed**.

References

Amigó, E., J. Gonzalo, and F. Verdejo. 2013. A general evaluation measure for document organization tasks. In *Proceedings of the 36th International ACM SIGIR Conference*, pages 643–652, New York, USA.

Cohen, J. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37.

Cormack, G. V. and T. R. Lynam. 2005. Trec 2005 spam track overview. In *TREC*.

Ghosh, J. 2003. Scalable clustering methods for data mining. In N. Ye, editor, *Handbook of Data Mining*. Lawrence Erlbaum.

Halkidi, M., Y. Batistakis, and M. Vazirgiannis. 2001. On Clustering Validation Techniques. *J. of Int. Inf. Syst.*, 17(2-3):107–145.

Hull, D. A. 1998. The TREC-7 filtering track: description and analysis. In *Proc. of TREC-7, 7th Text Retrieval Conf.*, pages 33–56.

Järvelin, K. and J. Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20:422–446, October.

Meila, M. 2003. Comparing clusterings. In *Proceedings of COLT 03*.

Moffat, A. and J. Zobel. 2008. Rank-biased precision for measurement of retrieval effectiveness. *ACM Trans. Inf. Syst.*, 27(1):2:1–2:27, December.

Steinbach, M., G. Karypis, and V. Kumar. 2000. A comparison of document clustering techniques.

Xu, W., X. Liu, and Y. Gong. 2003. Document clustering based on non-negative matrix factorization. In *Proc. of the 26th annual Int. ACM SIGIR Conf.*, pages 267–273.