Report on the Evaluation-as-a-Service (EaaS) Expert Workshop

Frank Hopfgartner, Allan Hanbury, Henning Müller, Noriko Kando, Simon Mercer, Jayashree Kalpathy-Cramer, Martin Potthast, Tim Gollub, Anastasia Krithara, Jimmy Lin, Krisztian Balog, Ivan Eggel
info@eaas.cc

07 April 2015

Abstract

In this report, we summarize the outcome of the "Evaluation-as-a-Service" workshop that was held on the 5th and 6th March 2015 in Sierre, Switzerland. The objective of the meeting was to bring together initiatives that use cloud infrastructures, virtual machines, APIs (Application Programming Interface) and related projects that provide evaluation of information retrieval or machine learning tools as a service.

1 Introduction

The standard approach to evaluating Information Retrieval (IR) systems involves distributing the data to the groups developing the systems so that they perform the evaluation locally. However, this approach of distributing data is often not practical, as the data may be:

- Huge – In order to obtain realistic evaluation results for IR, the evaluation should be done on realistic amounts of data. In the case of web search, this could be Petabytes of data. The current common approach of sending this data on hard disks through the postal service or via download has its limitations.
- Non-distributable – In many cases, it is not permitted to distribute data due to privacy, terms of service, or commercial sensitivity of the data. Privacy is the major concern for patient records. Even though law permits the distribution of anonymized medical records, large-scale anonymization can only be accomplished automatically, which data owners usually do not trust. For example, the Twitter Terms of Service forbid redistribution of tweets, while query logs are not made available for researchers after the AOL debacle in 2006. Distribution of company documents for the evaluation of enterprise search would not be permitted due to the commercial sensitivity of the data.
- Real-time – Companies working on real-time systems, such as recommender systems, are often not interested in evaluation results obtained on static historical data, in particular if this data has to be anonymised to allow distribution, as these results are too far removed from their operative requirements.

A number of initiatives are currently working to solve the above challenges. These initiatives all basically implement Evaluation-as-a-Service (EaaS), either making available APIs to access the data in a controlled way, or Virtual Machines (VMs) on which systems should be deployed. In order to organize these evaluation services, various aspects need to be considered. An overview of these aspects is given in Figure 1.

Figure 1: Overview of important aspects of evaluation-as-a-service (EaaS). Aspects are grouped into five dimensions: technology, people, policy, research, and business. At the bottom of the graphic, the nine EaaS grassroot initiatives that were presented at the workshop are listed.

In March 2015, Henning Müller and Allan Hanbury hosted a two-day workshop at the University of Applied Sciences Western Switzerland in Sierre in order to learn from the experiences of the organizers of these initiatives in tackling these aspects. The main aim of the workshop was to compile best practice guidelines, leading to the publication of a white paper. In this article, we first outline existing initiatives that were presented at the workshop (Section 2). Then, we summarize the main challenges for the implementation of Evaluation-as-a-Service that were discussed in Section 3. Follow-up plans are outlined in Section 4. The workshop was funded by the ELIAS[1] project of the European Science foundation and the FP7 project VISCERAL.[2]

[1] http://elias-network.eu/
[2] http://visceral.eu/

2 Existing Initiatives using EaaS Aspects

On the first workshop day, the participants presented individual evaluation initiatives and projects that implement aspects of shared computing and evaluation-as-a-service. In the remainder of this section, we briefly outline these initiatives. Fur further details, the interested reader is referred to the provided references.

2.1 TREC Microblog Task

The TREC Microblog tracks began in 2011 to explore search tasks and evaluation methodologies for information seeking behaviors in microblogging environments such as Twitter. This year (TREC 2015) marks the fifth iteration of the track. For the past four years, the core task has been temporally-anchored ad hoc retrieval, where the putative user model is as follows: "At time T, give me the most relevant tweets about an information need expressed as query Q." Since its inception, the track has had to contend with challenges related to data distribution, since Twitter's terms of service prohibit redistribution of tweets. For TREC 2011 and 2012 [12], the track organizers devised a solution whereby NIST distributed the *ids* of the tweets, rather than the tweets themselves. Given these ids and a downloader program (also developed by the track organizers), a participant could "recreate" the collection [11]. This approach adequately addressed the no-redistribution issue, but was not scalable. TREC 2013 [10] implemented an entirely different solution, which was to provide an API through which participants could complete the evaluation task. That is, the organizers gathered a collection of tweets centrally, but all access to the collection was mediated through the API, such that the participants could not directly interact with the raw collection. The search API itself was built using Thrift[3] and the Lucene search engine,[4] which are both widely-adopted open-source tools. A nice side-effect of the API approach is that common infrastructure promotes reproducibility [14] and sharing of open-source software components.

2.2 TIRA

The TIRA experimentation platform is a web service that supports organizers of shared tasks in computer science to accept the submission of executable software [5].[5] Traditionally, most shared tasks merely ask participants to submit the output of their software when run on a pre-published test dataset (a so-called "run"). This approach, however, has several shortcomings, including a complete lack of reproducibility of the shared task, and the necessity to publish test datasets prematurely, albeit sans ground truth. Notwithstanding these shortcomings, the organizers of shared tasks frequently employ run submission for its minimal organizational overhead, which used to be much smaller than that of software submission until now: TIRA automates software submission to a point at which it imposes no significant overhead, anymore, on organizers and participants alike. From the start, TIRA has been in active use: for the third year in a row, TIRA is employed for the four shared tasks of the PAN evaluation lab on digital text forensics [13], and as of this year, TIRA hosts the annual shared task of the CoNLL conference.

[3] `http://thrift.apache.org/`
[4] `http://lucene.apache.org/`
[5] `http://www.tira.io/`

2.3 BioASQ

The FP7 BIOASQ project aims to push research towards highly precise biomedical information access systems by establishing a series of challenges in which systems from teams around the world compete [2]. BIOASQ provides data, software and the evaluation infrastructure for the challenge. By these means, the project ensures that the biomedical experts of the future can rely on software tools to identify, process and present the fragments of the huge space of biomedical resources that address their personal questions. BIOASQ comprises two tasks. In Task A systems are required to automatically assign MESH terms to biomedical articles, thus assisting the indexing of biomedical literature. Systems participating in the task are given newly published MEDLINE articles, before the NLM curators have assigned MESH terms to them. The systems assign MESH terms to the documents, which are then compared against the terms assigned by the NLM curators. Task B focuses on obtaining precise and comprehensible answers to biomedical questions. The systems that participate in Task B are given English questions written by biomedical experts that reflect real-life information needs. For each question, the systems are required to return relevant articles, snippets of the articles, concepts from designated ontologies, RDF triples from Linked Life Data, an 'exact' answer (e.g., a disease or symptom), and a paragraph-sized summary answer [1].

2.4 Visceral

The FP7 project VISCERAL[6] is organizing a series of benchmarks on the processing of large-scale 3D radiology images [9]. The tasks include the segmentation of images, the detection of lesions in the images and the retrieval of similar cases including images and semantic terms as queries. VISCERAL is making use of an innovative cloud-based evaluation approach where all data are stored in the cloud. Participants in the tasks get virtual machines (VMs) to install their software and access to training data via the cloud. For the test phase the virtual machines are blocked for the participants and the organizers take over the VMs and run the executables in a defined format connecting the VM to a different storage with the test data. The idea is to bring the algorithms to the data instead of bringing the data to the algorithms [6]. The approach has several advantages as it first avoids sending hard disks with large amounts of data and allows working on confidential data as participants only get to see the training data set. In terms of science the availability of the data set and a working executable allows reproducibility of the approaches. The executables are also used in collaboration with the participants to run the algorithms on more non-annotated data sets with a goal to use label fusion and create more ground truth by fusing the output of all participant approaches. The ground truth created in this way is called the *silver corpus*, as opposed to the *gold corpus* that is created through manual annotation of the images by radiologists.

2.5 CLEF NewsREEL

The News REcommendation Evaluation Lab (NewsREEL)[7] is a campaign-style evaluation lab of CLEF. It implements the idea of *living laboratories* where researchers gain access to the resources of a company to evaluate different information access techniques using A/B

[6] `http://visceral.eu/`

[7] `http://clef-newsreel.org/`

testing [7]. The infrastructure is provided by plista GmbH, a company that provides a recommendation service for online publishers. Whenever a user requests an article from one of their customers' web portals, plista recommends similar articles that the user might be interested in. In NewsREEL, plista outsourced this recommendation task to interested researchers. Using plista's Open Recommendation Platform (ORP) [4], participants can register different recommendation algorithms and benchmark their performance over a longer period of time. One of the main requirements of this scenario is that recommendations have to be provided in almost real-time. Considering that a constant data stream [8] is exchanged between ORP and the participants' server, network latency becomes an actual issue since it reduces the amount of time remaining to compute recommendations. In order to avoid this time loss due, plista allows participants to run their algorithms on VMs in their data center.

2.6 CLEF LL4IR

Living Labs for Information Retrieval (LL4IR)[8] is an effort similar to NewsREEL, also running as a CLEF lab, but focusing on retrieval as opposed to recommendation. LL4IR provides a benchmarking platform where researchers can gain access to privileged commercial data (click and query logs) and can evaluate their ranking systems in a live setting, with real users, in their natural task environments. The first edition of the lab focuses on three specific use-cases: product search (on an e-commerce site), local domain search (on a university' website), and web search (through a major commercial web search engine). A key idea to removing the harsh requirement of providing rankings in real-time for query requests is to focus on head queries [3]. Participants can produce rankings for each query offline and upload these to the commercial provider. The commercial provider then interleaves a given participant's ranked list with their own ranking, and presents the user with the interleaved result list. Finally, feedback is made available to participants to facilitate improved offline ranking generation. Data exchange between live systems and participants is orchestrated by a web-based API.

2.7 CodaLab

The CodaLab platform[9] is an ongoing open-source development project with the goal of encouraging researchers to share and interact with datasets and algorithms through the medium of online scientific competitions. Written in Python, CodaLab both supports the standard academic model of competition in which participants download a common dataset, execute their algorithm locally and upload their results, but at the discretion of the competition owner it can also use the Microsoft Azure cloud to provide a standardized execution environment.

Any user can create a competition, defining multiple phases and automating the evaluation criteria needed to pass from one phase to the next. This may be done using either the editor provided or by uploading an appropriately-structured file - extensive documentation is available in the GitHub repository.[10] While the medical image analysis community were early adopters of CodaLab, the system offers sufficient flexibility to be useful to the scientific community in general and is now used more widely.

[8] `http://living-labs.net/`
[9] `http://www.codalab.org/`
[10] `https://github.com/codalab/codalab/wiki/`

2.8 C-BIBOP

Cloud-based Image Biomarker Optimization Platform (C-BIBOP)[11] is being developed as a technical resource for the cancer research community to support the development and assessment of quantitative imaging biomarkers. Lesion segmentation is a critical step in the development and use of imaging biomarkers in cancer. Another task that is organized as part of C-BIBOP requires the analysis of Magnetic Resonance Imaging (MRI) to identify biomarkers that best correlate with clinical outcomes. C-BIBOP is being developed to support reproducible science by enabling researchers to compare the performance of their image analysis algorithms that are co-located with large medical imaging datasets. The size of the datasets as well as the concerns about the sensitive nature of the data has highlighted the need for cloud-based solutions. Evaluation-as-a-service allows the challenge organizers to customize the evaluation methods for the clinical questions being addressed. Currently, C-BIBOP is built on the CodaLab plaform and plans to integrate key aspects from the VISCERAL project.

2.9 NTCIR

Since 1997, the NTCIR (short for NII Testbeds and Community for Information access Research) project has promoted research efforts for enhancing Information Access technologies such as Information Retrieval, Text Summarization, Information Extraction, and Question Answering techniques. Together with TREC and CLEF, it can be seen as one of the main venues for the organization of shared tasks. The general aim of NTCIR is to offer a research infrastructure that allows researchers to conduct large-scale evaluation of IA technologies; form a research community in which findings based on comparable experimental results are shared and exchanged, and develop evaluation methodologies and performance measures of information access technologies.

Differing from TREC and CLEF, NTCIR is following a two-cycles approach over a period of 18 months per cycle. Novel shared tasks are first organized as *pilot* tasks. More established tasks are organized as *core* tasks. Results are presented at the NTCIR conference in Tokyo, Japan. While NTCIR initially focused on Asian languages, it now accepts topics with a much broader focus. The current evaluation cycle (NTCIR-12) consists of five core tasks and three pilot tasks.

3 Challenges

After presenting the individual evaluation and benchmarking initiatives, the workshop participants identified and discussed a variety of aspects and challenges that need to be addressed in order to implement evaluation-as-a-service. These aspects and challenges, visualized in Figure 1, can be grouped into the five dimensions people, technology, policy, research, and business.

The main stakeholders in the organization of an EaaS activity include Task Organizers, Data Providers, Infrastructure Providers and Researchers. Naturally, these stakeholders are tightly integrated with the technological and political EaaS dimensions. Focusing on these stakeholders, we briefly summarize main issues that were discussed at the workshop. A more detailed discussion will be presented in a forthcoming white paper on evaluation-as-a-service.

[11] `http://cbibop.org/`

3.1 Organizers

The organizers ensure that data and tasks are provided on the EaaS infrastructure, and provide support for the participants. Challenges faced by the organizers include:

- The rules of participation in EaaS are still evolving, so designing EaaS activities requires more time. Questions to be considered include: How to allow participants to withdraw, can companies embargo their results, and the use of VMs containing participant submissions on data beyond that used in the evaluation (e.g., for silver corpus creation).
- Organizers feel that they have to assume additional responsibility due to potential security problems with the provided infrastructure.
- Organizers have to additionally provide technical support for the infrastructure, going beyond what is required in a traditional evaluation campaign.
- Covering the costs for the infrastructure can be a challenge, in particular covering the costs over a long period of time to run a series of evaluations and ensure their sustainability.
- On a commercial cloud infrastructure, some participants cause unnecessary costs by leaving VMs running while not computing.

3.2 Data Providers

The data providers provide data under certain conditions for use in a specific EaaS activity. Challenges faced by the data providers include:

- Fears that sensitive data will be leaked due to failure of the security procedures of the infrastructure.
- Drawing up a sufficiently strict and consistent data usage agreement that still is flexible enough to allow researchers to use the data as required.
- Enforceability of participant agreements — what can be done if a participant that is very likely to be in a different country breaches the participant agreement about data use?

3.3 Infrastructure Providers

Infrastructure providers make available the cloud or other infrastructure on which the EaaS activity runs. Challenges faced by the infrastructure providers include:

- Building a secure data access protocol to ensure that sensitive data cannot be downloaded from the system.
- Fear of illegal activities carried out by people granted access to a VM on the infrastructure.
- Certification of infrastructure to deal with specific types of sensitive data.
- Enforceability of participant agreements — what can be done if a participant that is very likely to be in a different country breaches the participant agreement about allowable infrastructure use.

3.4 Participants

The participants attempt to carry out the tasks defined by the organizers on the provided data making use of the EaaS infrastructure. Challenges faced by the participants include:

- High entry barriers, as the participants need to get used to a new infrastructure and potentially install software on a new VM, or get used to a new API;
- Fear of losing control of the evaluation as the participants cannot directly access all data involved;
- VMs provided on the EaaS infrastructure are potentially not powerful enough, or do not have specific hardware such as GPUs, so participants feel that they have less flexibility.
- Participants may feel uncertain in uploading code to an external infrastructure.

4 Next Steps

In order to pursue the idea of evaluation-as-a-service further, we have set up a web page[12] on which the latest developments in EaaS will be published. Moreover, a white paper detailing a thorough discussion on aspects and issues arising from this novel evaluation idea is in preparation. We aim for a more detailed analysis of the current initiatives, will identify relevant roles and stakeholders, and finally present a road-map for the development of evaluation-as-a-service in the short, medium and long term. This road-map aims to present routes to solving the challenges listed in the previous section, propose incentives for making EaaS interesting for industry, and culminate in a situation in which EaaS contributes to reproducibility in computational science and to encouraging innovation in industry.

Acknowledgements

We acknowledge financial support by the European Science Foundation via its Research Network Program "Evaluating Information Access Systems" (ELIAS) and by the European Commission via the FP7 project VISCERAL (318068).

References

[1] Georgios Balikas, Anastasia Krithara, Ioannis Partalas, and Georgios Paliouras. BioASQ: A challenge on large-scale biomedical semantic indexing and question-answering. In *MRMD'15: Proceedings of the Multimodal Retrieval in the Medical Domain Workshop*, 2015.

[2] Georgios Balikas, Ioannis Partalas, Axel-Cyrille Ngonga Ngomo, Anastasia Krithara, and Georgios Paliouras. Results of the BioASQ track of the question answering lab at CLEF 2014. In *CLEF'14: Proceedings of the 5th International Conference of the CLEF Initiative*, pages 1181–1193. Springer, 2014.

[3] Krisztian Balog, Liadh Kelly, and Anne Schuth. Head first: Living labs for ad-hoc search evaluation. In *CIKM'14: Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 1815–1818, 2014.

[12] `http://www.eaas.cc/`

[4] Torben Brodt and Frank Hopfgartner. Shedding Light on a Living Lab: The CLEF NEWSREEL Open Recommendation Platform. In *IIiX'14: Proceedings of Information Interaction in Context Conference*, pages 223–226. ACM, 08 2014.

[5] Tim Gollub, Benno Stein, and Steven Burrows. Ousting Ivory Tower Research: Towards a Web Framework for Providing Experiments as a Service. In *SIGIR'12: Proceedings of the 35th International ACM Conference on Research and Development in Information Retrieval*, pages 1125–1126. ACM, 2012.

[6] Allan Hanbury, Henning Müller, Georg Langs, Marc André Weber, Bjoern H. Menze, and Tomas Salas Fernandez. Bringing the algorithms to the data: cloud-based benchmarking for medical image analysis. In *CLEF'12: Proceedings of the 3rd International Conference of the CLEF Initiative*, pages 24–29. Springer Verlag, 2012.

[7] Frank Hopfgartner, Benjamin Kille, Andreas Lommatzsch, Torben Brodt, and Tobias Heintz. Benchmarking News Recommendations in a Living Lab. In *CLEF'14: Proceedings of the 5th International Conference of the CLEF Initiative*, pages 250–267. Springer Verlag, 09 2014.

[8] Benjamin Kille, Frank Hopfgartner, Torben Brodt, and Tobias Heintz. The plista dataset. In *NRS'13: Proceedings of the International Workshop and Challenge on News Recommender Systems*, pages 14–21. ACM, 10 2013.

[9] Georg Langs, Henning Müller, Bjoern H. Menze, and Allan Hanbury. Visceral: Towards large data in medical imaging — challenges and directions. In *MCBR–CDS'12: Proceedings of the Third MICCAI International Workshop*, pages 92–98. Springer, 2012.

[10] Jimmy Lin and Miles Efron. Overview of the TREC-2013 Microblog Track. In *TREC'13: Proceedings of the 22nd Text REtrieval Conference*, Gaithersburg, Maryland, 2013.

[11] Richard McCreadie, Ian Soboroff, Jimmy Lin, Craig Macdonald, Iadh Ounis, and Dean McCullough. On building a reusable twitter corpus. In *SIGIR'12: Proceedings of the 35th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1113–1114, Portland, Oregon, 2012.

[12] Iadh Ounis, Craig Macdonald, Jimmy Lin, and Ian Soboroff. Overview of the TREC-2011 Microblog Track. In *TREC'11: Proceedings of the 20th Text REtrieval Conference*, Gaithersburg, Maryland, 2011.

[13] Martin Potthast, Tim Gollub, Francisco Rangel, Paolo Rosso, Efstathios Stamatatos, and Benno Stein. Improving the Reproducibility of PAN's Shared Tasks: Plagiarism Detection, Author Identification, and Author Profiling. In *CLEF'14: Proceedings of the 5th Int. Conference of the CLEF Initiative*, pages 268–299. Springer Verlag, 2014.

[14] Jinfeng Rao, Jimmy Lin, and Miles Efron. Reproducible experiments on lexical and temporal feedback for tweet search. In *ECIR'15: Proceedings of the 37th European Conference on Information Retrieval*, pages 755–767, Vienna, Austria, 2015.