Pursuing a Moving Target: Iterative Use of Benchmarking of a Task to Understand the Task

Maria Eskevich[1], Gareth J. F. Jones[2], Robin Aly[3], Roeland Ordelman[3], Benoit Huet[4]

[1]Radboud University, The Netherlands; [2]ADAPT Centre, School of Computing, Dublin City University, Ireland
[3]University of Twente, The Netherlands; [4]EURECOM, Sophia Antipolis, France
m.eskevich@let.ru.nl; gjones@computing.dcu.ie; {r.aly, ordelman}@ewi.utwente.nl; Benoit.Huet@eurecom.fr

ABSTRACT

Individual tasks carried out within benchmarking initiatives, or campaigns, enable direct comparison of alternative approaches to tackling shared research challenges and ideally promote new research ideas and foster communities of researchers interested in common or related scientific topics. When a task has a clear predefined use case, it might straightforwardly adopt a well established framework and methodology. For example, an ad hoc information retrieval task adopting the standard Cranfield paradigm. On the other hand, in cases of new and emerging tasks which pose more complex challenges in terms of use scenarios or dataset design, the development of a new task is far from a straightforward process. This letter summarises our reflections on our experiences as task organisers of the *Search and Hyperlinking* task from its origins as a *Brave New Task* at the MediaEval benchmarking campaign (2011–2014) to its current instantiation as a task at the NIST TRECVid benchmark (since 2015). We highlight the challenges encountered in the development of the task over a number of annual iterations, the solutions found so far, and our process for maintaining a vision for the ongoing advancement of the task's ambition.

1. INTRODUCTION

Benchmark evaluation campaigns have become a key activity within a broad range of information processing disciplines, having demonstrated their critical impact on the fields' scientific progress especially for the information retrieval research community [11]. Individual benchmark tasks within these campaigns facilitate direct comparison of alternative approaches to specific technical challenges, encourage scientific innovation and, perhaps less obviously, enable understanding of what the task actually is. This last point is significant in the sense that the goal of a task can often be viewed as a "moving target" over successive (usually annual) iterations of the task. This situation arises over the period in which the task is active as the task organisers come to better understand what the task is seeking to achieve as a result of working to address questions raised by specification of the task itself, development of task datasets, the task participants feedback, and evaluation and analysis of the task results. In this letter we provide a brief review of our experiences of multiple iterations of the Search and Hyperlinking task developed within the MediaEval benchmark campaigns.

Copyright is held by the author/owner(s).
MediaEval 2016 Workshop, Oct. 20-21, 2016, Hilversum, Netherlands

2. SEARCH AND HYPERLINKING AT MEDIAEVAL

Our idea to define and shape an exploration of Search and Hyperlinking (S&H) through a benchmarking activity initially emerged from a diverse combination of reasons. A number of varied and challenging large scale multimedia data archives relevant to such a task were already becoming available, while the constantly increasing and diverse deluge of new multimedia content being produced, stored and shared by both non-, semi- and professionals meant that there was a compelling motivation to explore methods to search and manage this content. At the same time, scientific advances had reached the stage where algorithms with the potential to address more creative tasks that could encompass known-item and ad hoc retrieval of specific parts of content, as well as personalised collection exploration, were becoming available. Embarking on this adventure was also appealing since various aspects of the overall S&H task had already been investigated or tested in smaller scale tasks, e.g. the MediaEval 2011 Rich Speech Retrieval (RSR) Task [6] and the VideoCLEF 2009 Linking Task [7].

3. FROM A BRAVE NEW TASK TO A BRAVE NEW WORLD

From a starting point of a use case for a new task, the development of an actual benchmark activity often appears straightforward. However, this is often not the case, and once the task organisers begin to operationalize their ideas technical and practical challenges begin to emerge. This means that the task released to the participants is generally a technical and practical compromise, often containing hidden questions that the task organisers are unable to answer based on their current understanding of the user behaviour model or technical issues of the task. Thus the current instance of a task can itself be designed to answer these questions in order to move the task forward towards its ultimate research goals by exploiting better use case definition and representation in a subsequent version of the task.

The S&H tasks were a classic example of this situation. Once we began to examine the scope of what the task required in terms of specification and implementation, we realised that there were many questions to be addressed in order to fully understand the task itself and how it should best be implemented to benchmark the usability of its outputs and the algorithmic contributions of the participants' solutions. The activity thus began as a relatively small scale Brave New Task at MediaEval 2011 [6]. The key issue ad-

dressed in the first iteration was the exploration of the potential of crowdsourcing technologies for the query creation stage for a given collection and for the ground truth definition [4]. Setting up a task, that we envisaged as inspired by users' potential interests and request creation, we wanted to engage the real users in both task definition and evaluation.

In subsequent years the task received the status of a Main Task, meaning that we were able to gather a group of core participants (at least five) who expressed their interest in participating each year. Being a Main Task did not mean that the task definition and evaluation were set in stone, and thus, we kept experimenting with the collection, the type of users and their requests, evaluation metrics each year.

In 2014, we felt that the innovative Video Hyperlinking subtask within the S&H task had reached a good level of maturity in terms of task infrastructure, i.e., task definition [8], data availability and evaluation procedure [1], but there were still many questions unanswered in terms of addressing the algorithmic challenges of the task. We therefore sought the opportunity to increase participation and the range of scientific input by offering the task at TRECVid 2015, subsequently accepted by the TRECVid chairs [9].

Although we took the task to another venue, where most of the evaluation is usually done by NIST experts, we adhered to the crowdsourcing anchor creation and evaluation procedures that were established within our MediaEval activities. This approach preserved our flexibility in terms of the creativity of the task definition, and we kept our commitment to have users involved at all stages of benchmarking.

4. ITERATIVE TASK EVOLUTION

Traditionally, well established tasks with a straightforward scenario follow a pattern of gradually growing their dataset with each year iteration, using the same evaluation metric or a set of metrics, sometimes running the same software on the revised dataset, in order to be able to carry out direct comparison between the technology performance over the years. In the case of a more exploratory and innovative task, that is being developed through collaboration and feedback with participants, the same broad user scenario can be tested under different conditions, e.g. diverse target users of the potentially developed approaches, different data sets and evolving evaluation metrics that cover aspects of the task that could not have been foreseen beforehand.

When the task is defined by a clear use case scenario existing within an industrial set up, the task can be promoted by these industrial partners via data provision and help with the on-site evaluation. As our research focus is on large video archives that are not always created and gathered with a clear monetization strategy in mind, often aiming at cultural heritage preservation (without predefined usage scenarios), we were more free in defining the framework. The feedback from the crowdworkers helped us to test algorithms addressing the task in a fast iterative way.

Another aspect that has to be taken into account when setting up a task with a large data collection in mind is the copyright question. When the task is in its initial early development stage, it is easier to use a Creative Commons dataset to initially test the task feasibility. This proof of concept of the task viability allows the organisers to demonstrate the soundness of the overall framework, and thus to engage potential industry partners. This was the case for the S&H task that started with the BlipTV collection [10], and then switched to a BBC dataset [3, 2]. However, the usage of professionally created and copyright material also makes the task more dependent on the external partners and liable to all potential changes of the legal status of the data. Overall, the opportunity to run the task with different datasets enriches the discussion of the scientific approaches.

On the other hand, crowdsourcing of the task definition and results evaluation keeps the focus of the task on the user, and allows us to relate the scientific methods under test to the current users technology expectations. This brings a practical insight into the impact of the performance improvements in algorithms on user experience. In a way, the workers become part of the organisers team, i.e., the task, although being envisaged by the scientists, is finally shaped and vetted by the real users.

5. THE VIEW FROM A NEW HOME

Having run the task already for two years at the TRECVid benchmark, we can compare our experiences and outline the differences. At both venues at the initial stage of the yearly cycle, the tasks get feedback from the overall benchmark organisers committee in terms of task feasibility and interest within the targeted scientific community. However, during the yearly cycle of actually running the task, within the MediaEval campaign the organisers of all the tasks are aware of task progress, raising issues and sharing their solutions via bi-weekly conference calls. This is especially helpful, when tasks are sharing the datasets, or when they are being run for the first time and the organisers lack experience.

As organisers of a creative novel task, we found that interaction within the community of task organisers and with the actual task participants proved to be very useful to enable us to react quickly to any issues arising with the task, from the data release to submissions and evaluation release. However, TRECVid allows organisers to delegate some organisational activities to the NIST, thus saving time.

Running the benchmarking task requires a lot of commitment from the organisers, and an interest and engagement of the scientific community. In our experience, the growing cycle in terms of interest and participation in the S&H task coincided with a number of related projects funded at a time which also meant that the ending of the funding cycle affected the number of participants, while the actual scientific findings and discussions were still on an upwards path. The move to the TRECVid allowed us to involve large labs and companies that often participate in this venue.

6. SUMMARY FUTURE OUTLOOK

We have presented the evolution of the S&H task to date. Despite operating at two benchmarking venues, future challenges remain, with the most critical issue being sustainability [5]. While research is often bound to projects of finite length, the organization of tasks should ideally be able to continue independent of these. This is challenging in particular in terms of human resources and technical resources.

7. ACKNOWLEDGMENTS

This work has been partially supported by: ESF Research Networking Programme ELIAS; BpiFrance within the NexGenTV project, grant no. F1504054U; Science Foundation Ireland (SFI) as a part of the ADAPT Centre at DCU (13/RC/2106); EC FP7 project FP7-ICT 269980 (AXES).

8. REFERENCES

[1] R. Aly, R. Ordelman, M. Eskevich, G. J. F. Jones, and S. Chen. Linking inside a video collection: what and how to measure? In *Proceedings of the 22nd International World Wide Web Conference (WWW '13), Companion Volume*, pages 457–460, 2013.

[2] S. Chen, M. Eskevich, G. J. F. Jones, and N. E. O'Connor. An Investigation into Feature Effectiveness for Multimedia Hyperlinking. In *MultiMedia Modeling - 20th Anniversary International Conference (MMM 2014), Proceedings, Part II*, pages 251–262, Dublin, Ireland, 2014.

[3] M. Eskevich, R. Aly, D. N. Racca, R. Ordelman, S. Chen, and G. J. F. Jones. The Search and Hyperlinking Task at MediaEval 2014. In *Working Notes Proceedings of the MediaEval 2014 Workshop*, Barcelona, Catalunya, Spain, 2014.

[4] M. Eskevich, G. J. F. Jones, M. Larson, and R. Ordelman. Creating a Data Collection for Evaluating Rich Speech Retrieval. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, pages 1736–1743, Istanbul, Turkey, 2012.

[5] F. Hopfgartner, A. Hanbury, H. Müller, N. Kando, S. Mercer, J. Kalpathy-Cramer, M. Potthast, T. Gollub, A. Krithara, J. Lin, K. Balog, and I. Eggel. Report on the Evaluation-as-a-Service (EaaS) Expert Workshop. *SIGIR Forum*, 49(1):57–65, June 2015.

[6] M. Larson, M. Eskevich, R. Ordelman, C. Kofler, S. Schmiedeke, and G. J. F. Jones. Overview of MediaEval 2011 Rich Speech Retrieval Task and Genre Tagging Task. In *Working Notes Proceedings of the MediaEval 2011 Workshop*, Santa Croce in Fossabanda, Pisa, Italy, 2011.

[7] M. Larson, E. Newman, and G. J. F. Jones. Overview of VideoClef 2009: New Perspectives on Speech-based Multimedia Content Enrichment. In *Proceedings of the 10th International Conference on Cross-language Evaluation Forum: Multimedia Experiments*, CLEF'09, pages 354–368, Berlin, Heidelberg, 2010. Springer-Verlag.

[8] R. J. F. Ordelman, M. Eskevich, R. Aly, B. Huet, and G. J. F. Jones. Defining and Evaluating Video Hyperlinking for Navigating Multimedia Archives. In A. Gangemi, S. Leonardi, and A. Panconesi, editors, *Proceedings of the 24th International Conference on World Wide Web Companion (WWW 2015), Companion Volume*, pages 727–732, Florence, Italy, 2015. ACM.

[9] P. Over, J. Fiscus, D. Joy, M. Michel, G. Awad, W. Kraaij, A. F. Smeaton, G. Quénot, and R. Ordelman. TRECVID 2015 – an overview of the goals, tasks, data, evaluation mechanisms and metrics. In *Proceedings of TRECVID 2015*. NIST, USA, 2015.

[10] S. Schmiedeke, P. Xu, I. Ferrané, M. Eskevich, C. Kofler, M. A. Larson, Y. Estève, L. Lamel, G. J. F. Jones, and T. Sikora. Blip10000: a social video dataset containing SPUG content for tagging and retrieval. In *Multimedia Systems Conference 2013 (MMSys '13)*, pages 96–101, Oslo, Norway, 2013.

[11] C. V. Thornley, A. C. Johnson, A. F. Smeaton, and H. Lee. The scholarly impact of trecvid (2203 – 2009). *Journal of the American Society for Information Science and Technology*, 62(4):613–627, 2011.