Benchmarking Multimedia Technologies with the CAMOMILE Platform: the Case of Multimodal Person Discovery at MediaEval 2015

Johann Poignant[1], **Hervé Bredin**[1], **Claude Barras**[1],
Mickael Stefas[2], **Pierrick Bruneau**[2], **Thomas Tamisier**[2]
1. LIMSI, CNRS, Univ. Paris-Sud, Université Paris-Saclay, F-91405 Orsay, firstname.lastname@limsi.fr
2. LIST, Esch-sur-Alzette, Luxembourg, firstname.lastname@list.lu

Abstract

In this paper, we claim that the CAMOMILE collaborative annotation platform (developed in the framework of the eponymous CHIST-ERA project) eases the organization of multimedia technology benchmarks, automating most of the campaign technical workflow and enabling collaborative (hence faster and cheaper) annotation of the evaluation data. This is demonstrated through the successful organization of a new multimedia task at MediaEval 2015, Multimodal Person Discovery in Broadcast TV.

Keywords: evaluation campaign, collaborative annotation, multimedia

1. Introduction

For decades, NIST evaluation campaigns have been driving research in the field of human language technology (Martin et al., 2004), recently followed by the CLEF (Peters and Braschler, 2002) and ESTER/ETAPE (Gravier et al., 2004) initiatives. The concept has been successfully transposed to other research areas, such as image recognition (ImageNet Large Scale Visual Recognition Challenge (Russakovsky et al., 2015)), video (TRECVID (Smeaton et al., 2006)) or multimedia indexing (MediaEval (Larson et al., 2015)). More generally, evaluation campaigns allow the assessment of experimental research in fields where human perception and decision must be reproduced by machine learning algorithms (Geoffrois, 2008).

The general workflow of *à la NIST* evaluation campaigns comprises the following stages (Martin et al., 2004): specification of the task; definition of the evaluation metric and provision of an automatic scoring software; design and annotation of the training, development and evaluation corpora; definition of evaluation rules, schedule, protocols and submission formats; sharing of participant results through system descriptions and workshop communications.

Automatic scoring is made possible by the manual annotation of the data according to the task definition. Costly and time-consuming, this annotation step usually is the main bottleneck of evaluation campaigns. When addressing new tasks in multimodal perception, it becomes challenging (if not impossible) to pre-annotate the ever-increasing volume of multimedia data. A compromise has been successfully explored in the TREC and TRECVid campaigns, where the annotation of a small (but carefully chosen (Yilmaz and Aslam, 2006)) subset of the test data is bootstrapped by the participants' submissions.

In this paper, we claim that the CAMOMILE collaborative annotation platform (developed in the framework of the eponymous CHIST-ERA project) eases the organization of multimedia technology benchmarks, automating most of the campaign technical workflow and enabling collaborative (hence faster and cheaper) annotation of the evaluation data. This is demonstrated through the successful organization of a new multimedia task at MediaEval 2015, Multimodal Person Discovery in Broadcast TV (Poignant et al., 2015b).

2. Multimodal Person Discovery in Broadcast TV

The objective of this new task is to make TV archives fully exploitable and searchable through people indexing. Participants were provided with a collection of TV broadcast recordings pre-segmented into shots. Each shot had to be automatically tagged with the names of people both speaking and appearing at the same time during the shot.

Since one cannot assume that biometric models of persons of interest are available at indexing time, the main novelty of the task was that the list of persons was not provided a priori. Biometric models (either voice or face) could not be trained on external data. The only way to identify a person was by finding their name in the audio (using speech transcription - ASR) or visual (using optical character recognition - OCR) streams and associating them to the correct person – making the task completely unsupervised with respect to prior biometric models.

To ensure that participants followed this strict "no biometric supervision" constraint, each hypothesized name had to be backed up by an "evidence": a unique and carefully selected shot proving that the person actually holds this name (e.g. a shot showing a text overlay introducing the person by their name). In real-world conditions, this evidence would help a human annotator double-check the automatically-generated index, even for people they did not know beforehand.

Participants were provided with a fully functional baseline system, allowing them to only focus on some aspects of the task (e.g. speaker diarization) while still being able to rely on the baseline modules for the other ones (e.g. optical character recognition). The task was evaluated as a standard information retrieval task using a metric derived from mean average precision. Nine teams (Nishi et al., 2015; Budnik et al., 2015; Lopez-Otero et al., 2015; India et al., 2015; Poignant et al., 2015a; Bendris et al., 2015; dos Santos Jr et al., 2015; Le et al., 2015) managed to reach the

submission deadline, amounting to a total of 70 submitted runs. For further details about the task, dataset and metrics, the interested reader can refer to (Poignant et al., 2015b).

3. Person Discovery made easy with CAMOMILE

The CAMOMILE platform was initially developed for supporting collaborative annotation of multimodal, multilingual and multimedia data (Poignant et al., 2016). The data model was kept intentionally simple and generic, with four types of resources: corpus, medium, layer and annotation.

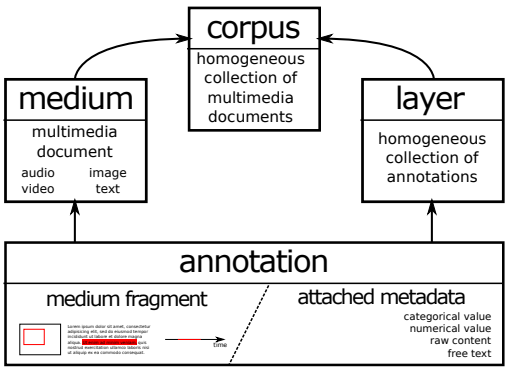


Figure 1: CAMOMILE data model

A corpus is a set of media (e.g. the evaluation corpus made of all test videos). An annotation is defined by a fragment of a medium (e.g. a shot) with attached metadata (e.g. the name of the current speaker). Finally, a layer is an homogeneous set of annotations, sharing the same fragment type and the same metadata type (e.g. a complete run submitted by one participant). All these resources are accessible through a RESTful API (clients in Python and Javascript are readily available), with user authentication and permission management.

A generic queueing mechanism is also available on the CAMOMILE backend as a means to control the workflow. The CAMOMILE platform is distributed as open-source software at the following address: `http://github.com/camomile-project/camomile-server`.

3.1. Automating the benchmarking workflow

The upper part of Figure 2 depicts the technical workflow of the proposed evaluation campaign.

The lower parts of Figure 2 summarize how we relied on the CAMOMILE platform and its Python and Javascript clients to automate most of the workflow.

3.1.1. Registration

After the task was advertised through the MediaEval call for participation, we relied on MediaEval standard registration procedure (i.e. filling an online form and signing dataset usage agreements) to gather the list of participating teams. Through a web interface, users and groups management features of the CAMOMILE platform were used to create one group per team and one user account for each team member.

3.1.2. Distribution

Due to technical (limited internet bandwith) or copyright concerns (datasets distributed by third parties), the development and evaluation datasets were not distributed through the CAMOMILE platform. Instead, ELDA and INA took care of sending the datasets to the participants. Nevertheless, corresponding metadata for corpora (development and test sets) and layers (for each video) were created as CAMOMILE resources with read permissions for each team, then bound to a local copy of the videos.

3.1.3. Submission

While the standard MediaEval submission procedure is to ask participating teams to upload their runs into a shared online directory, we chose to distribute to all participants a submission management tool, based on the CAMOMILE Python client. This command line tool would automatically check the format of the submission files, authenticate users with their CAMOMILE credentials and creates a new layer (and associated annotations) for each submission, with read/write permissions to (and only to) every team member.

3.1.4. Evaluation

For the duration of the submission period, a continuous evaluation service based on the CAMOMILE Python client would update a live leaderboard computed on a secret subset of the evaluation dataset – providing feedback to participants about the performance of their current submissions.

These four modules could easily be adapted to other benchmarking campaigns, as long as the reference and submissions can follow the CAMOMILE data model.

3.2. Collaborative annotation

While the development dataset had already been annotated in the framework of the past REPERE evaluation campaigns, the evaluation dataset was distributed by INA without any annotation. Thanks to the CAMOMILE platform, we were able to setup a collaborative annotation campaign where participants themselves would contribute some time to annotate the evaluation dataset.

3.2.1. Annotation interfaces

Two dedicated and complementary annotation web interfaces were developed, both based on the CAMOMILE Javascript client. The first one is dedicated to the correction of the "pieces of evidence" submitted by participants. For each correct evidence, annotators had to draw a bounding box around the face of the person and spellcheck their hypothesized name (`firstname_lastname`). The second one relies on the resulting mugshots to ask the annotator to decide visually if the hypothesized person is actually speaking and visible during a video shot. Moreover, a monitoring interface was also accessible to the organizers to quickly gain insight into the status of the annotation campaign (e.g. number of shots already annotated).

3.2.2. Backend

As shown in Figure 4, both annotation interfaces relied on the CAMOMILE queueing feature, thanks to a submission

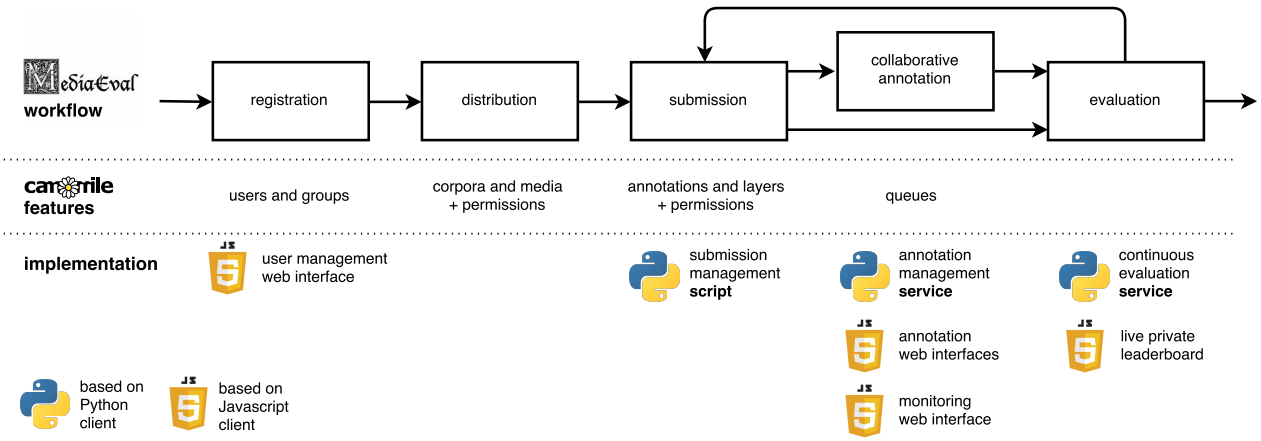
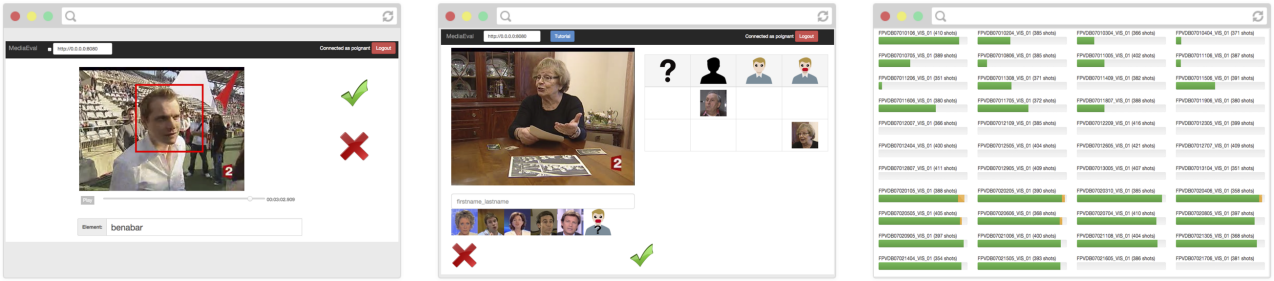


Figure 2: Workflow automation with the CAMOMILE platform



Evidence annotation | Label annotation | Annotation monitoring

Figure 3: Annotation web interfaces

monitoring service that would continuously watch for new submissions and update annotation queues accordingly. Every time a new run was submitted, the annotation management service would push not-yet annotated evidences into the CAMOMILE queue used as input for the evidence annotation interface. Corresponding mugshots (*i.e.* small picture depicting the person's face) would then be extracted automatically for later use in the label annotation interface. Similarly, not-yet annotated shots would be added into the CAMOMILE queue used as input of the label annotation interface. Once a consensus is reached (*cf.* next section), thoses shots would be added to the CAMOMILE groundtruth layer. Finally, a submission scoring daemon would continuously evaluate each submission, providing scores displayed by the live leaderboard.

3.2.3. Reaching consensus...

Table 1 summarizes the amount of work done during the annotation campaign. 7k+ "evidence" annotations were performed by 3 organizers while 66k+ "label" annotations were gathered from 20 team members – leading to the annotation of half of the evaluation corpus in less than a month. While the annotation of "evidence" was done by the organizers themselves, we wanted to guarantee the quality of the "labels" annotation done by the participants themselves. To that end, each shot was required to be annotated at least twice. Additional annotation of the same shot were requested until a consensus was found. Tables 2 and 3 show that, thanks to a simple, focused and dedicated "label" interface, the average number of required annotations for each shot is close to the minimum (2).

	Evidence	Label
# annotators	3	20
# annotations	7337	66089
Median duration	10.2s	4.4s

Table 1: Amount and median duration of annotations for both interfaces

	# shots
with 2+ annotations	28231 (100.0%)
with consensus	27873 (98.7%)
without consensus	358 (1.3%)

Table 2: Proportion of shots with/without consensus

# annotations	# shots
2	22770 (81.7%)
3	4257 (15.3%)
4	658 (2.4%)
5+	188 (0.6%)

Table 3: Number of annotations per shot with consensus

A quick look at the few shots with 4 or more annotations reveals a few ambiguous cases that were not forecast when designing the "label" annotation interface: people singing or dubbed, barely audible speech, etc.

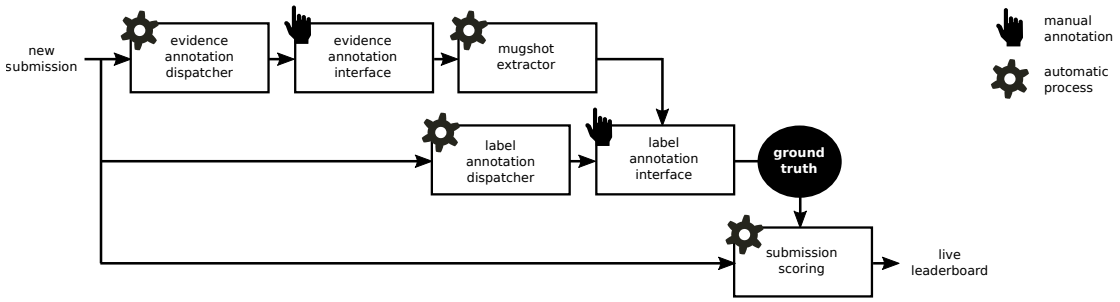


Figure 4: Annotation management service

4. Conclusion

Relying entirely on the CAMOMILE annotation platform, a team of two people was able to manage a large scale multimedia technology benchmark (nine teams, 70 submissions, 30k shots) – including the development of the submission management script, the leaderboard service and the whole annotation campaign. Everything was hosted on a virtual private server with 2 cores and 2 GB of RAM and resisted the load even during the peak submission time (right before the deadline) and the concurrent collaborative annotation period.

All the scripts and interfaces related to this campaign are publicly available on the CAMOMILE GitHub page. Though some were designed specifically for the proposed MediaEval Person Discovery task, we believe that a significant part of the approach is generic enough to be easily ported to a different task where manual and automatic annotation of audio-visual corpora is involved.

5. Acknowledgements

This work was supported by France "Agence Nationale de la Recherche" (ANR) under grant ANR-12-CHRI-0006-01 and Luxembourg "Fonds National de la Recherche" (FNR). We thank ELDA and INA for supporting the task with development and evaluation datasets.

6. Bibliographical References

Bendris, M., Charlet, D., Senay, G., Kim, M., Favre, B., Rouvier, M., Bechet, F., and Damnati, G. (2015). Percolatte : A multimodal person discovery system in tv broadcast for the medieval 2015 evaluation campaign. In *MediaEval*.

Budnik, M., Safadi, B., Besacier, L., Quénot, G., Khodabakhsh, A., and Demiroglu, C. (2015). Lig at mediaeval 2015 multimodal person discovery in broadcast tv task. In *MediaEval*.

dos Santos Jr, C. E., Gravier, G., and Schwartz, W. (2015). Ssig and irisa at multimodal person discovery. In *MediaEval*.

Geoffrois, E. (2008). An economic view on human language technology evaluation. In *LREC*.

Gravier, G., Bonastre, J., Galliano, S., Geoffrois, E., Mc Tait, K., and Choukri, K. (2004). The ester evaluation campaign of rich transcription of french broadcast news. In *LREC*.

India, M., Varas, D., Vilaplana, V., Morros, J., and Hernando, J. (2015). Upc system for the 2015 mediaeval multimodal person discovery in broadcast tv task. In *MediaEval*.

Larson, M., Ionescu, B., Sjöberg, M., Anguera, X., Poignant, J., Riegler, M., Eskevich, M., Hauff, C., Sutcliffe, R., Jones, G., Yang, Y.-H., Soleymani, M., and Papadopoulos, S. (2015). Working notes proceedings of the mediaeval 2015 workshop.

Le, N., Wu, D., Meignier, S., and Odobez, J.-M. (2015). Eumssi team at the mediaeval person discovery challenge. In *MediaEval*.

Lopez-Otero, P., Barros, R., Docio-Fernandez, L., González-Agulla, E., Alba-Castro, J., and Garcia-Mateo, C. (2015). Gtm-uvigo systems for person discovery task at mediaeval 2015. In *MediaEval*.

Martin, A., Garofolo, J., Fiscus, J., Le, A., Pallett, D., Przybocki, M., and Sanders, G. (2004). Nist language technology evaluation cookbook. In *LREC*.

Nishi, F., Inoue, N., and Shinoda, K. (2015). Combining audio features and visual i-vector at mediaeval 2015 multimodal person discovery in broadcast tv. In *MediaEval*.

Peters, C. and Braschler, M. (2002). The importance of evaluation for cross-language system development: the clef experience. In *LREC*.

Poignant, J., Bredin, H., and Barras, C. (2015a). Limsi at mediaeval 2015: Person discovery in broadcast tv task. In *MediaEval*.

Poignant, J., Bredin, H., and Barras, C. (2015b). Multimodal person discovery in broadcast tv at mediaeval 2015. In *MediaEval*.

Poignant, J., Budnik, M., Bredin, H., Barras, C., Stefas, M., Bruneau, P., Adda, G., Besacier, L., Ekenel, H., Francopoulo, G., Hernando, J., Mariani, J., Morros, R., Quénot, G., Rosset, S., and Tamisier, T. (2016). The CAMOMILE Collaborative Annotation Platform for Multi-modal, Multi-lingual and Multi-media Documents. In *LREC 2016*.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A., and Fei-Fei, L. (2015). Imagenet large scale visual recognition challenge. In *IJCV*.

Smeaton, A., Over, P., and Kraaij, W. (2006). Evaluation campaigns and trecvid. In *MIR*.

Yilmaz, E. and Aslam, J. (2006). Estimating average precision with incomplete and imperfect judgments. In *In Proceedings of the 15th ACM International Conference on Information and Knowledge Management*.