Overview of MediaEval 2011 Rich Speech Retrieval Task and Genre Tagging Task

Martha Larson[*]
m.a.larson@tudelft.nl

Maria Eskevich[†]
meskevich
@computing.dcu.ie

Roeland Ordelman[‡]
roeland.ordelman@utwente.nl

Christoph Kofler[*]
c.kofler@tudelft.nl

Sebastian Schmiedeke[§]
schmiedeke@nue.tu-berlin.de

Gareth J. F. Jones[†]
Gareth.Jones@
computing.dcu.ie

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Indexing methods*

General Terms

Measurement, Performance

1. INTRODUCTION

The MediaEval 2011 Rich Speech Retrieval Tasks and Genre Tagging Tasks are two new tasks offered in MediaEval 2011 that are designed to explore the development of techniques for semi-professional user generated content (SPUG). They both use the same data set: the MediaEval 2010 Wild Wild Web Tagging Task (ME10WWW). The ME10WWW data set contains Creative Commons licensed video collected from blip.tv in 2009. It was created by the PetaMedia Network of Excellence (http://www.petamedia.eu) in order to test retrieval algorithms for video content as it occurs 'in the wild' on the Internet and, in particular, for user contributed multimedia that is embedded within a social network. The data set was initially described in [3]. In this overview paper, we repeat the essential characteristics of the data set, describe the tasks and specify how they are evaluated.

2. THE ME10WWW BLIP.TV DATA SET

The ME10WWW data set consists of video episodes from blip.tv and a social network comprised of Twitter users tweeting about them. Videos were collected for shows for which the link to one of their episodes had been tweeted. Topsy (http://topsy.com) was used to collect blip.tv links from tweets. Their licenses were checked to confirm that they were Creative Commons and then the videos were downloaded from blip.tv. Topsy was then searched again to gather all users mentioning any one of the videos. We crawled the tweets and social network of these users for two steps (i.e., collected their interlocutors and their interlocutors's interlocutors). The data set contains 1974 episodes (247 development and 1727 test) comprising a total of ca. 350 hours of data. The development set is small with respect to the test set and is not intended for training, but rather for parameter tuning. The episodes were chosen from 460 different shows—shows with less than four episodes were not considered for inclusion in the data set. Each video is associated with metadata record including uploader assigned information (title, description, license, tags, uploaded ID/series ID).

The data set is accompanied by automatic speech recognition (ASR) transcripts [2], which were generously provided by LIMSI (http://www.limsi.fr/) and Vocapia Research (http://www.vocapia.com/) to MediaEval. In order to be included in the ME10WWW set, a video needed to have been transcribed by the ASR-system with an average word-level confidence score of > 0.7. The set is predominantly English with approximate 6 hours of non-English content divided over French, Spanish and Dutch. Specifically for 2011, LIMSI/Vocapia also provided a second set of "confusion networks", meaning that for a given time-point (time code), the transcripts may contain more than one hypotheses of the ASR-system.

The data set is also accompanied by the output of a shot detection system developed at the Technische Universität Berlin [1]. Both shot boundaries and extracted keyframes (one per shot) are included.

3. RICH SPEECH RETRIEVAL TASK

The Rich Speech Retrieval task is a known-item retrieval task and requires participants to return a ranked list of results in response to a query. The features used can be derived from speech, audio, visual content or metadata. The task can be considered to be a 'new generation' spoken content retrieval task in two respects. First, instead of requiring the identification of spoken documents or speech segments, the task requires the return of *jump-in* points, time points in the video at which users must start watching to view material relevant to the query. Second, the queries used express the user information need along three dimensions: the topical content, the speech act and the visual content.

The speech act dimension is, to our knowledge, investigated for the first time with this data set. When speakers speak, they are, on the one hand, pronouncing words, but on the other hand they are also actually 'doing' something. Treating spoken content in terms of 'illocutionary speech acts' (http://en.wikipedia.org/wiki/Speech_acts) emphasizes what speakers are accomplishing by speaking. The five speech acts used are 'apology', 'definition', 'opinion', 'promise' and 'warning'. These categories are similar to the 'speech-act

[*]Delft University of Technology, Netherlands

[†]Dublin City University, Ireland

[‡]Netherlands Institute for Sound & Vision and University of Twente, Netherlands

[§]Technische Universität Berlin, Germany

MediaEval 2011 Workshop, September 1-2, 2011, Pisa, Italy

like' units that have ben used in dialogue act modeling [5]. We are motivated to investigate this dimension because we conjecture that a connection exists between the reason for which a speaker makes an utterance and the reason for which a user would later search for the utterance. We assume that this connection could be used to improve spoken content retrieval systems.

The data set includes 30 queries associated with the development set and 50 queries associated with the test set. The form of queries includes a long form (`<title>`), a short form (`<short_title>`) and a label indicating the speech act type (`act`). An example of an apology is the following. Long form: 'How does Peter Busch, staff member of the Morning Swim Show, save face after the faux pas he made during his interview with Terry Denison?'. Short form: 'Peter Busch president chairman Denison morning Swim Show'. Although some queries in the data set make reference to visual characteristics (i.e., the speaker's clothing), most resemble this example and do not clearly need a visual contribution. Participants are required to make one submission for all test topics that uses only the provided ASR transcript (2010), and a second one using any combination of techniques which they believe will give the best overall performance. In the required run the participants must use the full queries.

The query set was created by having human annotators locate portions of the videos that they would be interested in sharing online and then formulating a comment on what the video portion was about (long form) and a query that would allow them to re-find jump-in points corresponding to those portions (short form). Online-sharing was chosen as the scenario, since it seemed to be the most natural, widely-understood reason for which users would be attempting to re-find previously seen video segments (i.e., known items). Access to a sufficient number of human annotators was secured by using a crowdsourcing platform, Amazon's Mechanical Turk (http://www.mturk.com). The task was carefully designed, the context of online video-sharing was clearly described and the language kept simple, i.e., "When you come across something interesting you might want to share on Facebook, Twitter or your favorite social network." Examples were given of quotes falling into each of the speech act categories. We note that 'opinions' were much more frequently located in the corpus than other categories such as 'warning' or 'promise'. The pairs of query+jump-in-point returned by the crowdsourcing workers were subjected to a standard quality control procedure (which eliminated ca. 40% of the results) and then by a screening for suitability and completeness (which eliminated a further ca. 40%).

The official evaluation metric of the task is mGAP, cf. [4], which generalizes the relevance of hypothesized jump-in points in relation to ground truth points by imposing a symmetric step-wise linearly decaying penalty function within a window of tolerance (10s, 30s, 60s windows are used). Since RSR is a known-item task, the metric is effectively a 'mGRR' (mean Generalize Reciprocal Rank).

4. GENRE TAGGING TASK

Genre information in the form of genre tags can provide valuable support for users searching and browsing the Internet for video. However, much video—and especially SPUG video—is not accurately or adequately tagged. This task attempts to automatically generate genre labels such as they are used to organize videos on video platforms such as blip.tv. Genre tagging is related to the genre classification task set out by Google as an ACM Multimedia Grand Challenge task in 2009 and 2010.

The Genre Tagging task requires participants to automatically assign genre tags to videos using features derived from speech, audio, visual content or associated textual or social information (Twitter social network). The tag set users are required to predict contains 26 genre tags.[1] Each video is associated with only one genre. Genre tags were collected for each episode using the API provided by blip.tv. A genre tag is represented by the field `categoryName` in the JSON output provided by the API. Whitespace in the genre tags was replaced by the underscore character ('_'), ampersands ('&') by the word 'and'.

Participants submit up to five runs in total representing five different approaches to the task (i.e., experimental conditions). They are required to complete one run using ASR transcripts only and one run including metadata. Within the metadata, we assume that user-assigned tags might already contain explicit genre information. Since we are interested in algorithms that work under the (likely) event that no tags are available, one run is also required with metadata, but no tags. Upon request, groups with a particular focus were excused from specific required runs. The official evaluation metric is MAP (Mean Average Precision).

5. ACKNOWLEDGMENTS

Thank you to Isabelle Ferrané (IRIT) who served as organizer of the Genre Tagging Task on the behalf of Quaero (http://www.quaero.org) and gave the task overview at the MediaEval 2011 workshop. Thank you to Thijs Verschoor (Twente) for his contribution to the RSR Task. This MediaEval effort received funding from EC FP7 NoE PetaMedia and Science Foundation Ireland RFP 2008.

6. REFERENCES

[1] P. Kelm, S. Schmiedeke, and T. Sikora. Feature-based video key frame extraction for low quality video sequences. In *WIAMIS '09*, pages 25–28, 2009.

[2] L. Lamel and J.-L. Gauvain. Speech processing for audio indexing. In *Advances in Natural Language Processing*, LNCS 5221, pages 4–15. Springer, 2008.

[3] M. Larson, M. Soleymani, P. Serdyukov, S. Rudinac, C. Wartena, V. Murdock, G. Friedland, R. Ordelman, and G. J. F. Jones. Automatic tagging and geotagging in video collections and communities. In *ACM ICMR '11*, pages 51:1–51:8, 2011.

[4] B. Liu and D. W. Oard. One-sided measures for evaluating ranked retrieval effectiveness with spontaneous conversational speech. In *SIGIR '06*, pages 673–674, 2006.

[5] A. Stolcke, N. Coccaro, R. Bates, P. Taylor, C. Van Ess-Dykema, K. Ries, E. Shriberg, D. Jurafsky, R. Martin, and M. Meteer. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Comput. Linguist.*, 26:339–373, September 2000.

[1] art, autos_and_vehicles, business, citizen_journalism, comedy, conferences_and_other_events, default_category, documentary, educational food_and_drink, gaming, health, literature, movies_and_television, music_and_entertainment, personal_or_auto-biographical, politics, religion, school_and_education, sports, technology, the_environment, the_mainstream_media, travel, videoblogging, web_development_and_sites