The Search and Hyperlinking Task at MediaEval 2013

Maria Eskevich[1], Robin Aly[2], Roeland Ordelman[2], Shu Chen[3], Gareth J.F. Jones[1]

[1]CNGL Centre for Global Intelligent Content, Dublin City University, Ireland
[2]University of Twente, The Netherlands
[3]INSIGHT Centre for Data Analytics, Dublin City University, Ireland
{meskevich,gjones} @computing.dcu.ie
shu.chen4@mail.dcu.ie
{r.aly, ordelman} @ewi.utwente.nl

ABSTRACT

The Search and Hyperlinking Task formed part of the MediaEval 2013 evaluation campaign. The Task consisted of two sub-tasks: (1) answering known-item queries from a collection of roughly 1200 hours of broadcast TV material, and (2) linking anchors within the known-item to other parts of the video collection. We provide an overview of the task and the data sets used.

1. INTRODUCTION

The increasing amount of digital multimedia content available is inspiring new scenarios of user interaction. The Search and Hyperlinking Task at MediaEval 2013 envisioned the following scenario: a user is searching for a segment of video that they know to be contained in a video collection (henceforth the target "known-item"). If the user finds the segment, he may wish to find additional information about some aspect of this segment. Computer systems should support users in this use scenario by providing links to satisfy the user's information needs. This use scenario is a refinement of a similar task at MediaEval 2012, see [4] for an overview of employed techniques. This paper describes the experimental data set provided to task participants for MediaEval 2013 and details of the two subtasks and their evaluation.

2. EXPERIMENTAL DATASET

The dataset for both subtasks was a collection of 1,260 hours of video provided by the BBC. The average length of a video was roughly 30 minutes and most videos were in the English language. The collection was used both for training and testing of systems. The BBC kindly provided human generated textual metadata and manual transcripts for each video. Participants were also provided with the output of two automatic speech recognition (ASR) systems and visual analysis. We describe these information sources in the following subsections.

2.1 Speech recognition transcripts

The audio was extracted from the video stream using the *ffmpeg* software toolbox (sample rate = 16,000Hz, number of channels = 1). Based on this data, two sets of ASR transcripts were created:

(i) All audio files were transcribed by LIMSI-CNRS/Vocapia[1] using the VoxSigma vrbs_trans system (version eng-usa_4.0) [7]. The models used by the system have been updated with partial support from the Quaero program [6].

(ii) The LIUM system[2] [10] is based on the CMU Sphinx project, and was developed to participate in the evaluation campaign of the International Workshop on Spoken Language Translation 2011. LIUM generated an English transcript for each audio file successfully processed. These results consist of: (i) one-best hypotheses in NIST CTM format, (ii) word lattices in SLF (HTK) format, following a 4-gram topology, and (iii) confusion networks, in an ATT FSM-like format.

2.2 Video cues

In addition to spoken content, visual descriptions of video content can potentially help for searching and hyperlinking. We provided the participants with shot boundaries, one extracted keyframe per shot, as well as the outputs of concept detectors (see below) and face detectors (see below) for these keyframes.

For each video, shot boundaries were determined and a single key frame per shot was extracted by a system kindly provided by Technicolor [8]. The extracted frame was the most stable I-frame within its shot. In total, the system extracted approximately 1,200,000 shots/keyframes. Concept detection scores for a list of concepts were provided. These concepts were selected by extracting keywords from metadata and spoken content. We used the on-the-fly video detector Visor, which was kindly provided by the Computer Vision Group of University of Oxford [2]. To make the confidence scores comparable over multiple detectors, we used them as variables in a logistic regression framework, which ensures the scores lie in the range $[0:1]$. We set the logistic regression parameters to the expected value of the parameters from over 374 detectors on the internet archive collection used in TRECVid 2011.

The appearance of faces in videos can be helpful information for search and linking. INRIA [3] kindly provided possible bounding boxes in keyframes with a confidence score that the bounding box contains a face. Additionally, the tool also contained for each bounding box, the n most similar faces (bounding boxes) in the dataset.

[1]http://www.vocapia.com/
[2]http://www-lium.univ-lemans.fr/en/content/language-and-speech-technology-lst

MediaEval 2013 Workshop, October 18-19, 2013, Barcelona, Spain

3. USER STUDY

For the definition of realistic queries and anchors, we conducted a study with 30 users between the ages of 18 and 30. By browsing the collection, the users selected items, a segment of a video with a start and an end time, that were interesting to them. The users were then instructed to consider these items as a known-item which they have to refind. We asked the users to formulate text and visual queries that they would use in a search engine to carry our their refinding. The study resulted in 50 known-items and corresponding multimodal queries. Subsequently, we asked the users to mark so-called anchors, or segments, related to other items from within the collection within the known-item for which they would like to see links. A second session of the study was conducted after the Task participants submitted their results. A set of users partially overlapping with the first group (17 participants) were presented with the selected anchors and with the hyperlinks proposed by the participants. The users had to assess the suitability of the proposed hyperlinks. Returning users assessed the anchors that they defined themselves. The reader can find a more elaborate description of this user study in [1].

4. SEARCH SUBTASK

We are interested in cross-comparison of one method being applied on all three types of transcripts. Thus we required the participants to submit up to 5 different approaches or their combinations, each being tested on all three transcripts.

We used the following three metrics in order to evaluate the submissions of the workshop participants: mean reciprocal rank (MRR), mean generalized average precision (mGAP) and mean average segment precision (MASP). MRR assesses the ranking of the relevant units. mGAP [9] rewards techniques that not only find the relevant items earlier in the ranked output list, but also are closer to the ideal point to begin playback (the "jump-in" point) of the relevant content. MASP [5] takes into account the ranking of the results and the length of both relevant and irrelevant segments that need to be listened to before reaching the relevant content.

5. LINKING SUBTASK

For the Hyperlinking subtask, the workshop participants were provided with the so-called anchors created by the users in the user study at the BBC and had to generate link targets. To be more concrete, the participants had to return a list of potential video segment link targets ranked by the likelihood of being relevant to the anchor or to the anchor in the context of corresponding known-item segment (though always independently of the initial known-item query).

To evaluate the linking subtask we used crowdsourcing via Amazon's Mechanical Turk platform[3], whereas the second stage of user study at BBC allowed us to assess the reliability of the crowdsourcing results.

Due to time and resource constraints, we chose a random subset of 30 anchors out of initial 98 for the formal task assessment. For these anchors and potential links, we used a pooling method to group the videos from the top 10 ranks of no more than 5 submitted runs of each of the participants. Submission were selected to maximize the diversity of the linking methods used in the pools to be assessed. This resulted in 9195 anchor-target pairs, that represented 7637 different pairs for crowdsoucing assessment. Users at BBC studies evaluated only 1 run per each participant which resulted in 2081 pairs, with 2078 being diverse. The manual assessment of these links resulted in the ground truth used to calculate precision at fixed rank cutoffs and MAP for all the participants runs. Both mturk and BBC ground truths were released to the participants for further performance analysis.

[3]www.mturk.com

6. ACKNOWLEDGEMENTS

This work was supported by Science Foundation Ireland (Grant 08/RFP/CMS1677) Research Frontiers Programme 2008 and (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (CNGL) project at DCU, and by the funding from the European Commission's 7th Framework Programme (FP7) under AXES ICT-269980. The user studies were executed in collaboration with Jana Eggink and Andy O'Dwyer from BBC Research, to whom the authors are greatful.

7. 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 *WWW (Companion Volume)*, pages 457–460, 2013.

[2] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman. The devil is in the details: an evaluation of recent feature encoding method. In *British Machine Vision Conference (BMVC 2011)*, Dundee, United Kingdom, 2011.

[3] R. G. Cinbis, J. Verbeek, and C. Schmid. Unsupervised Metric Learning for Face Identification in TV Video. In *Proceedings of ICCV 2011*, Barcelona, Spain, 2011.

[4] M. Eskevich, G. J. Jones, R. Aly, R. J. Ordelman, S. Chen, D. Nadeem, C. Guinaudeau, G. Gravier, P. Sébillot, T. de Nies, P. Debevere, R. Van de Walle, P. Galuscakova, P. Pecina, and M. Larson. Multimedia information seeking through search and hyperlinking. In *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval*, ICMR'13, pages 287–294, 2013.

[5] M. Eskevich, W. Magdy, and G. J. F. Jones. New metrics for meaningful evaluation of informally structured speech retrieval. In *Proceedings of ECIR 2012*, pages 170–181, Barcelona, Spain, 2012.

[6] J.-L. Gauvain. The Quaero Program: Multilingual and Multimedia Technologies. *IWSLT 2010*, 2010.

[7] J.-L. Gauvain, L. Lamel, and G. Adda. The LIMSI Broadcast News transcription system. *Speech Communication*, 37(1-2):89–108, 2002.

[8] A. Massoudi, F. Lefebvre, C. Demarty, L. Oisel, and B. Chupeau. A video fingerprint based on visual digest and local fingerprints. In *International Conference on Image Processing (ICIP 2006)*, pages 2297–2300, 2006.

[9] P. Pecina, P. Hoffmannova, G. J. F. Jones, Y. Zhang, and D. W. Oard. Overview of the CLEF 2007 cross-language speech retrieval track. In *Proceedings of CLEF 2007*, pages 674–686. Springer, 2007.

[10] A. Rousseau, F. Bougares, P. Deléglise, H. Schwenk, and Y. Estève. LIUM's systems for the IWSLT 2011 Speech Translation Tasks. In *Proceedings of IWSLT 2011*, San Francisco, USA, 2011.