The Search and Hyperlinking Task at MediaEval 2014

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

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

ABSTRACT

The Search and Hyperlinking Task at MediaEval 2014 is the third edition of this task. As in previous versions, it consisted of two sub-tasks: (i) answering search queries from a collection of roughly 2700 hours of BBC broadcast TV material, and (ii) linking anchor segments from within the videos to other target segments within the video collection. For MediaEval 2014, both sub-tasks were based on an ad-hoc retrieval scenario, and were evaluated using a pooling procedure across participants submissions with crowdsourcing relevance assessment using Amazon Mechanical Turk.

1. INTRODUCTION

The full value of the rapidly growing archives of newly produced digital multimedia content and digitalisation of previously created analog audio and video material will only be realized with the development of technologies that allow users to explore them through search and retrieval of potentially interesting content.

The Search and Hyperlinking Task at MediaEval 2014 envisioned the following scenario: a user is searching for relevant segments within a video collection that address a certain topic of interest expressed in a query. If the user finds a segment which is relevant to their initial information need expressed through the query, they may wish to find additional information about some aspect of this segment.

The task framework asks participants to create systems that support the search and linking aspects of the task. The use scenario is the same as in the Search and Hyperlinking task 2013 [4] with the main difference being that the search sub-task has changed from known-item to ad-hoc. This paper describes the experimental data set provided to task participants for MediaEval 2014, details of the two sub-tasks, and their evaluation.

2. EXPERIMENTAL DATASET

The dataset for both subtasks was a collection of 4021 hours of videos provided by the BBC, which we split into a development set of 1335 hours, which coincided with the test collection used in the 2013 edition of this task, and a test set of 2686 hours. The average length of a video was roughly 45 minutes, and most videos were in the English language. The test collection was broadcast content of date spans 01.04.2008 – 11.05.2008 and 12.05.2008 – 31.07.2008 for the development and test sets respectively. The BBC kindly provided human generated textual metadata and manual transcripts for each video. Participants were also provided with the output of several content analysis methods, which we describe in the following subsections.

2.1 Audio Analysis

The audio was extracted from the video stream using the *ffmpeg* software toolbox (sample rate = 16,000Hz, no. of channels = 1). Based on this data, the transcripts were created using the following ASR approaches and provided to participants:

(i) LIMSI-CNRS/Vocapia[1], which uses the VoxSigma vrbs_trans system (version eng-usa_4.0) [7]. Compared to the transcripts created for the 2013 edition of this task, the system's models had been updated with partial support from the Quaero program [6].

(ii) The LIUM system[2] [10], is based on the CMU Sphinx project. The LIUM system provided three output formats: (1) one-best transcripts in NIST CTM format, (2) word lattices in SLF (HTK) format, following a 4-gram topology, and (3) confusion networks in a format similar to ATT FSM.

(iii) The NST/Sheffield system[3] is trained on multi-genre sets of BBC data that does not overlap with the collection used for the task, and uses deep neural networks [8]. The ASR transcript contains speaker diarization, similar to the LIMSI-CNRS/Vocapia transcipts.

Additionally, prosodic features were extracted using the OpenSMILE tool version 2.0 rc1 [5][4]. The following list of prosodic features were calculated over sliding windows of 10 milliseconds: root mean squared (RMS) energy, loudness, probability of voicing, fundamental frequency (F0), harmonics to noise ratio (HNR), voice quality, and pitch direction (classes falling, flat, raising, and direction score). Prosodic information was provided for the first time in 2014 to encourage participants to explore its potential value for the Search and Hyperlinking sub-tasks.

MediaEval 2014 Workshop, October 17-18, 2014, Barcelona, Spain

[1] http://www.vocapia.com/

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

[3] http://www.natural-speech-technology.org

[4] http://opensmile.sourceforge.net/

2.2 Video Analysis

The computer vision groups at University of Leuven (KUL) and University of Oxford (OXU) provided the output of concept detectors for 1537 concepts from ImageNet[5] using different training approaches. The approach by KUL uses examples from ImageNet as positive examples [11], while OXU uses an on-the-fly concept detection approach, which downloads training examples through Google image search [3].

3. USER STUDY

In order to create realistic queries and anchors for our test set, we conducted a study with 28 users between aged between 18 and 30 from the general public around London, U.K. The study was similar to our previous study carried out for MediaEval 2013 [2], with the main difference being the focus on information needs with multiple relevant segments. The study focused on a home user scenario, and for this, to reflect the current wide usage of computer tablets, participants used a version of the AXES video search system [9] on *iPads* to search and browse within the video collection. The user study consisted of the following steps: i) a participant defined an information need using natural language, ii) they searched the test set with a shorter query, one they might use to search of the *Youtube* video repository, 3) after selecting several possible relevant segments, they defined anchor points or regions within each segment and stated what kind of links they would expect for this anchor.

Users were then instructed to define queries that they expected to have more than one relevant video segment in the collection. These queries consisted of several terms, and were used as input to a standard online search engine, e.g. "sightseeing london". The study resulted in 36 ad-hoc search queries for the test set. The development set for the task consisted of 50 known-item queries from the MediaEval 2013 Search and Hyperlinking task.

Subsequently, as in the 2013 studies, we asked the participants to mark so-called anchors, or segments they would like to see links to, within some the segments that are relevant to the issued search queries. The reader can find a more elaborate description of this user study design in [2].

4. REQUIRED RUNS SUBMISSIONS AND EVALUATION PROCEDURE FOR THE SEARCH AND LINKING SUB-TASKS

For the 2014 task, as well as ad hoc search, we were interested in cross-comparison of methods being applied across all four provided transcripts: one manual and 3 ASR. Thus, we allowed participants to submit up to 5 different approaches or their combinations, each being tested on all four transcripts, for both sub-tasks. In case any of the groups based their methods on video features only, they could submit this type of run in addition as well.

To evaluate the submissions of the search and linking subtasks a pooling method was used to select submitted segments and link targets for relevance assessment. The top-N ranks of all submitted runs were evaluated using crowdsourcing technologies. We report precision oriented metrics, such as precision at various cutoffs and mean average precision (MAP), using different approaches to take into account segment overlap, as described in [1].

5. ACKNOWLEDGEMENTS

This work was supported by Science Foundation Ireland (Grants 08/RFP/CMS1677 and 12/CE/I2267) and Research Frontiers Programme 2008 (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 grateful.

6. REFERENCES

[1] R. Aly, M. Eskevich, R. Ordelman, and G. J. F. Jones. Adapting binary information retrieval evaluation metrics for segment-based retrieval tasks. Technical Report 1312.1913, ArXiv e-prints, 2013.

[2] 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.

[3] K. Chatfield and A. Zisserman. Visor: Towards on-the-fly large-scale object category retrieval. In *Computer Vision–ACCV 2012*, pages 432–446. Springer, 2013.

[4] M. Eskevich, R. Aly, R. Ordelman, S. Chen, and G. J. F. Jones. The Search and Hyperlinking Task at MediaEval 2013. In *Proceedings of the MediaEval 2013 Workshop*, Barcelona, Spain, 2013.

[5] F. Eyben, F. Weninger, F. Gross, and B. Schuller. Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proceedings of ACM Multimedia 2013*, pages 835–838, Barcelona, Spain.

[6] J.-L. Gauvain. The Quaero Program: Multilingual and Multimedia Technologies. In *Proceedings of IWSLT 2010*, Paris, France, 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] P. Lanchantin, P. Bell, M. J. F. Gales, T. Hain, X. Liu, Y. Long, J. Quinnell, S. Renals, O. Saz, M. S. Seigel, P. Swietojanski, and P. C. Woodland. Automatic transcription of multi-genre media archives. In *Proceedings of the First Workshop on Speech, Language and Audio in Multimedia (SLAM@INTERSPEECH)*, volume 1012 of *CEUR Workshop Proceedings*, pages 26–31. CEUR-WS.org, 2013.

[9] K. McGuinness, R. Aly, K. Chatfield, O. Parkhi, R. Arandjelovic, M. Douze, M. Kemman, M. Kleppe, P. van der Kreeft, K. Macquarrie, A. Ozerov, N. E. O'Connor, F. De Jong, A. Zisserman, C. Schmid, and P. Perez. The AXES research video search system. In *Proceedings of the IEEE ICASSP 2014*.

[10] A. Rousseau, P. Deléglise, and Y. Estève. Enhancing the ted-lium corpus with selected data for language modeling and more ted talks. In *The 9th edition of the Language Resources and Evaluation Conference (LREC 2014)*, Reykjavik, Iceland, May 2014.

[11] T. Tommasi, T. Tuytelaars, and B. Caputo. A testbed for cross-dataset analysis. *CoRR*, abs/1402.5923, 2014.

[5] http://image-net.org/popularity_percentile_readme.html