Search and Hyperlinking Task at MediaEval 2012

Maria Eskevich, Gareth J.F. Jones, Shu Chen
Dublin City University, Ireland
{meskevich,gjones}@computing.dcu.ie
shu.chen4@mail.dcu.ie

Robin Aly, Roeland Ordelman
University of Twente, The Netherlands
{r.aly, ordelman}@ewi.utwente.nl

Martha Larson
Delft University of Technology
The Netherlands
m.a.larson@tudelft.nl

ABSTRACT

The Search and Hyperlinking Task was one of the Brave New Tasks at MediaEval 2012. The Task consisted of two subtasks which focused on search and linking in retrieval from a collection of semi-professional video content. These tasks followed up on research carried out within the MediaEval 2011 Rich Speech Retrieval (RSR) Task and the VideoCLEF 2009 Linking Task.

Categories and Subject Descriptors

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

General Terms

Measurement, Performance

1. INTRODUCTION

The increasing amount of digital multimedia content available is inspiring potential new types of user experience with the data. The Search and Hyperlinking Task at MediaEval 2012 envisioned the following scenario: a user is searching for a known segment in a video collection. On occasion the user may find that the information in the segment may not be sufficient to address their information need or they may wish to watch other related video segments. The Search and Hyperlinking Task thus required Search for a known relevant segment and Hyperlinking of related video segments.

This paper describes the experimental data set provided to task participants and details of the two subtasks.

2. EXPERIMENTAL DATASET

The dataset for both subtasks was the blip10000 collection of semi-professional videos crawled from the Internet video sharing platform Blip.tv. The collection contained the videos and user generated textual metadata available for each video. Participants were also provided with the following additional data: two automatic speech recognition (ASR) transcripts (containing the 1-best output; lattices, and confusion networks), automatically identified shot boundary points with extracted key frames for each shot, and automatic face and visual concept detection results.

2.1 blip10000

The blip10000 dataset was created by the PetaMedia NoE. Dataset creation involved collecting 15,449 Creative Commons videos from blip.tv, along with corresponding user provided metadata and associated Twitter social network data (i.e., who tweeted what about a particular video).

The final version of the dataset contained 14,838 videos, or episodes, comprising a total of ca. 3,260 hours of data. These episodes were separated into development and test sets, containing 5,288 videos (having a runtime of 1,135 hours) and 9,550 videos (having a runtime of 2,125 hours), respectively. The proportion of videos in the development and test set is ca. 1:3. Compared to the original dataset, the separation between development and test set is more balanced, enabling the direct application of both retrieval and classification approaches to address the task. These episodes were taken from 2,349 different shows.

2.2 Speech recognition transcripts

Audio was extracted from all videos using a combination of the *ffmpeg* and *sox* software (sample rate = 16,000Hz, number of channels = 1). Due to the absence of the audio signal for some videos or corruption in the files this extraction resulted in 5,237 files for the development set and 7,215 for the test set.

Two sets of ASR transcripts were provided for this data by LIMSI/Vocapia Research[1] and LIUM Research team[2].

The data is predominantly English, but there are also small numbers of Dutch, French and Spanish shows present.

LIMSI/Vocapia:.

Firstly LIMSI/Vocapia system [5] used a language identification detector (LID) to automatically identify the language spoken in the whole video along with a language confidence score (lconf), however the LID results were not manually checked.

Each file with a language identification score equal or greater than 0.8 was transcribed in the detected language. The remaining files were transcribed twice, with the detected language as well as with the English system. The average word confidence scores (tconf) were compared. The transcription with the higher score was chosen. For files in languages that do not have a transcription system, no transcripts were provided. This resulted in 5,238 and 7,216 files for development and test sets respectively.

LIUM:.

The LIUM sytem [9] is based on the CMU Sphinx project, and was developed to participate in the evaluation campaign

[1]http://www.vocapia.com/
[2]5http://www-lium.univ-lemans.fr/en/content/language-and-speech-technology-lst

MediaEval 2012 Workshop, October 4-5, 2012, Pisa, Italy

of the International Workshop on Spoken Language Translation 2011. LIUM provided an English transcription for each audio file successfully processed, that is 5,084 from the development set and 6,879 from the test set. 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.

2.3 Video clues

In addition to spoken content, visual descriptions of video content can potentially help for searching and hyperlinking. We therefore provided the participants with shot boundaries, and the outputs of concept-based and face-based analysis of the videos.

For each episode, the shot boundaries were created by TU Berlin [4]. For each shot segment, with a keyframe extracted from the middle each shot. In total, this dataset includes approximately 420,000 shots/keyframes with an average shot length of about 30 seconds.

Concept-based descriptors based on a list of 589 concepts selected by extracting keywords from the metadata of videos. We used the on-the-fly video detector Visor, which was kindly provided by the University of Oxford [1]. To make the confidence scores of a detector comparable over multiple detectors, we performed a logistic regression to bring the scores into the interval [0:1]. We set the linear regression parameters to the expected value over 374 detectors on the internet archive collection used in TRECVid 2010 [7].

The appearance of faces in videos can also be helpful information for search and hyperlinking. To encourage participants to use this information we provided face detection results, which were kindly provided by INRIA [2]. The data contained for possibly multiple bounding boxes in a keyframe, a confidence score stating the likelihood that this bounding box contains a face. Additionally, the data also contained the visual similarity between the faces in each bounding box.

3. SEARCH SUBTASK

The Search subtask was an extension of the Rich Speech Retrieval (RSR) Task at MediaEval 2011 [6]. Participants were provided with a set of queries for a known-item search task with 30 text queries each for the development and test sets. The queries were collected via crowdsourcing using the Amazon MTurk platform[3]. Participants were required to make one submission for the 1-best output of each ASR transcript version. Additionally, participants could deliver up to two other runs for each ASR transcript version.

3.1 Evaluation

We used three metrics in order to evaluate task results: mean reciprocal rank (MRR), mean generalized average precision (mGAP) and mean average segment precision (MASP). MRR assesses the ranking of the relevant units. mGAP [8] awards runs that not only find the relevant items earlier in the ranked output list, but also are closer to the jump-in point of the relevant content. MASP [3] 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 item.

[3]http://www.mturk.com/

4. LINKING SUBTASK

This subtask used the ground truth of the Search subtask as anchor videos from which links to other video should be formed. Participants were required to return a ranked list of video segments which were potentially relevant to the information in this video segment (independent of the initial textual query). Additionally they were allowed to use their own output of the Search subtask as the video anchors for the Linking subtask.

4.1 Evaluation

There was no ground truth created for this task beforehand. Top 10 ranks of the submitted result runs of all participanta were evaluated using the Mechanical Turk platform, and the ground truth based on the manual assessment of a combination of participants results was released for further error analysis. Once ground truth was collected, mean average precision (MAP) was used to compare the results.

5. ACKNOWLEDGEMENTS

This work was funded by a grant under the Science Foundation Ireland Research Frontiers Programme 2008 Grant No: 08/RFP/CMS1677, and by the funding from the European Commission's 7th Framework Programme (FP7) under AXES ICT-269980 and CUbRIK ICT-287704.

6. REFERENCES

[1] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman. The devil is in the details: an evaluation of recent feature encoding method. In *Proceedings of BMVC 2011*, 2011.

[2] 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.

[3] 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, 2012.

[4] P. Kelm, S. Schmiedeke, and T. Sikora. Feature-based video key frame extraction for low quality video sequences. In *Proceedings of WIAMIS 2009*, pages 25–28, 2009.

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

[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 *Proceedings of the MediaEval 2011 Workshop*.

[7] P. Over, G. Awad, J. Fiscus, B. Antonishek, M. Michel, A. Smeaton, W. Kraaij, and G. Quénot. Trecvid 2010–an overview of the goals, tasks, data, evaluation mechanisms, and metrics. 2011.

[8] 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.

[9] 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.