DCU Search Runs at MediaEval 2012: Search and Hyperlinking Task

Maria Eskevich Gareth J. F. Jones
Centre for Digital Video Processing & Centre for Next Generation Localisation
School of Computing, Dublin City University, Dublin 9, Ireland
{meskevich,gjones}@computing.dcu.ie

ABSTRACT

We describe the runs for our participation in the Search sub-task of the Search and Hyperlinking Task at MediaEval 2012. Our runs are designed to form a retrieval baseline by using time-based segmentation of audio transcripts incorporating pause information and a sliding window to define the retrieval segments boundaries with a standard language modelling information retrieval strategy. Using this baseline system runs based on transcripts provided by LIUM were better for all evaluation metrics, than those using transcripts provided by LIMSI.

Categories and Subject Descriptors

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

General Terms

Measurement, Performance

1. INTRODUCTION

The full potential of the constantly expanding archives of digital multimedia content can be only fully realised when effective retrieval technologies are available to enable users to locate interesting materials. The Search sub-task of the Speech and Hyperlinking Task at MediaEval 2012 was a continuation of the Rich Speech Retrieval (RSR) Task at MediaEval 2011 [5] with the objective of continuing to advance research in speech retrieval. The RSR Task explored spoken content search in conditions of highly varying semi- and non-professional videos from the Internet TV sharing platform blip.tv with the focus on queries that are associated with certain types of speech act.

While the MediaEval 2012 Search and Hyperlinking Task kept the same source of the dataset (blip.tv), the volume of the collection expanded to use the full blip10000 dataset consisting of 5,288 and 9,550 videos for the development and test sets respectively. In addition to the increased size of the test collection size other changes distinguish the research challenges of the MediaEval 2012 task. While the use of natural language textual queries remains a reasonable scenario for the task, the results of the previous RSR task showed that the relevance of the video segments was unaffected by the associated speech act, but rather that the relevant segment appeared more generally to have been chosen by the query creator as a highlight of an interesting video. The distinctive feature of the task remains the use of queries for which content is relevant primarily to the audio content with less focus on the visual information stream. Therefore the use of transcripts provided by automatic speech recognition (ASR) systems continued to be the primary source of indexing information. In 2012 two groups provided the ASR transcripts for the data: LIMSI/Vocapia [4] and LIUM [7], the transripts of the latter contained lattices and confusion networks in addition to the 1-best transcript, whereas the former had confusion networks and 1-best transcripts.

The task was evaluated using three metrics: mean reciprocal rank (MRR) which scores the rank of the retrieved segment containing relevant content, mean generalized average precision (mGAP) which combines the rank of the relevant segment and distance to the jump-in point at the start of the relevant content within the segment [6], and mean average segment precision (MASP) which combines the rank of the relevant segment with (ir)relevant length of the segment.[3].

Since our runs were planned to be rather baseline-style, we used the 1-best transcripts only. More details on the task dataset and evaluation metrics can be found in [2].

2. RETRIEVAL FRAMEWORK

As part of the task organizers group we focused our attention on carrying out baseline-style retrieval runs. Since the Search sub-task was a continuation of the RSR Task at MediaEval 2011 [5] we defined our runs based on the analysis of previous RSR task submissions.

In general for retrieval purposes the videos need to be segmented into smaller retrieval units which focus on distinct semantic elements. Analysis of runs submitted for the 2011 RSR task showed better performance for those using the pause information provided in the transcript as additional boundaries [1]. Another finding highlighted the usefullness of using a sliding window segmentation approach in order to identify segments that tend to capture semantically coherent content in a segment [8]. Thus we segmented each 1-best transcript as follows: (1) time segmentation into 180 seconds length segments, addition of the pause information (if pause between recognized words is > 0.5 seconds, a boundary is assigned); (2) time segmentation into 180 seconds length segments with a sliding window at a distance of 60 seconds from the end of each segment, addition of the pause information. This resulted in four index sets with varying numbers of segments. see Table 1. In general the LIUM transcripts

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

Table 1: Details of document specifications for submitted runs

Run parameters			No of	No of
Transcript type	Pause	Sliding window	documents	terms
LIUM	+	–	865130	57639
LIUM	+	+	914085	
LIMSI	+	–	331702	73601
LIMSI	+	+	355844	

Table 2: Evaluation results for submitted runs using alternative evaluation metrics

Run parameters			MRR			mGAP			MASP		
Transcript type	Pause	Sliding Window	60	30	10	60	30	10	60	30	10
LIUM	+	–	0.28	0.23	0.03	0.15	0.06	0.0	0.09	0.10	0.01
LIUM	+	+	0.27	0.22	0.02	0.15	0.05	0.0	0.08	0.09	0.01
LIMSI	+	–	0.13	0.09	0.03	0.07	0.04	0.03	0.03	0.03	0.03
LIMSI	+	+	0.10	0.06	0.02	0.04	0.02	0.01	0.006	0.006	0.003

were observed to contain more 0.5 seconds pauses, resulting in around three times more segments than for the LIMSI transcripts. At the same time we can see that the total number of distinct terms in the LIUM transcript is around 16,000 less, i.e. there is less variability in the transcription. However, it should be noted that the LIMSI system transcribed a number of files in languages other than English, and therefore the difference in size of vocabulary for the files in English language might be smaller. Unfortunately we do not know the the number of files in foreign languages and the number of files correctly labelled as non-English files. Thus this issue needs further investigation.

For retrieval we used the open-source Terrier information retrieval platform[1] with a standard language modelling method, with *lamda* equal to 0.35.

3. RESULTS, CONCLUSIONS AND FURTHER WORK

Table 2 shows results for all metrics. For all window sizes runs using the LIUM transcripts outperformed the LIMSI ones, while the additional boundaries added using the sliding window decreased performance for both transcripts, and more significantly for the LIMSI transcripts.

MRR highlights the fact that when the smaller window size is used (10 seconds), there is no difference in transcript performance. While MRR shows poor average ranking of the relevant segment, mGAP shows that when the smaller window size is used the distance to the beginning of the relevant content is smaller for LIMSI transcript based runs.

MASP was developed for ad-hoc search, for known-item search it simply reflects how much irrelevant content the user has to listen to before reaching one relevant segment.

Since the LIUM segments are on average three times shorter than the LIMSI ones, for the case of the window equal to 60 and 30 seconds (when ranking of LIUM runs is better than ranking of LIMSI runs, see MRR) LIUM-based runs score around three times higher than LIMSI ones. However in the case of the smallest window of 10 seconds, when the ranking is equally poor for all runs, the LIMSI-based runs show higher MASP and mGAP scores, meaning that these segmentations are closer to the jump-in point when assessed in those conditions.

The search sub-task used two types of transcript on the larger dataset for the first time. The reason for the difference in vocabulary size needs to be explained, since this may impact on the results. Although our baseline approach performed reasonably, comparison to other run submissions for the task shows that a more elaborate method of segmentation needs to be implemented.

[1] http://www.terrier.org

4. 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.

5. REFERENCES

[1] M. Eskevich and G. J. F. Jones. DCU at MediaEval 2011: Rich Speech Retrieval (RSR) Task. In *Proceedings of the MediaEval 2011 Workshop.*

[2] M. Eskevich, G. J. F. Jones, S. Chen, R. Aly, R. Ordelman, and M. Larson. Search and Hyperlinking Task at MediaEval 2012. In *Proceedings of the MediaEval 2012 Workshop.*

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

[5] 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.*

[6] 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.

[7] 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.

[8] C. Wartena and M. Larson. Rich Speech Retrieval Using Query Word Filter. In *Proceedings of the MediaEval 2011 Workshop.*