Time-based Segmentation and Use of Jump-in Points in DCU Search Runs at the Search and Hyperlinking Task at MediaEval 2013

Maria Eskevich
CNGL Centre for Global Intelligent Content
School of Computing
Dublin City University, Dublin, Ireland
meskevich@computing.dcu.ie

Gareth J.F. Jones
CNGL Centre for Global Intelligent Content
School of Computing
Dublin City University, Dublin, Ireland
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 2013. Our experiments investigate the affect of using information about speech segment boundaries and pauses on the effectiveness of retrieving jump-in points within the retrieved segments. We segment all three available types of transcripts (automatic ones provided by LIMSI/Vocapia and LIUM, and manual subtitles provided by BBC) into fixed-length time units, and present the resulting runs using the original segment starts and using the potential jump-in points. Our method for adjustment of the jump-in points achieves higher scores for all LIMSI/Vocapia, LIUM, and subtitles based runs.

1. INTRODUCTION

The constant growth in the size and variability of digital multimedia content being stored requires the development of techniques that not only identify files containing relevant content, but also bring the user as close as possible to the beginning of the relevant passage within this file to maximize the efficiency of information access. This starting point, referred to as the *jump-in* point, cannot simply be related to the locations of the words of interest being spoken, since the user may need to listen to the whole utterance in which the words were used, or slightly bigger passages, in order to get the idea of the context. Thus we assume that these jump-in points should occur at the beginning of the speech segments or utterances, and might be expressed by a pause in the speech signal. This idea underlies our experimental setup. We create one retrieval run for each fixed-length segmentation unit, but present it in two ways for further comparison: with the initial boundaries of the segments, and with adjusted jump-in points.

2. DATASET AND EVALUATION METRICS

The Search and Hyperlinking Task at MediaEval 2013 uses television broadcast data provided by BBC, and enhanced with varying additional content such as automatic speech recognition (ASR) transcripts [2]. The collection consists of circa 1260 hours of data that represent 6 weeks of broadcast content, including news programs, talk shows, episodes of TV series, etc. The 50 test set queries for the known-item retrieval task were created during user studies at the BBC [1].

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 ideal 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].

Table 1: Number of documents

Transcript	Window Size (seconds)		
Type	60	90	180
LIMSI	96 418	64 403	31 907
LIUM	95 091	63 308	31 210
Subtitles	82 220	54 698	26 742

3. RETRIEVAL FRAMEWORK

As the files in the collection vary in style and length, we decided to segment all the content into fixed length units. For these experiments we chose three values for segment length: 60, 90, and 180 seconds. These time units were the same for all types of transcripts. However, the transcripts do not always cover the spoken content in the same way: the ASR system might recognise some noise as words, or humans who create the manual subtitle transcripts might consider certain parts of the video non relevant for transcription. This explains the difference in the number of documents for diverse types of transcripts and time units given in Table 1.

At the segmentation stage we stored the information about potential jump-in points within each segment in a separate file. The LIMSI/Vocapia transcript contains speech segments boundaries predicted by their system [4]; whereas the LIUM transcript [7] has only time stamps for the words in the transcript; and manual subtitles have time stamps assigned on the utterance level. Thus for the official submission to the task we used as potential jump-in points the speech segments in the LIMSI transcript, pauses that are longer than 0.5 seconds between words in case of LIUM transcript and utterances in case of manual subtitles. Ad-

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

Table 2: Metric results (window = 60 seconds)

Run parameters			MRR			mGAP			MASP		
Transcript Type	Speech Segment	Pause	60	90	180	60	90	180	60	90	180
LIMSI	–	–	0.241	0.266	0.185	0.132	0.133	0.089	**0.142**	0.138	0.010
LIMSI	+	–	**0.250**	0.295	**0.240**	**0.151**	**0.164**	0.132	0.132	**0.146**	**0.124**
LIMSI*	–	+	**0.258**	**0.305**	**0.240**	0.150	0.153	**0.135**	0.139	0.145	**0.124**
LIUM	–	–	0.265	0.298	0.205	0.124	0.152	0.094	**0.140**	0.169	0.103
LIUM	–	+	**0.284**	**0.317**	**0.254**	**0.146**	**0.163**	0.114	**0.138**	**0.173**	**0.126**
Subtitles	–	–	0.343	0.369	0.217	0.209	0.191	0.0.092	**0.223**	0.231	0.093
Subtitles	–	+	**0.365**	**0.376**	**0.280**	**0.211**	**0.221**	**0.154**	0.212	**0.220**	**0.116**

Table 3: MRR results with varying window size (window size = 60, 30, 10 seconds)

Run parameters			Unit Size = 60 sec			Unit Size = 90 sec			Unit Size = 180 sec		
Transcript Type	Speech Segment	Pause	metric window			metric window			metric window		
			60	30	10	60	30	10	60	30	10
LIMSI	–	–	0.241	0.169	0.090	0.266	0.195	0.059	0.185	0.107	0.041
LIMSI	–	–	0.250	**0.223**	0.091	0.295	**0.226**	**0.110**	**0.240**	**0.178**	0.080
LIMSI*	–	+	**0.258**	0.194	**0.109**	**0.305**	**0.226**	0.090	**0.240**	0.175	**0.081**
LIUM	–	–	0.265	0.157	0.071	0.298	0.204	0.080	0.205	0.116	0.041
LIUM	–	+	**0.284**	**0.182**	**0.106**	**0.317**	**0.213**	**0.110**	**0.254**	**0.146**	**0.081**
Subtitles	–	–	0.343	0.273	0.144	0.369	0.255	0.096	0.217	0.113	0.042
Subtitles	–	+	**0.365**	**0.300**	**0.155**	**0.376**	**0.292**	**0.141**	**0.280**	**0.193**	**0.120**

ditionally we created an unofficial run that uses the pauses in the LIMSI transcript in order to be able to make a better comparison with the other types of transcript.

We do not have access to details of the ASR transcription systems. However we can distinguish them by the size of the vocabulary they used for this collection which contain 36,815, 57,259, and 98,332 entries for LIMSI/Vocapia, LIUM, and subtitles respectively.

For indexing and retrieval experiments we used the open-source Terrier Information Retrieval platform[1] [5] with a standard language modelling method, with default *lamda* value equal to 0.15.

4. RESULTS, CONCLUSIONS AND FURTHER WORK

Table 2 shows the results obtained for all three types of transcript and different segmentation unit size. The lines for the same transcript represent the same retrieval run, with the second line (and the third one for LIMSI) representing the enhanced result list. We highlight in bold the runs for which the addition of the jump-in point information increases the effectiveness of the results. In the case of LIMSI transcript there is no consistency that indicates whether the use of speech segments or pauses is preferable. This may be caused by the fact that sometimes these potential jump-in points coincide. For shorter segments the use of speech segments has better mGAP scoring, meaning that the speech segment based jump-in point brings the user closer to the beginning of the relevant content.

Table 3 shows the MRR results for varying window used in the metric calculation. The LIUM and subtitles based runs show the same trend of improvement of the score from the use of pause information in calculating the jump-in point.

These results allow us to argue that the simple time based segmentation results can be improved by use of pauses information contained in all types of transcripts.

5. ACKNOWLEDGMENTS

This work was supported by Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (CNGL) project at DCU.

6. 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] M. Eskevich, G. J. Jones, S. Chen, R. Aly, and R. Ordelman. The Search and Hyperlinking Task at MediaEval 2013. In *MediaEval 2013 Workshop*, Barcelona, Spain, October 18-19 2013.

[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] I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, and C. Lioma. Terrier: A High Performance and Scalable Information Retrieval Platform. In *Proceedings of ACM SIGIR'06 Workshop on Open Source Information Retrieval (OSIR 2006)*, 2006.

[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. Estèv. LIUM's systems for the IWSLT 2011 Speech Translation Tasks. In *Proceedings of IWSLT 2011*, San Francisco, USA, 2011.

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