Comparing Retrieval Effectiveness of Alternative Content Segmentation Methods for Internet Video Search

Maria Eskevich, Gareth J.F. Jones
Centre for Digital Video Processing
Dublin City University, Dublin 9, Ireland
{meskevich, gjones}@computing.dcu.ie

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

Christian Wartena
Univ. of Applied Sciences and Arts Hannover*
Hannover, Germany
christian.wartena@fh-hannover.de

Robin Aly, Thijs Verschoor, Roeland Ordelman
University Twente, P.O. Box 217,
7500AE Enschede, The Netherlands
{r.aly, t.verschoor, ordelman}@ewi.utwente.nl

Abstract

We present an exploratory study of the retrieval of semi-professional user-generated Internet video. The study is based on the MediaEval 2011 Rich Speech Retrieval (RSR) task for which the dataset was taken from the Internet sharing platform blip.tv, and search queries associated with specific speech acts occurring in the video. We compare results from three participant groups using: automatic speech recognition system transcript (ASR), metadata manually assigned to each video by the user who uploaded it, and their combination. RSR 2011 was a known-item search for a single manually identified ideal jump-in point in the video for each query where playback should begin. Retrieval effectiveness is measured using the MRR and mGAP metrics. Using different transcript segmentation methods the participants tried to maximize the rank of the relevant item and to locate the nearest match to the ideal jump-in point. Results indicate that best overall results are obtained for topically homogeneous segments which have a strong overlap with the relevant region associated with the jump-in point, and that use of metadata can be beneficial when segments are unfocused or cover more than one topic.

1 Introduction

The volume of online multimedia data continues to grow at an increasing rate. As such, realizing the potential of this material requires efficient access methods. While much effort has been expended on image-based video retrieval, we focus here on search based on the spoken content stream. Much of this data is complex often involving multiple speakers speaking in an informal, unstructured manner. In terms of user needs, searchers may be interested in finding not only content that matches the terms in their query, but also the specific context in which the terms were uttered and the speaker's intended meaning for these terms.

One important way that the meaning of spoken content can transcend the individual words is in terms of its function. For example, a speaker might use the same words, but in one case be warning listeners about something and in another case may be promising the listener to do something. These functions correspond to speech acts, i.e., the different purposes that the speaker is aiming to achieve[1]. Research in the area of dialogue systems and understanding has drawn from the overall typology of speech acts, concentrating on those that are helpful in conversation processing and analysis. The speech acts studied here are also drawn from a subset of this overall typology.

Specifically we focus on *illocutionary acts*, which correspond to the intended meaning of the utterance and are independent of the actual psychological impact on the listener. Particular segments might be interesting for a user because they provide not only factual information, but also give the possibility to listen to the essence of the speakers attitude towards the information in question. To the best of our knowledge analysis of spoken data in terms of speech acts has not been explored in previous work on spoken content retrieval (SCR).

The study described in this paper is based on the MediaEval 2011[2] [11] Rich Speech Retrieval (RSR) task [10]. The task focuses on finding the relevant jump-in point for each of a set of search queries each for a single manually

*At the time this work was done the author was affiliated with Novay, Enschede (The Netherlands) and Delft University of Technology.

[1] http://en.wikipedia.org/wiki/Speech acts

[2] http://www.multimediaeval.org/

identified illocutionary act, where the jump-in point is defined as the ideal time to start playback of relevant content.

The paper compares the results of extended submissions to the RSR 2011 task from three of the participating groups. These used different methods to segment transcripts of the content created using automatic speech recognition (ASR) and combination with manually created metadata in SCR systems to label a best predicted jump-in point for each query. Our study seeks to identify the factors underlying effective search for jump-in points, and those which can reduce search effectiveness.

The paper is structured as follows: Section 2 describes the data set, Section 3 overviews the content segmentation and retrieval methods used, Section 4 defines the evaluation metrics used, Section 5 gives results and analysis of retrieval behaviour, and Section 6 concludes.

2 Data overview

The MediaEval 2011 RSR dataset is drawn from the ME10WWW collection consisting of semi-professional user-generated videos from blip.tv [12].The RSR dataset used the test set portion of this data set which includes 1727 videos (ca. 300 hours of data), that are predominantly in English. Each video is associated with metadata manually added by the person uploading it to blip.tv which could include: title, description, license, tags, uploaded ID/series ID, and is accompanied by an ASR transcript [9].

The RSR 2011 task was a known-item search for which a test set of 50 queries associated with 5 types of illocutionary speech acts was collected using the Amazon Mechanical Turk (MTurk) platform[3] [3] : 'apology' (1), 'opinion' (21), 'definition' (17), 'promise' (5), and 'warning' (6) [10]. The MTurk workers were asked to browse videos to find examples of the specified speech acts, then to label beginning and end points of relevant regions, and to provide a manual transcript of this region. They then had to create 2 types of queries suitable for refinding this segment: a full query statement and a terse web search type query. We use only results using the short web queries in this study.

The words present in the relevant content or metadata do not always overlap with the query terms. On average the overlap with the manual transcripts is 0.30, with ASR transcript - 0.25, and with metadata - 0.22 (after standard stopword removal). 19 queries were found not to have any overlap at all with the ASR transcripts, and 15 had no overlap with the manual transcripts. For all types of speech acts, there were queries that have an overlap with metadata attached to the document containing the relevant passage.

[3]https://www.mturk.com/mturk/welcome

3 Content Processing and Information Retrieval

In this section we outline the content processing and information retrieval (IR) methods used for the three sets of submissions by RSR participants examined in this study. Since the blip.tv videos can vary in length from a few minutes to a few hours, all participants chose to segment them prior to retrieval in order to better identify the jump-in point within the content. The methods used to segment the ASR transcripts can be summarised as follows: segmentation into short segments of length approximately the same as that of an average relevant segment [18], segmentation into consecutive silence bounded utterances from the same speaker [1], and segmentation based on the lexical cohesion within the ASR transcript [4]. The following descriptions outline the segmentation and IR methods of each participant. In each case the predicted jump-in point for the segment is identified as the beginning of the segment.

3.1 Sliding Window (SW)

In this method the ASR transcripts were first processed to tag and lemmatize the words using Mark Hepple's [6] part-of-speech (POS) tagger. All closed class words (i.e., prepositions, articles, auxiliaries, particles, etc.) were then removed. To compensate for POS tagging errors, additionally English and Dutch stop words (standard Lucene search engine stopword lists[4]) were removed.

For the segmentation stage, fragments were defined as a sequence of sentences of n non-stop-words. In this investigation n = 20, 40 were used, more on n variations in [17]. Sentences were derived on the basis of punctuation (full-stop = sentence end) hypothesized by the ASR system and included in the transcript. If a sentence was less than the set number of words in length, subsequent sentences were added until it approximately meets this target.

For retrieval these segments were ranked using BM25 [16]. Since the segments created may overlap, the inverse document frequency idf (Eq.1) was calculated on the basis of the sentences.

$$\mathrm{idf}(t) \quad = \quad \log \frac{N - df_t + 0.5}{df_t + 0.5} \qquad (1)$$

where, N is the total number of fragments, and df_t is the number of fragments in which term t occurs. The weight of each term in each fragment-document is given by $w(d,t)$,

$$w(d,t) \quad = \quad \mathrm{idf}(t) \frac{(k+1) * f_{dt}}{f_{dt} + k * (1 - b + b * \frac{l_d}{\mathrm{avgdl}})} \qquad (2)$$

where f_{dt} is the number of occurrences of term t in document d, l_d is the length of d, and $avgdl$ is the average document length. The scaler constants were empirically set as

[4]http://lucene.apache.org/core/

Table 1. Overview of Runs

Retrieval Model	Use of data sources		
	ASR Transcript	Metadata	Combination of ASR Transcript and Metadata
BM25	SW_asr SW_asr_sh Sp_asr	Sp_meta	Sp_asr_meta
BM25F		SW_meta SW_meta_r	SW_asr_meta SW_asr_meta_sh
Language Model	LC_asr_c99 LC_asr_tt	LC_meta	LC_asr_meta_c99 LC_asr_meta_tt

k=2 and b=0.75. The ranking score for each segment was then calculated as the sum of the term weights of the matching terms between the query and the segment. Repetition of words in a query was ignored.

An initial ranking was created by ordering all segments by their matching score. A filter was then applied to remove all segments with a starting time within a window of 60 seconds of a higher ranked segment.

For the runs that combined ASR transcripts and metadata together, the BM25 extension known as, BM25F (fields) [16] was used. In BM25F for each document an f_t value was created for each term by forming a weighted sum of the frequency of term t in each field $f_t(i)$. For this sum the weight of terms in the ASR = 1.0 and in the metadata = 0.5.

The resulting six Sliding Window (SW) runs are shown in the Table 1: SW_asr (ASR transcript use only, n=40), SW_asr_sh (ASR transcript, n=20), SW_meta (metadata use only, the whole video used as a document), SW_meta_r (metadata use only, each jump-in points within the video are at least 60 seconds ahead of the previous one and at the beginning of a sentence), SW_asr_meta (use of both ASR transcript and metadata), and SW_asr_meta (use both ASR transcript and metadata, n=20).

3.2 Speech Segments (Sp)

The second method for creating segments for the IR stage involved using speech fragments between silence points identified by the ASR system. This process simply formed search segments by concatenating consecutive fragments from the same speaker as labeled by the ASR system.

Experiments were carried out using the concatenated ASR segments and the metadata. For each query a separate ranking was generated for the separate evidence sources using BM25 [16]. The matching score for each evidence source was then combined using a simple weighted sum [13] to form the final matching score of the segment. The search engine PFTijah [8] was used for ranking.

This approach resulted in three runs shown in Table 1: Sp_asr (segments of concatenated ASR transcript fragments), Sp_meta (document metadata only, with each entry expanded to a list of all speech segments found in this document), Sp_asr_meta (combination of ASR transcript speech segments and metadata scores). For Sp_asr_meta, the matching scores of each source were weighted equally.

3.3 Lexical cohesion based segmentation (LC)

The third method used lexical cohesion based segmentation methods to divide the ASR transcripts into topically focused segments that were then used as the search units.

Two lexical cohesion based segmentation algorithms developed for text segmentation were explored: C99 [2] and TextTiling [5]. TextTiling computes the cosine similarity between adjacent fixed sized blocks of sentences. The C99 algorithm also calculates the similarity between sentences using a cosine similarity measure to form a similarity matrix with the cosine scores then replaced by the rank of the score in the local region and segmentation points assigned using a clustering procedure. Both of the algorithms work with the fundamental unit of the sentence. For these experiments the punctuation inserted by the ASR system was used as the sentence boundaries. Based on preliminary experiments, additional segmentation points were inserted where gaps of more than 0.5 seconds were observed between words, since these are likely to indicate points of topical change.

Retrieval experiments used a version of the SMART IR system[5] extended to use language modelling (a multinomial model with Jelinek-Mercer smoothing) with a uniform document prior probability [7]. Equation 3 shows how a query q is scored against a document d within SMART.

$$P(q|d) = \prod_{i=1}^{n} (\lambda_i P(q_i|d) + (1 - \lambda_i) P(q_i)) \qquad (3)$$

where $q = (q_1, \ldots q_n)$ is the query comprising of n query terms, $P(q_i|d)$ is the probability of generating the i^{th} query term from a given document d being estimated by the maximum likelihood, and $P(q_i)$ is the probability of generating it from the collection and is estimated by document frequency. The retrieval model used $\lambda_i = 0.3$ for all q_i, the particular value being optimized on the TREC-8 dataset. Stopwords were removed using the standard SMART stopword list. Words were stemmed using a variant of the Lovins stemmer [14], packaged in SMART by default.

This method gave five Lexical Cohesion based (LC) runs shown in Table 1: LC_asr_c99 and LC_asr_tt (ASR transcript segmented with C99 and TextTiling algorithms), LC_meta (metadata use only, with the whole video used as a document), LC_asr_meta_c99 and LC_asr_meta_tt (combination of ASR transcript segmented with C99 or TextTiling algorithms with the metadata). For the runs with metadata,

[5] ftp://ftp.cs.cornell.edu/pub/smart/

Table 2. Mean Reciprocal Rank (MRR) and mean Generalized Average Precision (mGAP)

RunName	WindowSize					
	60		30		10	
	MRR	mGAP	MRR	mGAP	MRR	mGAP
SW_asr	0.38	0.30	0.34	0.22	0.10	0.10
SW_asr_meta	0.39	0.33	0.39	0.28	0.15	0.15
SW_asr_sh	0.37	0.32	0.32	0.27	0.19	0.19
SW_asr_meta_sh	0.35	0.29	0.30	0.25	0.14	0.14
SW_meta	0.20	0.15	0.18	0.13	0.06	0.06
SW_meta_r	0.16	0.11	0.11	0.08	0.02	0.02
Sp_asr	0.34	0.27	0.27	0.22	0.16	0.16
Sp_asr_meta	0.34	0.25	0.26	0.21	0.15	0.15
Sp_meta	0.18	0.14	0.14	0.11	0.07	0.07
LC_asr_c99	0.28	0.19	0.21	0.15	0.07	0.07
LC_asr_meta_c99	0.33	0.21	0.23	0.16	0.08	0.08
LC_asr_tt	0.36	0.25	0.29	0.18	0.09	0.09
LC_asr_meta_tt	0.39	0.28	0.30	0.20	0.14	0.14
LC_meta	0.18	0.11	0.09	0.07	0.03	0.03

Table 3. Average Precision and Recall, Window size = 60 sec

Run Name	AVR Precision in time	AVR Recall in time
SW_asr	0.2103	0.3787
SW_asr_meta	0.2385	0.4337
SW_asr_meta_sh	0.2790	0.2549
SW_asr_sh	0.2741	0.2586
Sp_asr	0.2188	0.5074
Sp_asr_meta	0.2188	0.5074
LC_asr_c99	0.2326	0.5147
LC_asr_meta_c99	0.2023	0.4701
LC_asr_tt	0.2592	0.4867
LC_asr_meta_tt	0.2602	0.4898

the metadata of the video was added to each segment prior to indexing.

4 Evaluation metrics

Search effectiveness in the RSR 2011 task was evaluated in terms of the position of the predicted jump-in point in the segment against a manually labeled gold standard jump-in point, and is intended to measure the effort of the user in locating the jump-in point when browsing the retrieved video. This is important due to the inefficient nature of audio browsing compared to that of reading retrieved text documents. Since the RSR 2011 task was a known-item search, one useful evaluation metric is the Mean Reciprocal Rank (MRR); additionally we apply a metric that evaluates the ranking and takes account of the distance between the predicted and actual jump-in point (mean Generalized Average Precision (mGAP)) [15].

Although the official RSR 2011 measure was mGAP, our systems return both start and end points for each segment. This enables us to examine the segment-level precision and recall in terms of relevant data contained within each segment, and to explore their relationship to search behaviour.

We next give the formal definitions of MRR and mGAP:

Mean Reciprocal Rank (MRR) The reciprocal rank of a query response is the multiplicative inverse of the rank of the first correct answer. The mean reciprocal rank is the average of the reciprocal ranks of results for a sample of queries Q (Equation 4).

$$MRR = \frac{1}{|Q|} . \sum_{i=1}^{|Q|} \frac{1}{rank_i} \tag{4}$$

Mean Generalized Average Precision (mGAP) mGAP generalizes the relevance of hypothesized jump-in points in relation to ground truth points by imposing a symmetric step-wise linearly decaying penalty function within a window of tolerance (10s, 30s, 60s windows are used). Since RSR 2011 was a known-item task, the metric is effectively a 'mGRR' (mean Generalized Reciprocal Rank). The calculation of GAP for a single query is,

$$GAP = \frac{1}{n} . \sum_{r=1}^{N} P[r] \cdot \left(1 - \frac{Penalty \cdot Granularity}{Window}\right) \tag{5}$$

where $P[r]$ is the precision at rank r, $Granularity$ is the step that is used to measure how far the retrieved jump-in point is from the relevant one and to penalize the result for longer distance ($Granularity$ = 10 seconds in this experiment); $Window$ is the distance before and after the beginning of a relevant segment that the result should fit in, in order to be considered correctly retrieved; $Penalty$ is the number of times the user has to move in time within the $Window$ with the $Granularity$ step, in order to get to the actual relevant jump-in point. Thus segments that make the user wait for longer than $Window$ size before and after the actual relevant jump-in point are not considered relevant.

5 Results and Analysis

Table 2 shows MRR and mGAP scores for all the runs with window sizes of 60, 30, and 10 seconds. As would be expected, smaller window size decreases the scores, however the trend of the method effectiveness remains the same. When runs have a larger decrease between MRR and mGAP, the start point of the segment containing the relevant content is further from the jump-in point. For example

MRR for LC_asr_tt and SW_asr_sh is 0.36 and 0.37 (window size = 60s), but mGAP is 0.25 and 0.32, meaning that second run segments begin closer to the jump-in point.

Runs using only metadata (SW_meta, SW_meta_r, LC_meta, and Sp_meta) have lower performance with the same retrieval settings, and these differences are statistically significant as measured by Wilcoxon Test with confidence 95%. The difference between runs using ASR transcripts only and those combining them with metadata is not statistically significant for any of the retrieval frameworks. Since the runs using only metadata for retrieval show low retrieval scores, we omit them from the following discussion.

Table 3 shows the average precision and recall for the content in the individual segment containing the jump-in point. The mGAP metric is designed to reward approaches that identify the beginning of the relevant segment better. From the results, it can be seen that runs having higher mGAP values also have higher precision and lower recall values. However analysis of the results per query shows that the recall within the segment is important for the ranking of segments containing relevant content.

5.1 Relationship Between Retrieval Effectiveness and Segmentation Methods

This section presents an analysis of the results in terms of the relationship between retrieval behaviour and segmentation methods. We focus on the behaviour of the vocabulary of individual queries and content of the segment units. A more quantitative study analyzing different aspects of the text and the segments can be found in [17].

5.1.1 Non-zero overlap of query words with ASR transcript

For runs where there is non-zero overlap of the terms in a query and relevant content, there is no direct correlation between the ASR word error rate (WER) and the retrieval results. This is observed to be because for a segment containing relevant content to achieve high ranks requires good topic segmentation around the relevant region: if the non-relevant content present in the retrieved segment belongs to the same topic, a segment even with an ASR WER of 55, 46, 44 % is found to be retrieved at the 1st or 2nd rank, whereas when the same relevant content is contained in a segment that also contains an adjacent part of the transcript relating to a different topic, its retrieval rank is much lower. For example, cf. Table 4, for query 36 with WER = 46 %, LC_asr_tt and SW_asr runs have 100 % recall, high precision (30 and 56 %), and retrieve the relevant content at the 1st rank, whereas segments created by LC_asr_c99 and Sp_asr, although containing all the relevant content, have lower precision and cover one or more additional topics, thus the segment is observed to be found at a lower rank. The case of query 6, WER = 62 %, is an even stronger example because SW_asr, LC_asr_c99, and LC_asr_tt runs have the same level of precision (23 %). The segments of SW and LC runs overlap only within the relevant content: SW_asr segment begins with relevant content and includes subsequent non-relevant content on a different topic, the LC_asr_c99 and LC_asr_tt segments start before the relevant content, but cover the same topic. The non-relevant part of the SW_asr segment contains a change of the topic, which causes a reduction in retrieval rank to the15th position. The same trend is observed for queries with much lower ASR WER (even when it is equal to 0 %, as in case of query 24).

Query 24 shows another effect of transcript segmentation on the results ranking. When the relevant content is divided between two segments, this influences the retrieval results. The LC_asr_c99 segment is counted in the mGAP metric because the segment starts within a window of 30 seconds, however its ranking is lower due to lower recall.

5.1.2 Zero-overlap of query words words with ASR transcript

For 10 queries there is no overlap in content words with either the metadata or ASR transcript, and 9 queries that have an overlap with metadata, but not with the ASR transcript. For the first type of queries retrieval results depend strongly on the segmentation of the area surrounding the relevant content. If the topic of the discussion stays similar or the same to the relevant topic, then these segments are usually retrieved at the top of the list, and then if these segments are within the window used for mGAP metric, they are taken into account positively in the mGAP score, even though precision and recall of the retrieved segment are 0%. Results of runs for queries with an overlap with metadata, but not with the ASR transcript, are not affected by the use of metadata when the surrounding text in the segment relates to another topic or has vocabulary different to that of the query. In cases when the surrounding segments are on the same topic and fit within the mGAP window, then addition of the metadata is found to decrease the results.

5.1.3 Non-zero overlap of query words with both ASR transcript and metadata

Cases where the runs that have high recall of relevant content in segments cover more than one topic, are ranked better when metadata is added, see for example run SW_asr_meta versus SW_asr for query 6.

However, when the relevant segment is very short (query 46, type 'warning', 'Dear democrats! don't count your chickens before they hatch') and the query itself is very vague and short ('democrats'), even 100% overlap with the ASR transcript and metadata does not help in the retrieval.

Table 4. Example of MRR, Precision, Recall results for queries with different ASR WER

	query 24, WER = 0 %			query 36, WER = 46 %			query 6, WER = 62 %		
Run Name	MRR	Precision	Recall	MRR	Precision	Recall	MRR	Precision	Recall
SW_asr	1.0	0.37	1.0	1.0	0.56	1.0	0.07	0.23	0.83
Sw_asr_meta	1.0	0.37	1.0	1.0	0.56	1.0	1.0	0.23	0.83
Sp_asr	1.0	0.73	1.0	0.003	0.13	1.0	0.14	0.22	1.0
Sp_asr_meta	0.13	0.73	1.0	0.002	0.13	1.0	1.0	0.22	1.0
LC_asr_c99	0.14	0.16	0.45	0.11	0.13	1.0	1.0	0.23	1.0
LC_asr_c99_meta	0.50	0.16	0.45	0.16	0.13	1.0	0.5	0.0	0.0
LC_asr_tt	1.0	0.22	1.0	1.0	0.30	1.0	1.0	0.23	1.0
LC_asr_tt_meta	1.0	0.22	1.0	1.0	0.30	1.0	0.33	0.0	0.0

6 Conclusions and Further Work

In this paper we described an investigation of the retrieval behaviour of multiple approaches to retrieval of semi-professional user-generated Internet video for queries associated with relevant content of a specific speech act. The small size of the query set does not allow us to draw general conclusions on the difference based on the speech act type. However we can state that the segmentation of the content plays a significant role in retrieving relevant content. When the segments have high recall and precision, and the rest of the segment belongs to the same topic, the different ranking methods all succeed in ranking the relevant content highly. We can also conclude that related metadata is useful when a segment (with high recall and non relevant content, or with low recall) is ranked low in the list. These results clearly indicate the need for further research on ASR segmentation for SCR, and exploration of methods to better exploit metadata in supporting speech search.

7. Acknowledgments

This work is funded by a grant under the Science Foundation Ireland Research Frontiers Programme 2008 Grant No: 08/RFP/CMS1677, by the funding from the European Commission's 7th Framework Programme (FP7) under grant agreement no. 216444 (EU PetaMedia Network of Excellence), and AXES ICT-269980. .

References

[1] R. Aly, T. Verschoor, and R. Ordelman. UTwente does Rich Speech Retrieval at mediaEval 2011. In Larson et al. [11].

[2] F. Y. Y. Choi. Advances in domain independent linear text segmentation. In *Proceedings of NAACL 2000*, 2000.

[3] M. Eskevich, G. J. Jones, M. Larson, and R. Ordelman. Creating a data collection for evaluating rich speech retrieval. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 2012.

[4] M. Eskevich and G. J. F. Jones. DCU at MediaEval 2011: Rich Speech Retrieval. In Larson et al. [11].

[5] M. Hearst. Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64, 1997.

[6] M. Hepple. Independence and commitment: Assumptions for rapid training and execution of rule-based POS taggers. In *Proceedings of ACL 2000*, Hong Kong, 2000.

[7] D. Hiemstra. *Using language models for information retrieval*. PhD thesis, University of Twente, 2001.

[8] D. Hiemstra, H. Rode, R. V. Os, and J. Flokstra. PFTijah: text search in an xml database system. In *Proceedings of OSIR 2006*, pages 12–17, 2006.

[9] L. Lamel and J.-L. Gauvain. Speech processing for audio indexing. In *Advances in Natural Language Processing*, 2008.

[10] 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 Larson et al. [11].

[11] M. Larson, A. Rae, C.-H. Demarty, C. Kofler, F. Metze, R. Troncy, V. Mezaris, and G. J. F. Jones, editors. *Working Notes Proceedings of the MediaEval 2011 Workshop, Santa Croce in Fossabanda, Pisa, Italy*, volume 807 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2011.

[12] M. Larson, M. Soleymani, M. Eskevich, P. Serdyukov, R. Ordelman, and G. J. Jones. The community and the crowd: Developing large-scale data collections for multimedia benchmarking. *IEEE Multimedia Magazine*, to appear 2012.

[13] J. H. Lee. Analyses of multiple evidence combination. In *Proceedings of ACM SIGIR'97*, pages 267–276, 1997.

[14] J. B. Lovins. Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics*, 11, 1968.

[15] 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 the CLEF 2007 Workshop*, pages 674–686, 2007.

[16] S. Robertson, H. Zaragoza, and M. Taylor. Simple bm25 extension to multiple weighted fields. In *ACM CIKM 2004*, 2004.

[17] C. Wartena. Comparing segmentation strategies for efficient video passage retrieval. In *10th CBMI Workshop 2012*, 2012.

[18] C. Wartena and M. Larson. Rich Speech Retrieval using query word filter. In Larson et al. [11].