Segmentation Strategies for Passage Retrieval from Internet Video using Speech Transcripts

Christian Wartena
Department of Information and Communication
Hochschule Hannover – University of Applied Sciences and Arts
Expo Plaza 12
30539 Hannover
Germany
Chrisian.Wartena@hs-hannover.de

Journal of Digital
Information Management

ABSTRACT: *We compare the effect of different segmentation strategies for passage retrieval of user generated internet video. We consider retrieval of passages for rather abstract and complex queries that go beyond finding a certain object or constellation of objects in the visual channel. Hence the retrieval methods have to rely heavily on the recognized speech. Passage retrieval has mainly been studied to improve document retrieval and to enable question answering. In these domains best results were obtained using passages defined by the paragraph structure of the source documents or by using arbitrary overlapping passages. For the retrieval of relevant passages in a video no author defined paragraph structure is available. We compare retrieval results from 5 different types of segments: segments defined by shot boundaries, prosodic segments, fixed length segments, a sliding window and semantically coherent segments based on speech transcripts. We evaluated the methods on the corpus of the MediaEval 2011 Rich Speech Retrieval task. Our main conclusions are (1) that fixed length and coherent segments are clearly superior to segments based on speaker turns or shot boundaries; (2) that the retrieval results highly depend on the right choice for the segment length; and (3) that results using the segmentation into semantically coherent parts depend much less on the segment length. Especially, the quality of fixed length and sliding window segmentation drops fast when the segment length increases, while quality of the semantically coherent segments is much more stable. Thus, if coherent segments are defined, longer segments can be used and consequently fewer segments have to be considered at retrieval time.*

Categories and Subject Descriptors:
I.2.7[Artificial Intelligence]: Natural Language Processing – Text Analysis; **I.2.10 [Artificial Intelligence]**: Vision and Scene Understanding – Video Analysis; **H.3.3 [Information Storage and Retrieval]** Information Search and Retrieval - Search Process

General Terms:
Video Analysis, Image Segmentation, Passage Retrieval

Keywords: Text Segmentation, Video Segmentation, Multimedia Retrieval, User Generated Content

Received: 11 August 2013, Revised 15 September 2013, Accepted 19 September 2013

1. Introduction

Video content represents a fast growing part of the total amount of internet content. Audio-visual content is not restricted to entertainment, but includes also video lectures, instructional videos, interviews, documentaries and so on. Users in many cases do not watch these videos linearly but watch just selected fragments [19]. Thus we need methods to browse and search within a video and to find the relevant parts of a video. The retrieval of the relevant fragments or jump-in points is called passage retrieval. Passage retrieval raises a number of interesting questions like the relation between the passages and the video as a whole and the questions of determining the right segment boundaries and the relevant segments given some information need. It is this latter question that we address in this paper.

Segmentation of video and retrieval of fragments from video has been studied extensively with queries for objects that are depicted in the visual channel of the video, like "*vendor behind counter*" or "*elderly woman*" [10]. The type of queries we are considering is completely different and the relevance of the answer is not necessarily dependent on what is depicted in the visual channel. The queries we are considering are much more abstract and complex and the correct answer might be a longer passage consisting of multiple shots. An example of such a query is: "*How to right-click on a tablet PC* ". Here the user does not just want to find a picture of a tablet PC, but also wants the explanation about the right click. In fact, one could imagine, that the user does not bother, whether the answer is a video or a text.

For passage retrieval of written texts often the formatting of the text, as defined by the author is used as a base for defining passages [3] [6] [17]. Either the chapters or paragraphs are used directly as retrieval units or paragraphs are merged into larger units. Adjacent paragraphs are merged either because their topics are very similar or in other approaches simply in order to define passages of more or less constant length. For segmentation of video this type of formatting information usually is not available.

For retrieval and browsing video lectures usually the accompanying slides or textbook are used as a sole or additional source to segment the video stream [23] [9] [13]. In general, however, we do not have slides and slide transitions and we have to rely on the information that can directly be obtained from the audio and the video signal. In the following we will focus on segmentation and retrieval based on the video itself without the use of additional information. We will use the audio signal and the video signal for segmentation and the audio channel, especially the (automatic) speech transcripts, for retrieval.

There are basically three possibilities for the segmentation of a video: (1) using visual information, like shot boundaries, (2) using information from the audio channel; either prosodic information, like intonation, pauses and speaker turns or advanced segmentation techniques that find lexically (and semantically) coherent segments; or (3) using segments of fixed length. For the fixed length segments there are again two variants. Either the document is split up in segments of (almost) equal length, or a sliding window is used, defining many overlapping segments. Probably by the lack of good evaluation corpora, we find hardly any literature comparing the effectiveness of these methods for passage retrieval directly. Indirectly, passage retrieval is evaluated by its use for question answering and for improving document retrieval. Results in these domains suggest that the sliding window strategy performs best.

In the current study we compare the five segmentation strategies for passage retrieval on a video corpus of the MediaEval 2011 Rich Speech Retrieval task (GenreRSR2011). Thus we do not evaluate the quality of the segments directly, but indirectly by their usefulness in a passage retrieval task. In fact we even don't consider the whole segment for evaluation but only evaluate the start time of the segment, as it is the user himself who decides to pursue the listening or not.

We find that fixed length (either overlapping or non-overlapping) passages give best results. However, these results depend on the right estimation of the optimal segment length. Results of the coherent segment strategy are almost as good, but depend less on the choice of the correct parameter. Especially, the results remain stable for longer segments. Thus, the coherent segments give the possibility to work with longer and consequently less segments. In the present paper we focus on a quantitative analysis of the effects of segment length for a number of segmentation strategies. A more qualitative study analyzing different aspects of the text and the segments can be found in [1]. Some of the results presented in this paper were presented before at CMBI 2012 [20].

The remainder of the paper is organized as follows. In section 2we discuss related work. In section 3 we introduce the data we have used to test the segmentation strategies. Subsequently we sketch our general approach for passage retrieval. In section 4 we discuss the compared segmentation methods in more detail and show how they are applied to the used data set. Section 5 gives the results of the experiment. We finish the paper with a discussion and outlook for future work.

2. Related Work

Segmentation of spontaneous or planned speech has been studied mainly for lecture videos. The quality of the segmentation of these videos is usually accessed by a comparison with an available ground truth [23] [9] [13]. We are not aware of any evaluation of segmentation strategies in the context of passage retrieval for this type of data. Moreover most research on lecture video segmentation uses additional sources of information.

Since the early nineteens' passage retrieval for written text has received a lot of attention [16]. However, in most work passage retrieval is used to improve document retrieval [21] [3] [11] to improve query expansion [22] or it is used as an intermediate result for question answering [15] [17].

[3] introduce the text tiling algorithm that defines lexically coherent segments. They base document retrieval in various ways on passage retrieval. They report that the text tiling strategy outperforms fixed lengths segmentation. However, no significant differences are found with retrieval results based on the paragraph structure of the documents. [6] also compare effectiveness of different segmentation strategies for document retrieval based on passage retrieval. They introduce a further segmentation strategy with overlapping segments that does not only use a sliding window but also considers windows of different size. Kaszkiel and Zobel find that this strategy, that they call arbitrary segmentation, gives best results. Arbitrary segmentation gives also best results for question answering in experiments described by [17]. In this study no lexically coherent segmentation is evaluated. In another study [18] include text tiling but do not consider arbitrary segmentation. Now the sliding window supports the question answering task best.

The various studies all indicate that paragraph structure, if available, works very well. Best results are generally obtained with very flexible and redundant segmentation: the sliding window or arbitrary segments approach. Lexically coherent segments seem to have no advantages, but it should be noted that in all cases the text tiling algorithm [2] was used and that no variations of the granularity with which the algorithm should work were investigated. In the following we use another algorithm to build lexically coherent segments in which the number of segments is an explicit parameter. Like for the fixed length segments and the sliding window we can then vary the (average) segment length.

3. Experimental Setup

3.1 The MediaEval Dataset

We carry out experiments on a corpus with Creative Commons content collected from blip.tv. This data set was used for multiple tasks in the MediaEval benchmark (GenreRSR2011) [1].The collection contains 1974 episodes (247 development and 1727 test) comprising a total of ca. 350 hours of data. We have used the development set to test our algorithms. In the following we will report only on the test set. The spoken channel is a mixture of planned and spontaneous speech. Each episode is accompanied by automatic speech recognition (ASR) transcripts provided by CNRS-LIMSI and Vocapia Research [8] and also by metadata (descriptions, title and tags), added by the uploader. In the following we focus on segmentation of ASR transcripts and do not use the metadata to improve the retrieval.

The speech transcripts are divided into segments. Apparently segments boundaries are assumed at speaker turns and silences. Therefore we will refer to these segments as prosodic segments. The average lengths of these segments is 13,6 seconds. More details on the prosodic segments are given in section 3.3. The ASR transcripts also propose a further division of each prosodic segment into sentences. The average length of the proposed sentences is 6,1 seconds which corresponds to almost 18 words per sentence (on average 7,3 words after stop word removal). In general the proposed segment boundaries are very reasonable. Since we assume that words in a sentence have to be interpreted together, for all approaches we have aligned the boundaries of the retrieval units with the sentence boundaries. Given the short length of the sentences this means at most a few seconds deviation from originally computed boundaries.

The 2011 Rich Speech Retrieval task provided 80 queries (30 development and 50 test), each with both full and short forms. The set of queries was constructed by asking crowd source workers to find and mark passages for 5 different illocutionary acts and to provide a description and a query for each passage that could be used to find that passage [1] (Kofler 2011). Thus there is a user description of the target video segment (e.g., '*This is a clip from a George Carlin special in which he comments on why he does not vote*' and '*Andrew Magloughlen talks how Google can help advance government tech*') and a short query. The query is formulated to be directed at a general Web search engine (e.g., '*Voting Opinions*' and '*Google government projects*'). Because the queries are user generated, they can contain spelling or grammar errors. These, are not, however, a subject of investigation here. For our study of different segmentation methods we use only the short queries that are more like queries users normally give to search engines.

3.2 Approach

There are basically two approaches to passage retrieval (Roberts & Gaizauskas, 2004): Either all possible passages are ranked directly, or initially documents are retrieved and subsequently the most relevant passages within these documents are searched. We use the first approach here since there is no reason in this experiment to be very efficient and to limit the number of passages to be considered.

Before segmentation and ranking all words are stemmed and stop words are removed. Mark Hepple's [4] part-of-speech (POS) tagger is used to tag and lemmatize all words. We remove all closed class words (i.e., prepositions, articles, auxiliaries, particles, etc.). To compensate for POS tagging errors, we additionally remove stop words (standard Lucene search engine stop word lists). Word and sentence segmentation, POS-tagging and term selection are implemented as a UIMA (http://uima.apache.org) analysis pipeline. The ASR-transcripts of the test set (1727 videos) contain approximately 3, 07 million words. After filtering and stop word removal 1, 27 million words remain. This roughly gives a rate of a bit more than 1 content word per second. The average length of a video is 1782 recognized or 735 content words.

We carry out ranking using BM25. Since fragments may overlap, we calculate *idf* on the basis of the sentence, the basic organizational unit of the speech channel, as

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

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

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

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. In our experiments, we set $k = 2$ and $b = 0{,}75$, based on optimization of results on the development set.

The retrieval status value (RSV) of a document for query consisting of more than one word is defined as,

$$w\,(d, Q) = \sum_{t \in Q} w\,(d, t) \qquad (3)$$

We create an initial ranking by ordering all fragments by their RSV values. In order to generate our final results list, we remove all fragments with a starting time within a window of 60 seconds of a higher ranked fragment.

3.3 Evaluation Metric

Results are evaluated in terms of mean generalized average precision (mGAP). mGAP was designed for evaluation of passage retrieval and takes into account the effect that the found jump-in point often might be close to the desired point but not exactly the same [10].Thus it generalizes the calculation of the average precision 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. To be precise, the generalized average precision for each query is computed as

$$GAP = \frac{1}{N}\sum_{r=1}^{N} prec@r \left(1 - \frac{Penalty\ .\ Granularity}{Window}\right) \quad (4)$$

where $prec@r$ is the precision at rank r, N is the number of results returned, $Granularity$ is the step size used to measure the distance between the retrieved jump-in point and the relevant one, *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 and $Penalty$ 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. In the following we use a 60s tolerance window and 10s granularity.

4. Segmentation

If no additional information is available, like written plots, accompanying slides, etc., we can distinguish three basic ways to segment the video. The first possibility is to define arbitrary passages of fixed or arbitrary length not using any information from the content of the video. The second possibility is to use the video channel, especially by detecting shot boundaries. Finally, segmentation can be based on the audio channel and the recognized speech.

For some methods we can vary the length of the segments or the number of segments for a video. For the retrieval task as described above, long segments clearly have two disadvantages: longer segments have a higher risk of covering several subtopics and thus give a lower score on each of the included subtopics. In the second place, long segments run the risk that they include the relevant fragment but that the beginning of the segment is nevertheless too far away from the jump-in point that should be found. Short segments on the other hand might get high rankings based on just a view words. Furthermore, short segments make the retrieval process more costly. The ideal length should be learned on a test set. Here we are however not interested in determining the optimal length, but rather in studying the behavior of the retrieval under changing lengths.

In the following subsections we will present the compared segmentation strategies in more detail.

4.1 Using Shots as Segments

One of the most obvious ways to segment a video is by using the shot boundaries. For the Mediaeval 2011 data set the shot boundaries have been detected by [7]. There are 29 429 shots in the test set with an average length of 36,1 seconds. The length of the shots varies strongly with the type of video. There are 386 videos that consist of one single shot. An example of such a video is a book review, where the reviewer switches on his webcam, speaks into the camera, and when finished with his review switches the camera off. Here we see clearly differences by user generated video and professionally produced material. In other videos, like interviews or recorded discussions the succession of shots is much faster. For the longest videos, up to about 260 shots have been detected. More details on the distribution of shot lengths are given by the histogram in Figure 1.

4.2 Prosodic Segmentation

Another natural way to segment a video is to assume segment boundaries at speaker turns and at (longer) silences. The data set we have used (see section3.1) is distributed with transcripts from automatic speech recognition. These transcripts are divided into fragments based on prosodic information. Apparently fragment boundaries are assumed at each speaker turn in a video with several speakers and at longer silences. Exact

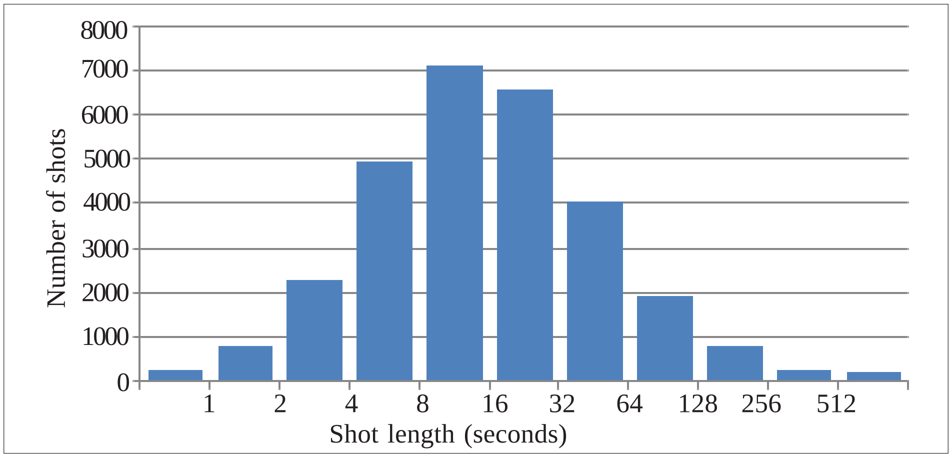

Figure 1. Histogram of shot lengths

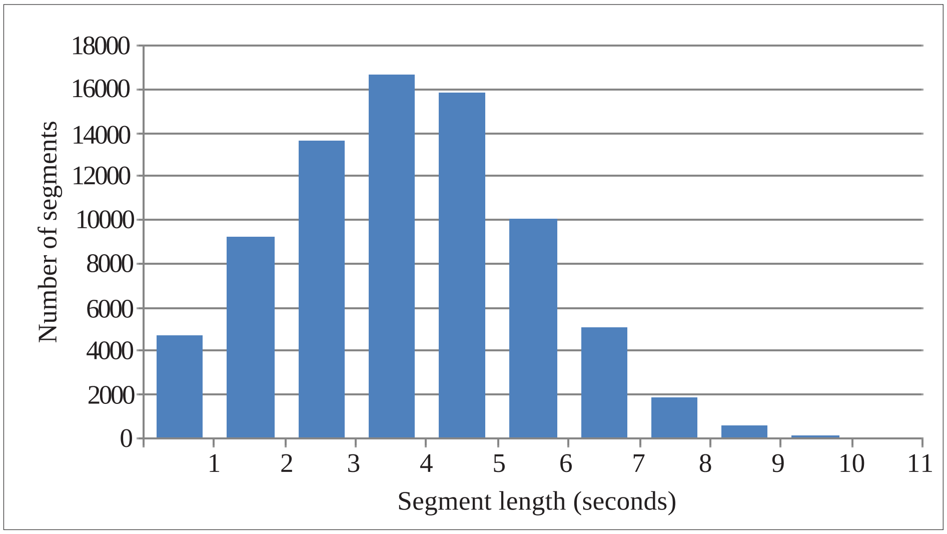

Figure 2. Histogram of lengths of prosodic segments

thresholds are not given. There are 77 878 fragments in the test set with an average length of 13,6 seconds. There are 396 videos that consist of one single prosodic segment while the longest videos have up to 1000 segments. The videos with only one prosodic segment in many cases are the same as the videos consisting of one shot. The overlap coefficient between the sets of videos consisting of one shot and those consisting of one prosodic segment is 0,56, where the overlap between two sets *A* and *B* is defined as

$$overlap\,(A, B) = \frac{|A \cap B|}{\min(|A|, |B)} \qquad (5)$$

The Jaccard coefficient is 0,38. The distribution of the lengths (see Figure 2) is very similar to the distribution of the shot lengths.

In general one could expect that there is a relatively large correspondence between the prosodic segments and the detected shots. A speaker turn could be accompanied by a new shot showing the new speaker and also silences could indicate topic boundaries that are marked as well by shot boundaries. Since both types of segments are normalized by the start of the first recognized sentence in that segment, the boundaries of both shots and segments can be compared easily. The overlap coefficient for these boundaries is 0,46, the Jaccard coefficient is 0,14. This is a relatively small correspondence, given the fact that the prosodic segments are very short and more than every third sentence start is a boundary of a prosodic segment. Thus we can conclude that, the shot boundaries and the boundaries of the prosodic fragments are quite independent.

4.3 Fixed Length Segmentation (Time Based)

The simplest way to split up a video is to divide it into segments of equal length. This kind of rigid segmentation would split up and hence destroy sentences and words that are needed subsequently for retrieval. Thus we try to find segments that are as close as possible to the target segment length but respect the sentence boundaries as proposed by the ASR. The length of these segments varies from 10 to 110 seconds.

4.4 Fixed Length Segmentation (Word based)

Segmentation into equal length segments alternatively can be based on the number of words in the segment instead of the duration of the segment. Thus segments roughly contain the same amount of information and have the advantage that the subsequent ranking algorithm does not have to deal with problems arising from length differences.

Before segmenting we lemmatize the speech transcripts and remove all stop words and non-content words (see section 3.2). We count the length of a passage in terms of content words rather than in terms of recognized words. Since a sentence reasonably is the smallest unit to be retrieved, we respect these boundaries. Thus, if we segment a transcript into parts of a certain length the actual length of each segment might be a few words longer or shorter. The segmentation algorithm always chooses the sequence of sentences with smallest absolute difference between the actual and the targeted length.

The test set comprises 199 140 sentences. On average each sentence has 15 recognized words and 6,4 content words.

4.5 Sliding Window Segmentation

The sliding window method uses fixed length segments based on words as well. The first segment is the same as in the fixed length approach. In order to find the next possible segment, the first sentence of the segment is removed, and one sentence at the end is added. If this new segment is longer than the target length, more sentences at the beginning are removed as long as the absolute difference with the target length decreases. If the segment is too short, in the same manner more sentences are added at the end. In case the target length is close to the average sentence length, the sliding window segmentation becomes almost the same as the fixed length segmentation.

4.6 Lexically Coherent Segmentation

The fixed length segments do not take into account the structure of the video. Ideally segmentation corresponds to rhetorical structure of the video or to subtopics in the video. Such segmentation then could be expected to give better results than fixed length segmentation, assuming that human annotators tend to choose the beginnings of these ‘*natural segments*’ as jump-in points.

A lot of research has been done into automatic segmentation of texts and speech transcripts. The basic idea is always to find regions that are lexically (and hence semantically) coherent. Lexically coherent passages can be understood as passages with a vocabulary that is distinct from adjacent regions or distinct from the overall vocabulary of the text. These regions usually tend to correspond very well to regions with a distinct subtopic. However, if we have spontaneous informal speech with smooth transitions of subtopics it is not that evident that always really natural segments are found, especially, if we try to find very short segments. Probably the most well-known method implementing this idea is Hearst’s text tiling algorithm [2].

The method for segmentation that we have used is the minimum cut model from [12]. This algorithm is based on sentence similarity. A cut has to be chosen, such that (length normalized) sum of the similarities between sentences to the right and to the left of the cut is minimal. If a text has to be split up in more than two segments the sum of the (normalized) cut values has to be minimal. In the original algorithm Malioutov and Barzilay do not use sentences but word sequences of fixed length. In our implementation we however stick to the sentences proposed by the speech recognizer. This raises the problem that a very short sentence between two long sentences is very likely to cause a break. To avoid this effect we use a relatively strong smoothing of word frequencies between adjacent sentences. To compute the sentence similarity we represent sentences as vectors of tf.idf values, smoothed as proposed by Malioutov and Barzilay, by setting

$$\bar{s}_l = \sum\nolimits_{j=i}^{i+k} e^{-\alpha(j-i)} s_j \tag{6}$$

where each s_i is a vector of tf.idf values, and α is the parameter that controls the degree of smoothing. In our experiments we have set $\alpha = 1$. For the computation of the $tf.idf$ we use the document frequency of a word in the whole test set.

Moreover we want to avoid that short sentences of one or two words with one common word, are tight much stronger together than long sentences with much more words in common. Thus we do not use cosine similarity but use the inner product of the $tf - idf$ vectors as a similarity measure.

In order to speed up the segmentation we do not consider segments shorter than two sentences and segments longer than half of the whole video. The algorithm finds an optimal segmentation into a given number of segments. Thus the number of segments (or equivalently the average segment length) is a parameter of the algorithm like in the fixed length and sliding window segmentation. However, we always split each video up in at least three segments. The segmentation algorithm cannot do anything useful if the targeted segment length becomes too small. Thus, we did not use this method for very short segments.

5. Results

We have run the same retrieval algorithm, for all different segmentation strategies and different (average) segment lengths.

The retrieval results for the 50 questions from the Media-Eval 2011 Rich Speech Retrieval task for all four segmentation methods and for varying segment lengths are given in Figure 3. The results are given for the actual average segment length, not for the targeted segment length. Segment lengths vary slightly because segments have to begin and end at sentence boundaries. For the longer fixed length segments the average length is strongly influenced by the last segment of each video that often is much smaller than the other segments. For all methods the segment length is expressed in seconds, independent of the fact whether the target length was defined as duration or as a number of words. The best results for each method and the length at which that result was achieved are given in Table 1. The coherent method reached its maximum value for several lengths in the range from 34,7 seconds to 75,1 seconds. Note however, that we do not know beforehand what the optimal segment length is.

Most pairwise differences are not significant. We used the Wilcoxon signed-rank test to compute significance. At the significance level of 0,05 the difference between the sliding window and the fixed length segments for length of 17 seconds (20 words) is significant as well as the difference between the coherent and the consecutive fixed length segments for the longest segment length (107s and 114s, resp.). Also the drop of mGAP between the values at e.g. 25s and 40s of the consecutive fixed length segments to the low values in the end is highly significant. The same holds for the sliding window. In contrast the pairwise differences between the results of the coherent segment strategy are not even significant at the level of 0,1.

For the fixed length and lexically coherent segmentation the number of segments that has to be ranked directly corresponds to the segment length. The range is for the fixed length segments is from 130 496 to 9 325. For the sliding window the number of segments ranges from 130 496 for the shortest segments to 99 865 for the longest segments.

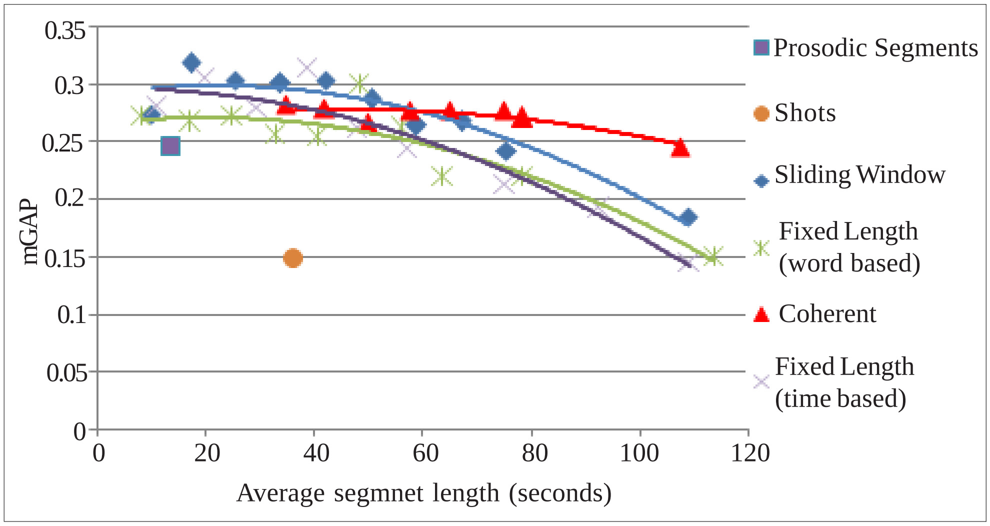

Figure 3. Results (observed values and trend lines) for the MediaEval2011 RSR task using 6 different segmentation methods with varying segment lengths

Method	Average segment length	mGAP
Shots	36,1	0,15
ASR Segments	13,6	0,25
Fixed length (time based)	29,3	0,31
Fixed length	48,4	0,30
Sliding Window	17,3	0,32
Coherent	34,7 (75,1)	0,28

Table 1. Results for the MediaEval2011 RSR task using 6 different segmentation methods with optimal segment lengths

6. Discussion

The first remarkable observation is, that the shot base segmentation performs significantly worse than the other methods. The results obtained with the prosodic fragments are reasonable, but not as good as those achieved with the other segmentation strategies. Apparently, neither shot boundaries nor pauses and speaker turns correspond very well with topic boundaries. This can be explained partly by the nature of semi-professional internet video e.g. when a single stationary camera is used. Also the genres present in the corpus might be a reason for the mediocre performance of the prosodic and visual segments. In a discussion or an interview there might be many shot boundaries and speaker turns when there is much interaction. But a high degree of interaction indicates that the participants are strongly involved in a discussion and probably are focusing on a specific topic. In instructional videos we see often an alternation of shots showing the speaker and the object he is talking about. Again these shot boundaries do not correspond to topic boundaries. Finally, it can be observed in many videos that a change of a topic is made very explicitly, i.e. the presenter is saying that a topic is finished and that he is switching to a new topic. At such a moment there is usually neither a shot boundary nor a longer break in the audio signal.

The best results are obtained using the sliding window. However, for most segment lengths the differences are not significant and the price in terms of number of segments that has to be considered for retrieval is high. The results of the fixed lengths segmentation is almost as good, but seems to be very sensitive to the exact value of the segment length. We do not find any remarkable differences between the segments based on a fixed duration and those based on a fixed number of content words. The problem for this type of segmentation in general is, that in a number of cases a passage containing the correct jump-in point is ranked very high, but that the beginning of the passage is too far away from this jump in point. The beginning of the passage is not determined by a change of vocabulary (like for the lexically coherent segments) nor by an optimal match (like for the sliding window), but by a rigid division into equal length segments. The segmentation method using the minimum cut model of Malioutov and Barzilay gives also similar results for segments up to a length of about 80 content words (i.e. about 70 seconds). This method does not have the disadvantages of the other methods: it seems much less dependent on the exact value of the segment length and it does not leave all the labor to the ranking algorithm. For the longer segments we see that the results obtained with the lexically coherent segments are also more stable and do not drop as fast as the fixed length and sliding window segments. Thus this segmentation strategy allows working with much less segments for retrieval.

Our results suggest that there is a clear advantage of using sophisticated segmentation methods for passage retrieval. This is quite surprising, since in most research on passage retrieval the advanced segmentation methods did not significantly perform better than other methods. As noted before most research on passage retrieval was done to improve document retrieval. In those studies the results are evaluated by the relevance of the retrieved documents only. A correct prediction of the jump-in point or of the relevant passage is not necessary. It seems to be exactly for the matching of the jump-in points that the semantic segmentation performs better than the fixed length strategy.

In the present study we tested only one method for non-trivial segmentation. Also the used method could be improved by including information about relations between words in the computation of sentence similarities: often passages are not coherent because the same words are used all over the passage, but because the words in the passage are related to each other. Thus, further improvement of retrieval results using this segmentation approach can be expected [14].

7. Conclusion

We have made some interesting first observations on the usefulness of different strategies for the segmentation of ASR transcripts for video passage retrieval. As more similar data sets will become available, more experiments should be done to substantiate these observations.

Having made this reservation, we found (1) that a segmentation based on shots or prosodic information is not adequate for the type of user generated video used here; that (2) the results of video passage retrieval perform best with methods using fixed length segments (consecutive or overlapping) or lexically coherent segments but that the results also strongly depend on the right choice of the segment length; and that (3)the segmentation method based on lexically coherent segments has some clear advantages: In the first place retrieval results based on this segmentation strategy are not as dependent on the choice of the correct segment length as those using a fixed length segmentation strategy. In the second place it does not produce as many candidate segments for retrieval as the sliding window approach and, finally, it also gives reasonable results for longer segments.

8. Acknowledgement

I would like to thank Martha Larson and Maria Eskevich for many helpful discussions.

References

[1] Eskevich, M., Jones, G. J., Larson, M., Wartena, C., Aly, R., Verschoor, T., Ordelman, R. (2012). Comparing Retrieval Effectiveness of Alternative Content Segmentation Methods for Internet Video Search. 10th International Workshop on Content-Based Multimedia Indexing (CBMI 2012). *IEEE*.

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

[3] Hearst, M., Plaunt, C. (1993). Subtopic structuring for full-length document access. *In*: Proceedings of the 16th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, (S. 59-68).

[4] Hepple, M. (2000). Independence and Commitment: Assumptions for Rapid Training and Execution of Rule-based POS Taggers. *In*: Proccedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL 2000).

[5] Huurnink, B., Snoek, C., de Rijke, M., Smeulders, A. (2010). Today’s and tomorrow’s retrieval practice in the audiovisual archive. *In*: Proceedings of the ACM International Conference on Image and Video Retrieval, (S. 18-25).

[6] Kaszkiel, M., Zobel, J. (2001). Effective Ranking with Arbitrary Passages. *Journal of the American Society of Information Science*, 52 (4) 344-364.

[7] Kelm, P., Schmiedeke, S., Sikora, T. (2009). Feature-based video key frame extraction for low quality video sequences. 10th Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS’09), (S. 25-28).

[8] Lamel, L., Gauvain, J. -L. (2008). Speech Processing for Audio Indexing. *Advances in Natural Language Processing*, (S. 4-15). Springer.

[9] Li, H., Dong, A. (2006). Hierarchical Segmentation of Presentation Videos through Visual and Text Analysis. *In*: Proceedings of the IEEE International Symposium on Signal Processing and Information Technology.

[10] Liu, B., Oard, D. W. (2006). One-sided measures for evaluating ranked retrieval effectiveness with spontaneous conversational speech. *In*: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR ’06), (S. 673-674).

[11] Liu, X., Croft, W. B. (2002). Passage retrieval based on language models. *In*: Proceedings of the 2002 ACM CIKM International Conference on Information and Knowledge Management (CIKM 2002), November 4-9, (S. 375-382).

[12] Malioutov, I., Barzilay, R. (2006). Minimum Cut Model for Spoken Lecture Segmentation. *In*: Procceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (ACL 2006). The Association for Computer Linguistics.

[13] Repp, S., Meinel, C. (2008). Segmentation of Lecture Videos based on Spontaneous Speech Recognition. *In*: Proceedings of the 10th IEEE International Symposium on Multimedia.

[14] Riedl, M., Biemann, C. (2012). Text Segmentation with Topic Models. *JLCL*, 27 (47-69), 13-24.

[15] Roberts, I., Gaizauskas, R. J. (2004). Evaluating Passage Retrieval Approaches for Question Answering. Advances in Information Retrieval - Proceedings of the 26th European Conference on IR Research (ECIR 2004).2997, S. 72-84. Springer.

[16] Salton, G., Allan, J., Buckley, C. (1993). Approaches to passage retrieval in full text information systems. *In*: Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval, (S. 49-58).

[17] Tiedemann, J. (2007). Comparing Document Segmentation Strategies for Passage Retrieval in Question Answering. *In*: Proceedings of the Conference on Recent Advances in Natural Language Processing (RANLP '07).

[18] Tiedemann, J., Mur, J. (2008). Simple is best: Experiments with different document segmentation strategies for passage retrieval. *In*: Proceedings of the 2nd workshop on Information Retrieval for Question Answeringa (IRQA '08), (S. 17-25).

[19] Van Houten, Y., Schuurman, J. G., Verhagen, P. (2004). Video Content Foraging. *In*: Proceedings of the Third International Conference on Image and Video Retrieval (CIVR 2004).3115, S. 15-23. Springer.

[20] Wartena, C. (2012). Comparing Segmentation Strategies for Efficient Video Passage Retrieval. *In*: Proceedings of the 10th Workshop on Content-Based Multimedia Indexing. Annecy.

[21] Wilkinson, R. (1994). Effective retrieval of structured documents. *In*: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval, (S. 311-317).

[22] Xu, J., Croft, W. B. (2000). Improving the effectiveness of information retrieval with local context analysis. *ACM Trans. Inf. Syst*., 18 (1) 79-112.

[23] Yamamoto, N., Ogata, J., Ariki, Y. (2003). Topic segmentation and retrieval system for lecture videos based on spontaneous speech recognition. *In*: Proceedings of the 8th European Conference on Speech Communication and Technology, EUROSPEECH 2003 - INTERSPEECH.