Segmentation Strategies for Passage Retrieval in Audio-Visual Documents

Petra Galuščáková*

Faculty of Mathematics and Physics
Charles University in Prague

Abstract

In this paper we deal with Information Retrieval from audio-visual recordings. Such recordings are often long and a user may want to know the exact starting point of each relevant passage. Therefore, we apply Passage Retrieval on the recordings. The recordings are automatically divided into smaller parts, on which we apply standard retrieval techniques. In this work, we study several techniques for segmentation of audio visual recordings and focus on strategies which create passages that are semantically coherent and more suitable for retrieval.

1 Introduction

Information Retrieval (IR) is an essential task which enables extracting particular document corresponding to a given query from data. In this work we are focused on IR from audio-visual recordings. This task is even more demanding than IR in textual documents. The semantic content of the recording needs to be at first mined using Automatic Speech Recognition (ASR) system from the audio track or using video content analysis from the visual track. The recordings have linear structure and, compared to texts, they are harder to skim, which is a problem especially for long recordings. Therefore, we apply Passage Retrieval on the audio-visual recordings – a process which divides long documents into smaller passages which serve as individual documents in further IR setup. This enables users to find the exact relevant segments in a

*galuscakova@ufal.mff.cuni.cz

collection of long audio-visual documents and should reduce time demanded to find requested information. Moreover, the text has usually structure (paragraphs, section, ...) defined by the author but no such structure is given in audio-visual recordings. To some extent, the structure of the recordings could be derived from audio and visual features (e.g. shots, length of the silence).

The remainder of the paper is organized as follows. In the following section we describe Passage Retrieval – a technique which makes use of splitting large documents into smaller parts for more efficient retrieval. In section 3, the process of splitting documents into smaller, semantically coherent, parts is described. Our setup and preliminary results of passage retrieval based on different segmentation strategies are described in section 4. In sections 5 and 6 we discuss our conclusions and plans for future work.

2 Passage Retrieval

Information Retrieval is a process of searching through a collection of data which finds the documents relevant to a users' query and returns full documents as a result. However, retrieval of full documents is sometimes insufficient. Passage Retrieval makes use of splitting texts into smaller units which are then used as documents in information retrieval process.

Passage Retrieval is used in numerous IR applications. It is especially utilized in Question Answering. Question Answering is a subtask of IR focused on retrieval an exact answer to question in natural language. To obtain an answer for the assigned question, a relevant passage must be identified first. Therefore, information retrieval is applied on the recordings and segments relevant to the question are marked, e.g. Roberts and Gaizauskas (2004); Melucci (1998); Tellex et al. (2003). Then, the answer is mined from the retrieved segment and presented to the user. If no relevant segment is retrieved, whole system is unable to return the right answer. Therefore, the IR quality is a bottleneck of question answering (Tiedemann and Mur, 2008). Question answering is very sensitive to the length of the relevant segment (Melucci, 1998). It needs to be long enough to contain all relevant information but it should not include other information. In some cases, whole retrieved passage could be considered as the answer. According to Lin et al. (2003) users even prefer whole passages to exact sentence to be retrieved because passages are embedded in context, which makes the answer more trustworthy and simplifies finding answers to related questions.

Another possible application of Passage Retrieval is automatic query expansion (Papka and Allen, 1997; Allan, 1995; Xu and Croft, 1996). Queries could be expanded by adding related words which occur in the same documents as the query terms. Query expansion could be improved by locating related words only in the close surrounding of the original query, i.e. if they occur in the same segment.

Passage Retrieval could also help "classical" IR setup in several ways. The first advantage of Passage Retrieval is that the position of the word occurrence could also be used in the retrieval (Mittendorf and Schäuble, 1994) – e.g. we could define that words occurring in the beginning of the document have bigger weight.

In the second case, Passage Retrieval could improve results of IR when the document is long and it contains a large range of topics. If the document contains a short relevant passage along many irrelevant passages, the relevant document is often identified as irrelevant. In Passage Retrieval, searched words must appear within a short distance. Several authors (Salton et al., 1993; Hearst and Plaunt, 1993; Kaszkiel and Zobel, 1997) show improvement when segmentation is employed comparing with case when full documents are retrieved; according to Kaszkiel and Zobel (2001), this improvement depends on passage type, collection, and query set.

Various techniques for evaluating the document relevance (Callan, 1994; Kaszkiel and Zobel, 1997; Tellex et al., 2003; Hearst and Plaunt, 1993; Buckley et al., 1994) are examined. The document could be scored according to its highest ranked passage, the scores of the relevant segments could sum or they could be combined with the score of whole document.

Length normalization is a next advantage of Passage Retrieval. In Passage Retrieval we could influence the length of the segments which we process. Kaszkiel and Zobel (1997) claim that Passage Retrieval could help length normalization of the documents from various sources which could be useful especially in the case when some measures (e.g. cosine) prefer shorter documents.

Fourth advantage of Passage Retrieval in IR setup is identification of the exact relevant passage in long documents when we want to save time needed to find relevant information. It is even more important for long audio-visual recordings in which skimming is time-demanding. The segmentation of audio-visual recordings is not widely studied, some experiments are performed by Eskevich et al. (2012c) and Wartena (2012). Wu and Yang (2008) work with

audio-visual recordings as well but they use it for question answering (for Chinese) in video documents, in which they utilize captions recognised by OCR.

Kaszkiel and Zobel (2001) divide segmentation into three groups – window-based (passages are created regularly as overlapping windows of fixed length, measured in term of words), structure-based (defined by the author of the document), and semantic-based. Semantic-based segmentation should correspond to the real topical structure of documents and we could find it using segmentation algorithm (e.g. TextTiling, C99). Some authors also use arbitrary segmentation in which segments could start on arbitrary word in the sentence and could last any long (Liu and Croft, 2002).

In text, the majority of authors show (Wartena, 2012; Kaszkiel and Zobel, 1997; Callan, 1994; Kaszkiel and Zobel, 2001; Tiedemann and Mur, 2008) that the segmentation using flowing window and creating overlapping segments of a regular length is the most successful approach to segmentation and its subsequent usage in IR. Authors also show that this approach is sensitive to the window length which needs to be tuned on training data. For instance, Callan (1994) supposes use of window with about 200–250 words, similarly Kaszkiel and Zobel (2001) achieves the best results for 150–300 words and Wartena (2012) achieves the best result with about 20 content words. Kaszkiel and Zobel (2001) also claim that the segmentation preference depends on the type of the query: for short queries and long documents structure-based segmentation achieve good results, whereas for the documents with the uniform length ignorance of document structure is preferred.

As it was said, window-based segmentation outperforms structure-based segmentation in most cases, which is slightly surprising. One possible explanation is that structure-based segmentation achieves worse results because lengths of the segments significantly varies (Kaszkiel and Zobel, 2001). According to Tiedemann and Mur (2008), the segmentation approach is not crucial, the length of the segment is more important. Callan (1994) also experiments with bounded-paragraph passages in which he merged too short paragraphs and split too long paragraphs. He gives another reason for insufficient structure-based segmentation results – topics are often formatted just for presentation purposes and they are not based on semantics. Problem arises with titles, captions, tables, and other similar units which are usually treated as ordinary paragraphs (Tiedemann, 2007).

However, Tiedemann and Mur (2008) shows that semantic-based segmentation (which utilizes coreference chains and TextTiling algorithm) outperforms segmentation which is based on paragraphs and sections defined by the author and could improve the results in question answering. In their experiments segmentation based on coreference chains even outperforms regular segmentation.

Window-based approach achieves the best results also in audio-visual retrieval in the experiments by Eskevich et al. (2012c) who compares segmentation techniques of the participants of the Rich Speech Retrieval Track in MediaEval Benchmarking in 2011. Wartena (2012) compares four segmentation approaches and evaluates the segmentation quality of audio-visual data. He examines non-overlapping fixed length segments, a sliding window, semantically coherent segments, and prosodic segments. He concludes that the quality of retrieval is sensitive to segment length. The best result is achieved using sliding window but the segmentation into topically coherent segments is more robust and less sensitive to the predefined average length of the segment, achieves better results for segments of higher lengths and thus enables the reduction of the number of segments.

Other experiments on Passage Retrieval include experiments by Tellex et al. (2003) who try several IR algorithms, Mittendorf and Schäuble (1994) who introduce IR method based on Hidden Markov Models in which Passage Retrieval is easily applicable and Liu and Croft (2002) who apply language modeling technologies.

As we can see, Passage Retrieval is helpful in many applications. However, segmentation is an extra step, comparing with IR. Also, the number of segments which need to be examined rises in Passage Retrieval and, therefore, it could be more computationally expensive than IR.

3 Semantic Segmentation

In this work we are aimed at retrieval from the audio-visual recordings which have no predefined document structure. We can use speech transcripts automatically extracted from audio track and segmentation designed for textual documents. Other information such as sound and video is available. In semantic segmentation we can influence the length and the nature of the segment. Thus, semantic segmentation could be effective method for splitting audio-visual documents and further application of information retrieval which is very sensitive to the segment properties. According to our best knowledge, only a

few experiments have been done in this area yet.

In this section we describe algorithms for semantic segmentation. In our former work (Galuščáková, 2012) we overview the methods used for semantic segmentation. Found passages must be semantically coherent and each passage should cover single topic. Semantic content is naturally organized hierarchically. According to Kaszkiel and Zobel (2001) "sentences should convey a single idea; paragraphs should be about one topic; and sections should be about one issue." Melucci (1998) even shows an improvement of Passage Retrieval if text is hierarchically organized. Whereas most algorithms for segmentation extract segments in linear fashion, several approaches output segmentation as the hierarchical structure. For instance, Song et al. (2011) create binary tree. They iteratively break the documents into segments at the weakest points using two similarity measures. But what is important is that segmentation should be consistent with the task in which it will be applied. If the segmentation is further used in Passage Retrieval, detected sections should correspond to the expected answers.

Segmentation approaches are diverse; Kauchak and Chen (2005) divide approaches into similarity-based, lexical-chain-based, and feature-based. We describe these approaches in following sections, divided according to the modality of processed data.

3.1 Text-based Segmentation

Many algorithms which utilize only textual information are based on measuring similarity between potential segments (usually determined by cosine distance). Segments should have high intra-similarity (they should be coherent) and low inter-similarity (they should differ from other segments) (Malioutov and Barzilay, 2006).

3.1.1 Similarity-based Algorithms

Probably most often used algorithms for semantic segmentation, TextTiling (Hearst, 1997) and C99 (Choi, 2000), are both similarity based; both calculating cosine distance between segments. In the C99 algorithm similarity matrix is created according to the similarity between each pair of sentences, regions with high similarities are then identified in the matrix and boundaries are set between regions with high intra-similarity. Reynar (1994) uses graphical algorithm called dotplotting (Church, 1993) and, similarly to C99, areas with

high density of words' repetition are identified as coherent segments. In TextTiling algorithm, distance between each two adjacent segments is calculated and points with the lowest values are considered as boundaries. TextTiling is also subsequently used in Passage Retrieval by Hearst and Plaunt (1993) and Eskevich et al. (2012c).

3.1.2 Lexical-chain-based Algorithms

Both similarity-based and lexical-chain-based algorithms make use of lexical cohesion in topical segments. Lexical-chain-based algorithms detect lexically related words – the amount of related words within one segment is typically higher then the amount between adjacent paragraphs. Kauchak and Chen (2005) define lexical chain as "a sequence of lexicographically related word occurrences". Segment boundary could be detected at the place where large number of lexical chains begin and end. Repetition of the lexical items could be detected easily and this approach could be improved by using synonyms and subordinates. Morris and Hirst (1988) determines lexically close words from Roget's thesaurus, Nguyen et al. (2011) further utilizes word collocations, Mohri et al. (2010)calculate coocurrence statistics and Kozima (1993) estimates similarities for pair of words and use them to find a sequence of lexical cohesiveness. Ponte and Croft (1997) propose a method for detection of small segments which share few common words. They use Local Content Analysis, which detects essential concept (bag of words, which describes topic) of two passages. Thus, passages does not have to contain common words but they need to have similar concept.

Lexical cohesion is also employed in Bayesian approach (Eisenstein and Barzilay, 2008; Jeong and Titov, 2010). Some authors use Latent Dirichlet Allocation (LDA), generative unsupervised model of the topic and use Gibbs sampling to estimate this model (Nguyen et al., 2011; Misra et al., 2009). Other approaches are based on Hidden Markov Models (Blei and Moreno, 2001; Mittendorf and Schäuble, 1994).

3.1.3 Feature-based Algorithms

Feature-based algorithms in text usually make use of cue phrases. Ballantine (2004) defines cue phrases as words and phrases which "serve primarily to indicate document structure or flow, rather than to impart semantic information about the current topic" (e.g. Good evening, well, so, . . .). Thus, they easily indicate the beginning or the end of a segment. Beeferman et al. (1999) study

the most influential lexical features – the most efficient feature is information whether a word appears up to five words in the past. Based on selected features with the highest gain, the probability that a topic ends is assigned to each sentence end and a decision about segment break is taken according to the assigned probability. Other used features include for instance the presence of pronouns and named entities.

3.2 Segmentation in Audio-Visual Recordings

Comparing with text-based segmentation algorithms, most algorithms focused on the audio-visual recordings are feature-based: they use supervised machine learning techniques which utilize wide range of textual, acoustic, and visual features.

Yun Hsueh and Moore (2007) integrate all kind of features and apply Maximum Entropy classifier on them. They also examine different combinations of the features and prove that application of multimodal features improves system with only lexical features by reducing overprediction. According to them, lexical features (i.e. cue words) are the most powerful ones but they need to be combined with audio and visual features. Conversational features (such as silence and speaker activity change, cue words and amount of overlapping speech) appear to be the most useful, followed by contextual features (dialogue act type and speaker role), prosodic features (e.g. fundamental frequency and energy level in audio track), and motion features (detected movements, frontal shots, hand movements). Maximum entropy classifier is also used by Hsu et al. (2004) who use features like anchor face, commercial detection, pitch jump, silence, speech segments defined by ASR system, speech rapidity and their combinations.

Tür et al. (2001) combine lexical cues with prosodic ones. Prosodic cues include energy patterns around segment boundaries, duration features (duration of pauses, duration of final vowels and final rhymes, and their normalized versions), and pitch features (fundamental frequency patterns around the boundary, pitch range). Decision tree and Hidden Markov Models are applied on these features. Similar features are also used by Dielmann and Renals (2005) but they apply them in dynamic Bayesian Network to solve segmentation of recordings of meetings.

Pye et al. (1998) combine audio segmentation algorithm based on the change in acoustic characteristics and on Kullback-Leibler distance between frames.

Their shot segmentation is based on color histogram of video. The audio breaks are essential in their work, visual breaks are used to support them. Hauptmann and Witbrock (1998) are especially interested in visual features, they use them also for commercials detection. Among scene cuts they use also black frames (which often precede commercials), frame similarity (color histogram similarity and face similarity), and motion information. They also integrate information from captions. Other applicable features count hand gestures, corresponding slides, and notes from meetings, if they are available. Malioutov et al. (2007) introduce the approach which does not require the transcript, they just use audio track and analyze the occurrence of acoustic patterns.

Textual features in audio-visual segmentation need to be acquired using ASR system. However, the quality of the transcripts varies and arises the question how does the quality of the transcripts influence the IR. Yun Hsueh and Moore (2007) show that despite word recognition error of 39% word error rate, none of their system performs significantly worse on ASR transcripts than on reference transcripts. They also observe one possible explanation: the same word is misrecognized by the same way in different parts of corpus and thus, the cohesion is not influenced. Utilization of multimodal features could also reduce the influence of the transcript quality. The quality could also be improved by using lattices instead of single one-best hypothesis of ASR system (Mohri et al., 2010).

3.3 Evaluation of Segmentation Quality

Segmentation is evaluated using standard *Precision* and *Recall* measures. We count the number of cases from all marked boundaries in which the segment boundary really occurs (*Precision*) and the number of cases from all possible boundaries (e.g. after each word or sentence) in which the boundary is marked (*Recall*). But the number of possible boundaries could be huge comparing to the number of real segment boundaries, which could cause the unsuitable high recall values.

Therefore, two measures are especially proposed to estimate the quality of segmentation system: P_k (Beeferman et al., 1999) and *WindowDiff* (Pevzner and Hearst, 2002). P_k reports the probability that two sentences randomly selected from the text are correctly determined to belong to the same or different segments. However, Pevzner and Hearst (2002) found that the measure penalizes “false negatives more heavily than false positives” and “over-penalizes

near-misses". Therefore, they modified P_k measure and proposed *WindowDiff* measure. In their proposal, a fixed-length window is slid through the document and number of times in which number of marked segment boundaries differs from real segment boundaries inside of the window is calculated.

In our experiments we use extrinsic evaluation. We do not evaluate segmentation directly but we evaluate it in use – we evaluate applied IR. Methods, used for evaluation of IR are described in section 4.1.3.

4 Experiments

In this section, we describe our experiments with Passage Retrieval in which we employ several types of segmentation and examine various IR techniques to tune our system. All our experiments are performed within Search and Hyperlinking Task in MediaEval Benchmarking[1] .

4.1 Test Collection and Evaluation Methods

The test collection which we use in this work was published in MediaEval Benchmarking. Similarly, the evaluation methods which we apply were used for evaluation of Search and Hyperlinking Task. In this section data collection used in our experiments and evaluation methods applied on our results are described.

4.1.1 MediaEval Benchmarking

MediaEval is benchmarking aimed at development, comparison, and improvement of strategies for processing and retrieving multimedia content. One of the organized tasks focused on Search and Hyperlinking.

The main aim of the Search and Hyperlinking Task is to find solution of the following problem: users want to find the passage relevant to their interest in a large set of audiovisual recordings. Subsequently, users want to find more passages similar to the retrieved ones. Thus, we would like to help users to find the relevant information quickly and then easily navigate through related passages.

The task consists of two subtasks: search subtask in which we retrieve the exact passage relevant to the user submitted query and hyperlinking in which we retrieve more passages similar to the retrieved one. The search subtask

[1] `http://www.multimediaeval.org/`

coincides with our problem of finding relevant segment in collection of audio-visual data, therefore, we participated in it.

4.1.2 Test Collection

The video collection used in the task was created from semi-professional videos published on Blip.tv[2] under the Creative Common license. Data significantly vary in format (e.g. local television news, interviews, culinary shows, personal blogs), length, and quality. The collection was divided into development and test sets (Table 1). In the following experiments, the results are reported on the test set. Details about the data are in the task description (Eskevich et al., 2012a).

	Dev Data	Test Data
Number of Documents	5288	9550
Hours of Video	1143.2	2144.6
LIMSI Sentences	369444	456732
LIUM Speech Segments	349502	705441

Table 1: Statistics of the test collection.

The queries were created using crowdsourcing. Participants were asked to find remarkable passages in the recordings and comment them shortly (Larson et al., 2011). This process differs from the usual query input in which the user first specifies the query and then judges the retrieved passages. The reversed procedure could cause higher overlap of the queries and relevant passages because the users tend to use the vocabulary from the recording. On the other hand, the queries may be more diverse. Totally, 60 queries were collected; 30 were used as the development set and the rest was used for testing. The queries consist of "Title", which shortly describes passage, "Short Title" which describes the passage in a more "search engine" style, information whether the segment contains a face, main color of the segment, and the main visual concept (e.g. Rocky Mountains, Volcanoe, Chair, Piano), if there is any. An example of a question and a relevant segment is presented in Table 2.

The recordings are published with two transcripts created by the LIMSI/ Vocapia (Lamel and Gauvain, 2008) system and the LIUM system (Rousseau et al., 2011), metadata, shot boundaries (Kelm et al., 2009), face clustering,

[2] http://blip.tv/

Title	Profit Partner programe talks about growing business faster.
Short Title	the profit partner growing business faster mortgages
Face	Yes
Colours	Dark
Video Content	Chair, Woman
Relevant Segment	Welcome to the Profit Partner where we help you grow six figure businesses in twelve months or less. My name is Cheree Warrick and I am the Profit Partner and I am so very honoured today to the interviewing Sarah Pichardo of George Mason Mortgage.

Table 2: Query example.

and visual concepts. The LIUM transcripts consist of one-best hypothesis, word-lattices, and confusion networks and the LIMSI transcripts include word variations with their confidence score. LIMSI first detects the language of the recording before processing it, so the transcripts could be in several languages, whereas LIUM transcribes only into English. In the LIMSI transcripts, the segmentation into sentences is available and the transcripts are divided into speech segments; each speech segment corresponds to continuous utterance of one speaker. Published data are further are described in task description (Eskevich et al., 2012a).

4.1.3 Evaluation Methods

Three measures were employed for evaluating the Search subtask: Mean Reciprocal Rank (mRR), Mean Generalized Average Precision (mGAP), and Mean Average Segment Precision (MASP). Each measure was applied with three window lengths: 60, 30, and 10 seconds; the window length is a parameter of each measure. In the following experiments we use 60-seconds-length only.

Reciprocal Rank is calculated as a reciprocal value of the rank of the first correctly retrieved document; in our case, document correctly retrieved inside of a window with the given length. mRR (Voorhees, 1999) is then calculated as the average of Reciprocal Ranks over the set of queries.
The GAP (Liu and Oard, 2006) measure also employs the exact jump-in point, which represents the start of the relevant segment. In our experiments, it is calculated as follows:

$$GAP = \frac{1}{rank} Penalty(distance) \quad (1)$$

where *rank* is the rank of the fist correctly retrieved document and *Penalty* assesses the quality of the jump-in point. The *Penalty* value is estimated according to the *Penalty Function*, based on the distance between the starting point of the relevant segment and the starting point of the retrieved segment. The shape of the *Penalty Function* is triangular and it depends on a given window width. In our previous work, we have proposed the adapted *Penalty Function* (Galuščáková et al., 2012), which should better correspond with satisfaction of users when they are searching for particular information. Similarly to mRR, mGAP is calculated as an average of GAP values over the set of the queries.
The MASP (Eskevich et al., 2012d) measure exploits the precision of whole retrieved segment; i.e. both starting and ending points of the relevant segment are taken account. MASP is calculated as the average of Average Segment Precision values over the set of queries. The Average Segment Precision is in our case calculated as the length of the first relevant retrieved segment (first document correctly retrieved inside of a given window) over the length of the relevant segment. mGAP and MASP are explained in Figure 1.

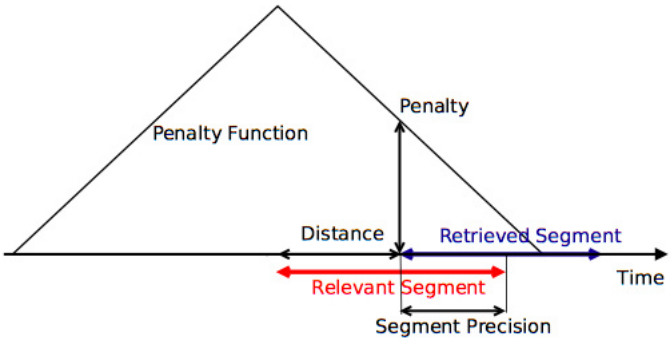

Figure 1: Explanation of the mGAP and MASP measures.

4.2 System Description

In all experiments, we employ the Terrier IR system[3] on the segmented transcripts of the recordings from the MediaEval Benchmarking. In this section we

[3] http://terrier.org/

describe our Baseline Run and the experiments performed with three state-of-the-art search models.

4.2.1 Baseline Run

In our Baseline Run, we apply Hiemstra Language Model (Hiemstra, 2001) on regular 90-seconds-long segments with 30-seconds overlap. We do not use any tuninig (Section 4.2.3), no stemming and stopwords (Section 4.4.1), no metadata (Section 4.4.4), any type of query expansion (Section 4.4.5) and results filtering (Section 4.4.2). Only a "Title" field from the query is employed. The scores for Baseline Run for both transcripts are displayed in Table 3.

	LIMSI			LIUM		
	MRR	mGAP	MASP	MRR	mGAP	MASP
Baseline	0.195	0.131	0.049	0.242	0.155	0.062

Table 3: Results of Baseline Run on LIMSI and LIUM transcripts. Hiemstra LM is applied on regular 90-seconds-long segments with 30-seconds overlap, no LM tuning, metadata, stemming, stopwords, any type of query expansion and results filtering are applied. Only a "Title" field from the query is employed

4.2.2 IR Model

We examine three IR models: TF IDF (Manning et al., 2008, p. 118), Hiemstra Language Model (Hiemstra, 2001), and BM25 (Manning et al., 2008, p. 232), see Table 4. Each model is applied with implicit parameters, no tuning of the models is employed.

	LIMSI			LIUM		
	MRR	mGAP	MASP	MRR	mGAP	MASP
Hiemstra LM	**0.47**	**0.29**	**0.123**	**0.449**	**0.25**	**0.102**
TF IDF	0.428	0.256	0.103	0.418	0.239	0.087
BM 25	0.423	0.251	0.102	0.429	0.238	0.091

Table 4: Search models comparison. The results of all systems are without parameter tuning, for 90-seconds-long window with 30-seconds overlap, with stemming, stopwords, metadata and "removal of overlapping" filtering and both "Title" and "Short Title" fields employed. The best results are in bold.

Language model is further tuned and the results are described in Section 4.2.3. For both the LIMSI and LIUM transcripts, Langauge Model achieves highest

score even if no parameter tuning is performed. In case of the LIUM transcript, BM25 slightly outperforms TF IDF model. In case of LIMSI transcripts, TF IDF is slightly better than BM25 but the difference is minor.

4.2.3 Hiemstra Model Tuning

The results of Hiemstra Language modeling are strongly dependent on the parameter used in this method. According to Hiemstra (2001), the parameter expresses the importance of a query term in a document.

In the experiments, we find a connection between segment length and language model parameter. This behaviour is apparent in Figures 2, 3, and 4 (all experiments are performed on the LIMSI transcript). The presented values are for filtered results, therefore, we also examine the behaviour without any filtering but the trend is the same.

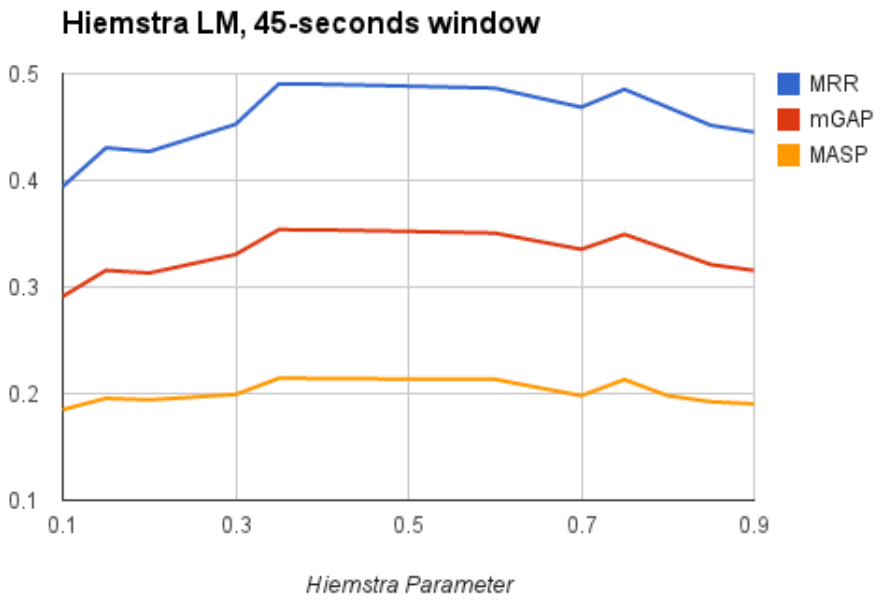

Figure 2: Behaviour of the parameter of Hiemstra Language Model on LIMSI transcripts for 45-seconds-long window with 15-seconds overlap, with stemming, stopwords, metadata and "removal of overlapping" filtering and both "Title" and "Short Title" fields employed.

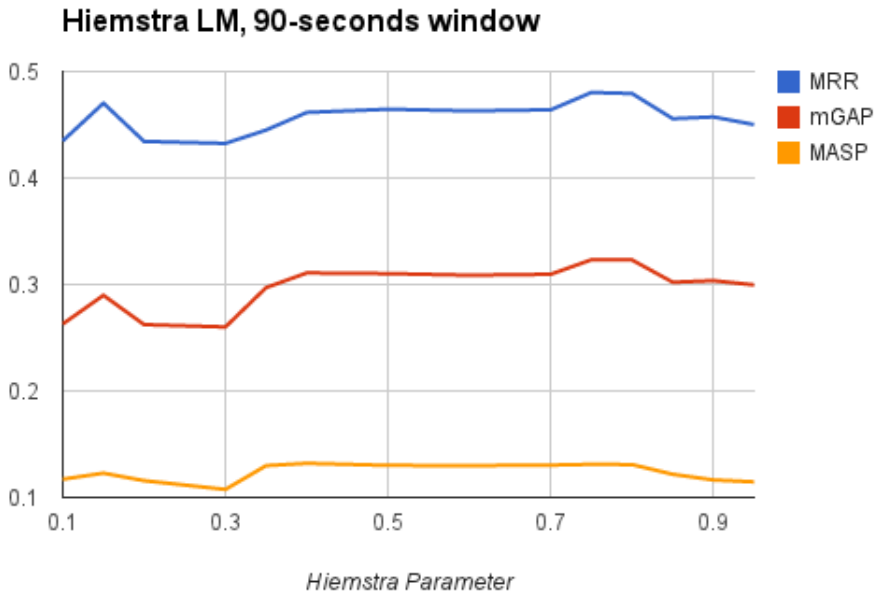

Figure 3: Behaviour of the parameter of Hiemstra Language Model for 90-seconds-long window with 30-seconds overlap, with stemming, stopwords, metadata and "removal of overlapping" filtering and both "Title" and "Short Title" fields employed.

Presented functions of the measures differ in maximum points. For the window-length of 45 seconds, the highest value for all measures is achieved at 0.35. The values for window of length of 90 seconds achieve the maximum for MRR and mGAP score at 0.75 and for the MASP score at 0.4. For window-length of 120 seconds, the maximal MRR score is achieved at 0.8 and the maximal mGAP and MASP score are achieved at 0.2. In all cases, there is a local maximum of the function at 0.15, then the function breaks around the point 0.35 and next local optimum occurs around point 0.75. Hiemstra (2001) also experimentally determined parameter 0.15 to perform well in general.

4.3 Segmentation

In this section we explore how does segmentation influence Passage Retrieval. Specifically, we study several parameters of window-based segmentation, described in Section 2 and two types of semantic segmentation: TextTiling and

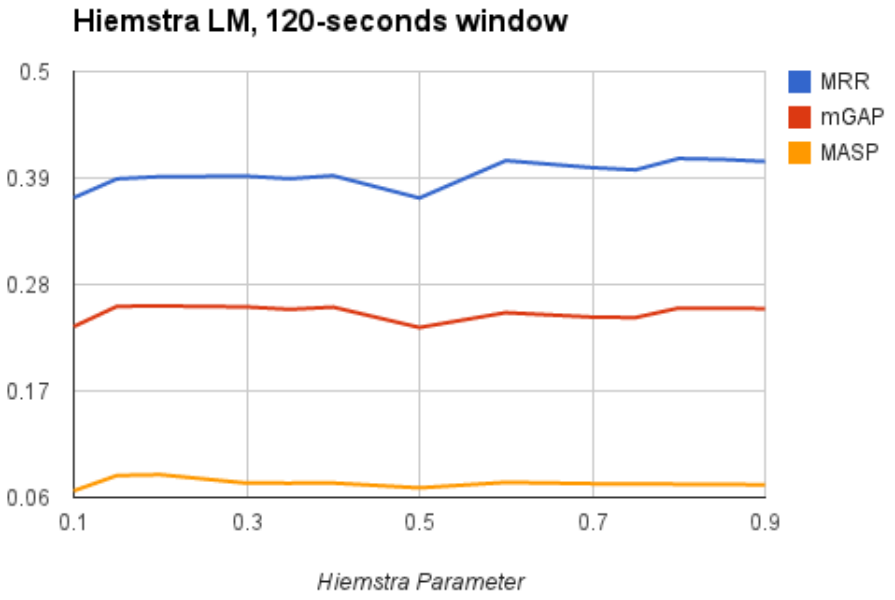

Figure 4: Behaviour of the parameter of Hiemstra Language Model for 120-seconds long window with 30-seconds overlap with stemming, stopwords, metadata and "removal of overlapping" filtering and both "Title" and "Short Title" fields employed.

feature-based segmentation employing decision trees.

4.3.1 Window-based Segmentation

Sliding windows of different sizes (time lengths) and overlaps are created in this approach. Comparing to former approaches which count number of words in segment (Wartena, 2012; Kaszkiel and Zobel, 2001; Callan, 1994), our window-based strategy utilizes time in the recordings. As simply as this approach is, it achieves the highest score in our experiments.

In our former work (Galuščáková and Pecina, 2012), we examined the connection between segment length and used measures. We used window of 45, 60, 90, and 120 seconds, with 30-seconds overlap (15 seconds for 45-seconds long window).

We also examine the effect of overlapping – we create a new window with a

given length (60- and 90-seconds long) with 10-, 15-, and 30-seconds overlap. The results of overlapping for Hiemstra Language Model is displayed in Figure 5, for TF IDF in Figure 6, and for BM 25 in Figure 7.

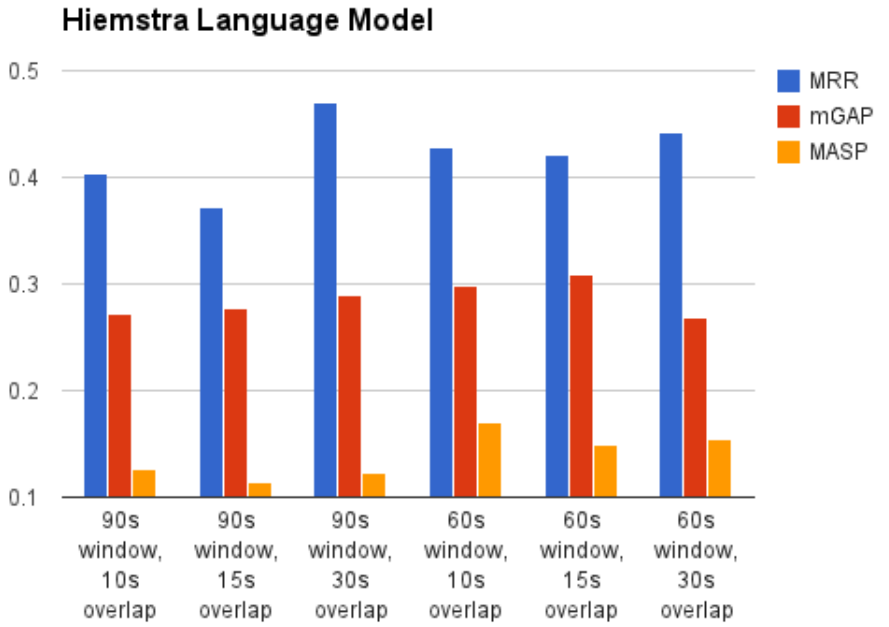

Figure 5: The effect of an overlapping on the LIMSI transcripts for Hiemstra Language Model for 60- and 90- seconds-long windows with various overlaps, with stemming, stopwords, metadata and "removal of overlapping" filtering and both "Title" and "Short Title" fields employed.

Surprisingly, the best results are achieved for 30-seconds overlap in all cases. For Hiemstra LM, window of 90 seconds outperforms window of 60 seconds but for TF IDF and BM 25, window of 60 seconds long window outperforms window of 90 seconds.

4.3.2 Semantic Segmentation

We explore two segmentation approaches based on the semantic content – TextTiling algorithm (Hearst, 1997) and feature-based segmentation. TextTiling algorithm is applied with settings set to correspond to regular segmentation with 90-seconds-long windows (one segment consists of 9 sentences and one sentence contains 27 words, in average).

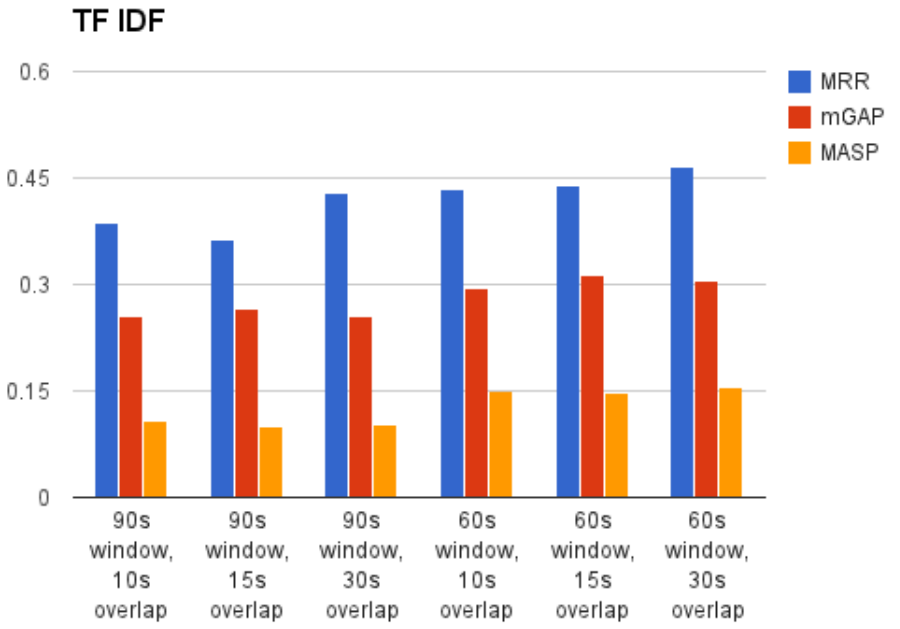

Figure 6: The effect of an overlapping on the LIMSI transcripts for TF IDF search engine for 60- and 90- seconds-long windows with various overlaps, with stemming, stopwords, metadata and "removal of overlapping" filtering and both "Title" and "Short Title" fields employed.

Last segmentation approach is based on feature-based semantic segmentation. We utilize decision trees (Breiman et al., 1984) in this approach: for each word, we decide whether the segment ends after this word or not. First, we manually mark the segments ends of several documents (about 4 hours of recordings were processed). Then, we use following features for training: shot segments, output of TextTiling algorithm, cue words (well, thanks, so, I, now), speech segments, sentence breaks, and the length of the silence after the previous word. As it could be hard to find the exact shot boundary, we employ also certain tolerance by setting the feature shot to "*true*" on small surrounding (one word before and after) of the breaking point. Then, we use these features for training decision trees, which are finally applied to segment the test data.

The results of semantic segmentation are shown in Table 5. The semantic-based segmentation underperforms window-based segmentation. However we believe that tuning of feature-based approach could further improve the results,

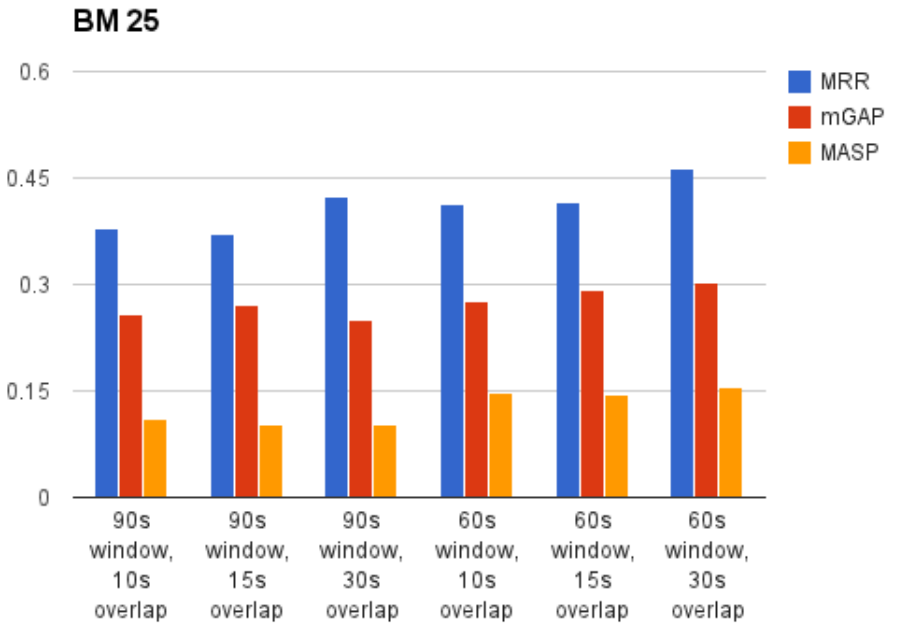

Figure 7: The effect of an overlapping on the LIMSI transcripts for BM25 search engine for 60- and 90- seconds-long windows with various overlaps, with stemming, stopwords, metadata and "removal of overlapping" filtering and both "Title" and "Short Title" fields employed.

especially in case of the mGAP and the MASP scores which are more sensitive to marking exact starting end ending points.

	MRR	mGAP	MASP
Baseline Run	0.195	0.131	0.049
Best Run	**0.489**	**0.352**	**0.214**
TextTiling	0.278	0.206	0.161
Feature-based	0.265	0.174	0.166

Table 5: Application of semantic segmentation on LIMSI transcripts. Hiemstra LM is applied on both TextTiling and feature-based approach, with stemming, stopwords, and both "Title" and "Short Title" fields employed. The best results are in bold.

4.4 System Tuning

We use several methods to improve the IR system performance: we apply stopwords and stemming, filter the retrieved results, use only English documents, utilize metadata, and explore automatic query expansion.

4.4.1 Stopwords, Stemming and Full Query Application

Stopwords and stemming are standard preprocessing procedures in IR, see Table 6. We use procedures available in Terrier: implicit stopwords list and Porter Stemmer. Application of stopwords and stemming increases the results; in some cases, almost by a factor of two.

In the Baseline Run, only "Title" field is used in the retrieval. Application of query "Short Title" field also improves the Baseline results; in case of the LIUM transcripts and for the MRR measure in case of the LIMSI transcripts, the usage of the "Short Title" field helps the Baseline even more than stopwords and stemming application.

	LIMSI			LIUM		
	MRR	mGAP	MASP	MRR	mGAP	MASP
Baseline	0.195	0.131	0.049	0.242	0.155	0.062
Stopwords + Stem.	0.291	**0.211**	**0.079**	0.313	0.164	0.064
Title + Short Title	**0.310**	0.188	0.077	**0.321**	**0.171**	**0.079**

Table 6: Employing stopwords and stemming on the LIMSI and LIUM transcripts. In both cases, Hiemstra LM is applied on 90-seconds-long segments with 30-seconds overlap. The best results are highlighted.

4.4.2 Filtering of Overlapping Results

As the window-based segmentation produces overlapping segments, overlapping passages are also contained in the retrieved results. We suppose that this overlap could cause decrease the MRR and mGAP scores, because the relevant segment could be postponed by many irrelevant overlapping segments. Therefore, we use several strategies to remove overlapping: we keep only the higher ranked segment from each document (one best), we filter out the segments which partially overlap with higher ranked segments (removal of overlapping) and we filter out all segments which lie in the surrounding of the higher ranked segments (window filtering), see Table 7.

The hypothesis that the overlapping segments in the retrieved results could

	MRR	mGAP	MASP
No Filtering	0.474	0.341	0.208
Removal of Overlapping	**0.489**	**0.352**	**0.214**
Window Filtering	0.486	0.35	0.212
One Best	0.469	0.335	0.207

Table 7: The effect of filtering the results on the LIMSI transcripts. In all cases, Hiemstra LM without parameter tuning is applied on 90-seconds-long window with 30-seconds overlap, with stemming, stopwords, metadata and both "Title" and "Short Title" fields employed. The best results are in bold.

decrease the overal score is confirmed. We discover that the most efficient strategy for results filtering is to remove all segments which are partially overlapping with higher ranked segment.

4.4.3 English-Only Files

MediaEval data contain documents in several languages including English, Spanish, Dutch, and French (Eskevich et al., 2012a) but all the queries and assessed relevant segments are in English only. Therefore, the list of documents which contain mainly English data was published by the organizers after the final submission. The LIUM system transcribes all documents into English and the results for all documents and English-only documents are almost identical for this transcript. However, the LIMSI system at first detects the language of the document and transcribes it into the most probable language. If the decision is not certain, document is transcribed into both languages in ask, probability of the document in both languages is estimated and the more probable transcript is selected. The comparison of all employed data and English-only files for the LIMSI transcripts for various runs, including Baseline, the Best Run, and runs from Galuščáková and Pecina (2012), is displayed in Figure 8.

Because the queries and relevant segments are in English only, filtering of the files in other languages is naturally expected to increase all scores. In most of the cases the difference between the results is minor but, surprisingly, for Run 1 and Run 3 the results for all files are slightly higher than for English-only files. This behaviour needs to be further examined.

4.4.4 Metadata Utilization

The recordings in the data collection are accompanied with further information such as title, description, and tags, provided by authors, and comments by

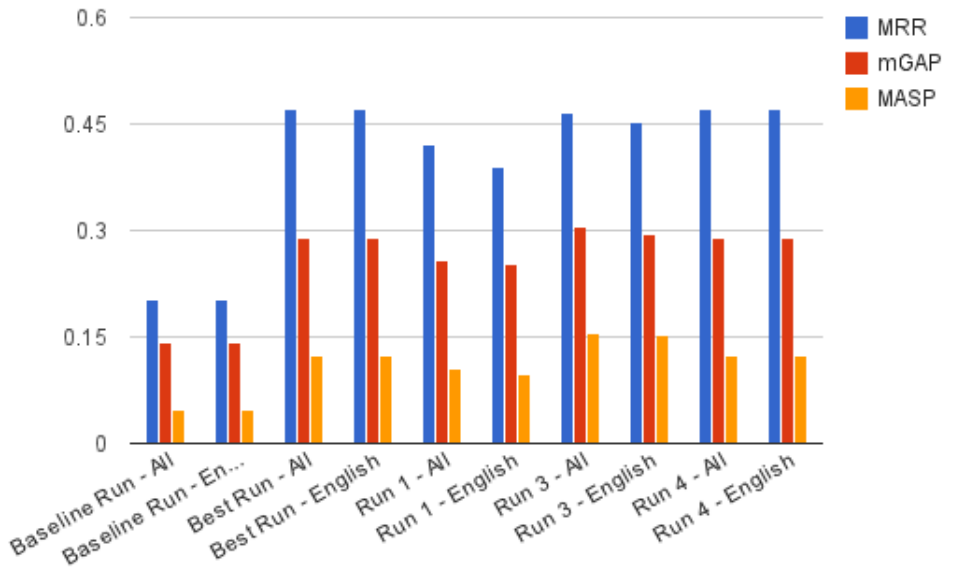

Figure 8: Results for all available documents and English-only files for various runs. Different runs use various search models, segmentation and various tuning was applied on them (Galuščáková and Pecina, 2012)

spectators.

First, for each segment we find metada belonging to the parental document of the segment and concatenate the segment and found metadata, see Figure 9. Thus, segments from the same file have the same metada. This approach improves results in all measures. Concatenation of description, tags, and comments very slightly outperforms concatenation of description, tags, comments, and filename for the MRR measure, for the mGAP, and the MASP measures the latter approach slightly wins.

4.4.5 Automatic Query Expansion

Query expansion is a technique which enables extension of a query by new words and thus overcomes the problem of small lexical overlap of the query and a relevant passage. We examine two approaches to expand the query – we

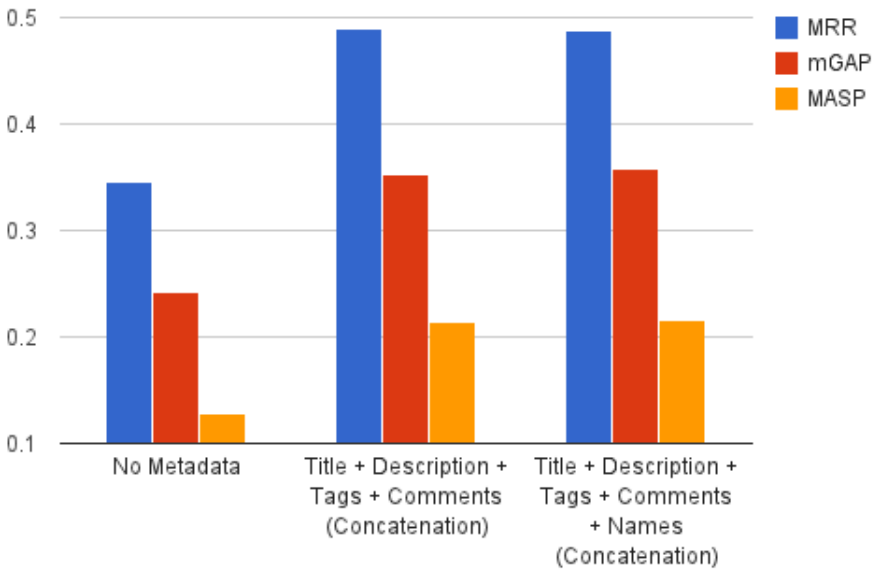

Figure 9: Employement of metadata on the LIMSI transcripts. The experiments are performed on the Best Run.

use a psuedo-relevance feedback and we expand the queries using WordNet[4].

In our experiments, a pseudo-relevance feedback (Manning et al., 2008, p. 187) increases MRR score and decreases mGAP and MASP scores of the Baseline, see Figure 10. If we employ pseudo-relevance feedback in case when we use metadata, pruning and also a "Short Title" of the queries, the scores drop. In this case, the query is already expanded by the "Short Title" and the relevant passage is expanded by metadata. Further expansion of the query carries irrelevant information and thus decreases the scores.

In the second case, we use WordNet to automatically expand the queries. For each query term we find a set of coordinated terms, derived words, hypernyms, hyponyms, and synonyms. The correct sense is not disambiguated and in each case, all possible senses are used. All the strategies decrease the score but derivation of nouns is the most promising strategy.

[4]`http://wordnet.princeton.edu/`

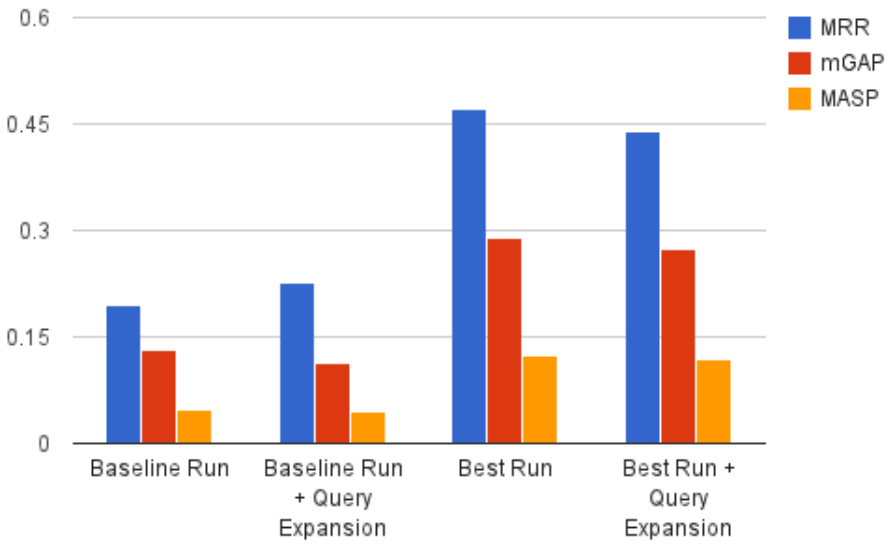

Figure 10: Employement of a pseudo-relevance feedback on the LIMSI transcripts.

4.4.6 Transcript Types

For the LIMSI transcripts we use all possible variations of each word provided in the transcripts and for LIUM we use the one-best possibility. The average word error rate of relevant passages in the LIMSI transcript is 0.317 and 0.404 for the LIUM transcript. As LIMSI offers more word varieties, the transcripts are more robust than LIUM one best transcript. The transcripts also differ in the vocabulary. The vocabulary of the LIMSI transcripts is bigger but it is mainly caused by transcribing into several languages. If we use English files only, the size of the vocabulary drops by more than a half, see 8.

	LIMSI	LIUM
Words Total	13.9 mil	10.3 mil
Words Unique	186 k	87 k
English Words Total	12.6 mil	9.1
English Words Unique	93 k	81 k

Table 8: Statistics of test data.

In our experiments, the LIUM transcripts outperform the LIMSI transcripts in the Baseline Run but in the tuned runs (stopwords, stemming, metadata, filtering, and short title employed), LIMSI achieves higher score, see Table 9.

	LIMSI			LIUM		
	MRR	mGAP	MASP	MRR	mGAP	MASP
Baseline Run	0.195	0.131	0.049	**0.242**	**0.155**	**0.062**
Tuned Run	**0.47**	**0.29**	**0.123**	0.449	0.255	0.102

Table 9: Comparison of LIMSI and LIUM for the Baseline and a tuned run. The results of the Hiemstra LM are without parameter tuning, for 90-seconds-long window with 30-seconds overlap, with stemming, stopwords, metadata and "removal of overlapping" filtering and both "Title" and "Short Title" fields employed. The best results are in bold.

4.5 Final Results

The best result was achieved on the LIMSI transcripts for Hiemstra LM with parameter 0.35, for 45-seconds-long window with 15-seconds overlap, with stemming, stopwords, metadata and "removal of overlapping" filtering and both "Title" and "Short Title" fields applied. The results for the best run are in Table 10

	MRR	mGAP	MASP
Best Run	0.489	0.352	0.214

Table 10: Results of the Best Run; achieved on the LIMSI transcripts for Hiemstra LM with parameter 0.35, for 45-seconds-long window with 15-seconds overlap, with stemming, stopwords, metadata and "removal of overlapping" filtering and both "Title" and "Short Title" fields applied.

In Figure 11 and Figure 12, the results for each query for both transcripts are drawn. The queries with the highest MRR score are displayed in Table 11 and with the lowest MRR score in Table 12.

Not surprisingly, LIMSI outperforms LIUM for most of the queries. However, for queries 10 and 19, LIUM achieves better score for all measures and for query 17 LIUM outperforms LIMSI in the MASP measure. Query 21 is also remarkable: it achieves the maximal MRR and mGAP score, but MASP is equal to zero. Generally, queries with high scores often contain specific words and proper names, which help to identify the segment of interest. Queries with low scores are very descriptive, especially query 20.

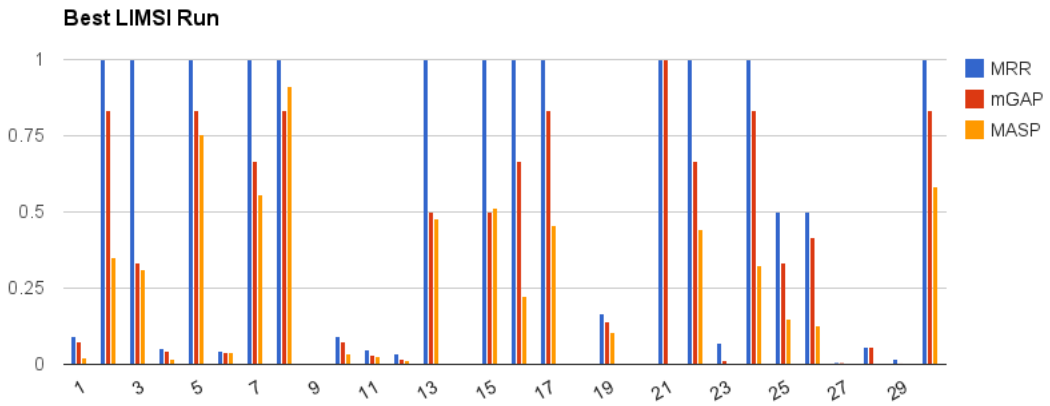

Figure 11: Results per query for the Best Run for LIMSI transcript.

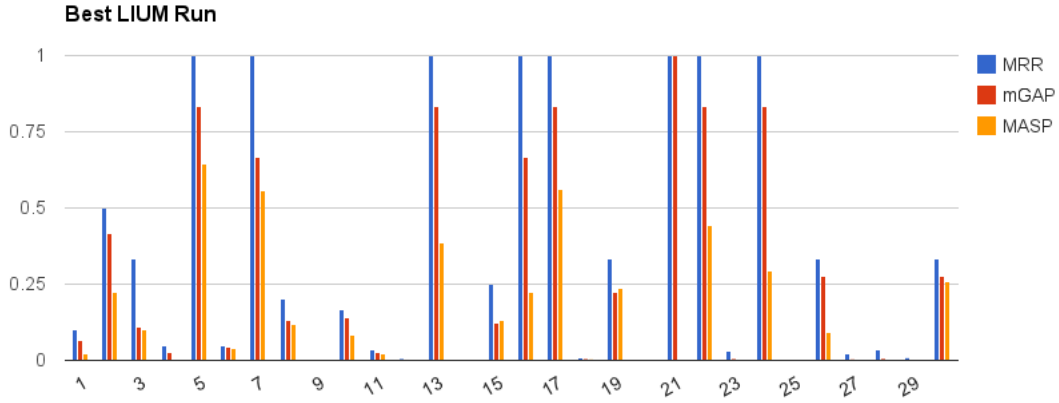

Figure 12: Results for each query for the Best Run for LIUM transcript.

5 Conclusion

In this work we study Information Retrieval from large collection of audio-visual data. We are especially interested in the impact of segmentation of recordings into smaller units on the Passage Retrieval quality. We study several approaches to such segmentation; the regular window-based segmentation using the time in the recordings outperforms semantic-based segmentation. We also study other techniques which influence retrieval quality: we explore query

Num	Query
2	Profit Partner programe talks about growing business faster.
3	Curtis Baylor of Allstate gives a small piece of planning advice for small business using his basic three factors.
5	One of the biggest problems with the EEE PC 900 laptop and how to solve it.
7	Its about an annual Brooklyn Blogfest where bloggers and fans meet each other and have fun.
8	"Hey guys, I thought this was pretty. . . interesting to listen to. Minus the fact it should be Judaism,and not Judism (sounded like Druidism HAH) I thought his reaction to the news of conversion was pretty funny."
13	Medical Marijuana clinics in California.
15	Its about wrong impressions created by artists on Angels and clarifies the authentic interpretation as per the Bible.
16	California to pass law intended to put an end to domestic violence by outing the abusers in public.
17	What an unusual painting interview
21	Too Big to Fail composed by Austin Launge Lizards
22	Its a Grit TV presentation on Green Party Presidential Candidate.
24	Sending automatic emails whenever you add new content to blogs or web sites.
30	What Would Google Do By Jeff Jarvis

Table 11: Queries with the highest MRR score (equal to 1) for the LIMSI transcripts.

expansion, various retrieval models and their tuning, utilization of metadata, and post-filtering of the retrieved results.

Evaluation of our experiments is carried out on the MediaEval 2012 Benchmarking collection. The highest score is achieved with the LISMI transcripts, for regular time segmentation with 45-seconds-long segments and 15-seconds-long overlap, employing metadata, stemming, and stopwords list, using Hiemstra Language Model with parameter 0.35, and applying simple filtering of overlapping results. In this case, the MRR value is 0.49, mGAP 0.35, and MASP 0.21. Query expansion is not applied because it helps only in case when no metadata and only a part of whole query are used. Similarly, the LIUM transcripts achieve better results for the Baseline Run (despite higher word error rate) but the best

Num	Query
9	Its of serious comics on science related subjects.
14	This is the process a comic book goes through before it's released.
18	"This is a video that includes two different poets, both doing readings of their work."
20	"I found this clip simple but very helpful. I couldn't remember how to create a new new pattern, but the steps were pretty simple and easy to follow. Hope it can help you guys out too! Enjoy."

Table 12: Queries with the lowest MRR score (equal to 0) for the LIMSI transcripts.

scores are achieved using the LIMSI transcripts.

Comparing the other approaches in MediaEval Benchmarking, our approach achieves the highest scores in all measures (Eskevich et al., 2012b). However, we hope that especially precision of our approach could still be improved.

6 Future Work

Our future plans could be divided into several points:

1) We would like to improve precision of the semantic segmentation. We believe that the precision could be increased by improving feature-based segmentation, e.g. more training data, more features (lexical features, various cue words, and possibly prosodic features such as energy of the utterance) can be employed and results improved.

2) As it was proved, the segment length is especially important in Passage Retrieval. Therefore we need to integrate the regulation of the length of the segments to our solution.

3) Next step will be implementation of the segmentation into a real world environment. All the used methods are independent from data type and language, except training data for the feature-based approach. However, the transcripts of the recordings are needed for the employment of the retrieval. The segmentation is assumed to be utilized in two projects: CEMI project and Dialogy corpus[5].

[5] http://ujc.dialogy.cz/

6.1 Time Schedule

In the first year of the study, an experimental system for the segmentation and further information retrieval was created. In the second year we would like to improve the segmentation, especially use machine learning techniques and advanced features. In the next year, we will start to work on the cooperation of the segmentation with project CEMI and Dialogy. The integration with the project will be finished in the fourth year of study and the final version of the thesis will be submitted.

References

Allan, J. (1995). Relevance feedback with too much data. In *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '95, pages 337–343, New York, NY, USA. ACM.

Ballantine, J. (2004). Topic Segmentation in Spoken Dialogue. Master's thesis.

Beeferman, D., Berger, A., and Lafferty, J. (1999). Statistical models for text segmentation. *Machine Learning*, 34(1-3):177–210.

Blei, D. M. and Moreno, P. J. (2001). Topic segmentation with an aspect hidden markov model. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '01, pages 343–348, New York, NY, USA. ACM.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and Regression Trees*. Wadsworth Inc.

Buckley, C., Salton, G., Allan, J., and Singhal, A. (1994). Automatic query expansion using SMART: TREC 3. In *Proceedings of Text REtrieval Conference*, pages 69–80. Department of Commerce, National Institute of Standards and Technology.

Callan, J. P. (1994). Passage-level evidence in document retrieval. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '94, pages 302–310, New York, NY, USA. Springer-Verlag New York, Inc.

Choi, F. Y. Y. (2000). Advances in domain independent linear text segmentation. In *Proceedings of the 1st North American chapter of the Association for*

Computational Linguistics conference, NAACL 2000, pages 26–33, Stroudsburg, PA, USA. Association for Computational Linguistics.

Church, K. W. (1993). Char_align: a program for aligning parallel texts at the character level. In *Proceedings of the 31st annual meeting on Association for Computational Linguistics*, ACL '93, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

Dielmann, A. and Renals, S. (2005). Multistream dynamic bayesian network for meeting segmentation. In *Proceedings of the First international conference on Machine Learning for Multimodal Interaction*, MLMI'04, pages 76–86, Berlin, Heidelberg. Springer-Verlag.

Eisenstein, J. and Barzilay, R. (2008). Bayesian unsupervised topic segmentation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 334–343, Stroudsburg, PA, USA. Association for Computational Linguistics.

Eskevich, M., Jones, G. J. F., Chen, S., Aly, R., Ordelman, R., and Larson, M. (2012a). Search and Hyperlinking Task at MediaEval 2012. In *MediaEval 2012 Workshop*, Pisa, Italy.

Eskevich, M., Jones, G. J. F., Chen, S., Aly, R., Ordelman, R., and Larson, M. (2012b). Search and Hyperlinking Task at MediaEval 2012. http://www.slideshare.net/MediaEval2012/search-and-hyperlinking-task-at-mediaeval-2012.

Eskevich, M., Jones, G. J. F., Wartena, C., Larson, M., Aly, R., Verschoor, T., and Ordelman, R. (2012c). Comparing retrieval effectiveness of alternative content segmentation methods for internet video search. In Lambert, P., editor, *CBMI*, pages 1–6. IEEE.

Eskevich, M., Magdy, W., and Jones, G. J. F. (2012d). New metrics for meaningful evaluation of informally structured speech retrieval. In Baeza-Yates, R. A., de Vries, A. P., Zaragoza, H., Cambazoglu, B. B., Murdock, V., Lempel, R., and Silvestri, F., editors, *ECIR*, volume 7224 of *Lecture Notes in Computer Science*, pages 170–181. Springer.

Galuščáková, P. (2012). Application of topic segmentation in audiovisual information retrieval. In Šafránková, J. and Pavlů, J., editors, *WDS'12 Proceedings*

of Contributed Papers, pages 118–122, Praha, Czechia. Matematicko-fyzikální fakulta Univerzity Karlovy, Matfyzpress.

Galuščáková, P. and Pecina, P. (2012). CUNI at MediaEval 2012 Search and Hyperlinking Task. In Larson, M., Schmiedeke, S., Kelm, P., Rae, A., Mezaris, V., Piatrik, T., Soleymani, M., Metze, F., and Jones, G., editors, *Working Notes Proceedings of the MediaEval 2012 Workshop*, volume 927 of *Workshop Proceeding*, Pisa, Italy. CEUR Workshop Proceedings.

Galuščáková, P., Pecina, P., and Hajič, J. (2012). Penalty functions for evaluation measures of unsegmented speech retrieval. In *CLEF*, volume 7488 of *Lecture Notes in Computer Science*, pages 100–111. Springer.

Hauptmann, A. G. and Witbrock, M. J. (1998). Story segmentation and detection of commercials in broadcast news video. In *Proceedings of the Advances in Digital Libraries Conference*, ADL '98, pages 168–179, Washington, DC, USA. IEEE Computer Society.

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

Hearst, M. A. and 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*, SIGIR '93, pages 59–68, New York, NY, USA. ACM.

Hiemstra, D. (2001). *Using Language Models for Information Retrieval*. PhD thesis, Enschede, Netherlands.

Hsu, W. H., Chang, S.-F., Huang, C.-W., Kennedy, L. S., Lin, C.-Y., and Iyengar, G. (2004). Discovery and fusion of salient multimodal features toward news story segmentation. In Yeung, M. M., Lienhart, R., and Li, C.-S., editors, *Storage and Retrieval Methods and Applications for Multimedia 2004, 20 January 2004, San Jose, CA, USA*, volume 5307 of *SPIE Proceedings*, pages 244–258. SPIE.

Jeong, M. and Titov, I. (2010). Multi-document topic segmentation. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10, pages 1119–1128, New York, NY, USA. ACM.

Kaszkiel, M. and Zobel, J. (1997). Passage retrieval revisited. In *Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '97, pages 178–185, New York, NY, USA. ACM.

Kaszkiel, M. and Zobel, J. (2001). Effective ranking with arbitrary passages. *Journal of the American Society for Information Science and Technology*, 52(4):344–364.

Kauchak, D. and Chen, F. (2005). Feature-based segmentation of narrative documents. In *Proceedings of the ACL Workshop on Feature Engineering for Machine Learning in Natural Language Processing*, FeatureEng '05, pages 32–39, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kelm, P., Schmiedeke, S., and Sikora, T. (2009). Feature-based video key frame extraction for low quality video sequences. In *WIAMIS*, pages 25–28. IEEE Computer Society.

Kozima, H. (1993). Text segmentation based on similarity between words. In *Proceedings of the 31st annual meeting on Association for Computational Linguistics*, ACL '93, pages 286–288, Stroudsburg, PA, USA. Association for Computational Linguistics.

Lamel, L. and Gauvain, J.-L. (2008). Speech processing for audio indexing. In *Proceedings of the 6th international conference on Advances in NLP*, GoTAL '08, pages 4–15. Springer-Verlag.

Larson, M., Eskevich, M., Ordelman, R., Kofler, C., Schmiedeke, S., and Jones, G. J. F. (2011). Overview of MediaEval 2011 Rich Speech Retrieval Task and Genre Tagging Task. In Larson, M., Rae, A., Demarty, C.-H., Kofler, C., Metze, F., Troncy, R., Mezaris, V., and Jones, G. J. F., editors, *MediaEval*, volume 807 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Lin, J. J., Quan, D., Sinha, V., Bakshi, K., Huynh, D., Katz, B., and Karger, D. R. (2003). What makes a good answer? The role of context in question answering. In Rauterberg, M., Menozzi, M., and Wesson, J., editors, *INTERACT*. IOS Press.

Liu, B. and 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, pages 673–674, New York, NY, USA. ACM.

Liu, X. and Croft, W. B. (2002). Passage retrieval based on language models. In *Proceedings of the eleventh international conference on Information and knowledge management*, CIKM '02, pages 375–382, New York, NY, USA. ACM.

Malioutov, I. and Barzilay, R. (2006). Minimum cut model for spoken lecture segmentation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, ACL-44, pages 25–32, Stroudsburg, PA, USA. Association for Computational Linguistics.

Malioutov, I., Park, A., Barzilay, R., and Glass, J. R. (2007). Making sense of sound: Unsupervised topic segmentation over acoustic input. In Carroll, J. A., van den Bosch, A., and Zaenen, A., editors, *ACL*. The Association for Computational Linguistics.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, New York.

Melucci, M. (1998). Passage retrieval: a probabilistic technique. *Information Processing and Management*, 34(1):43–68.

Misra, H., Yvon, F., Jose, J. M., and Cappe, O. (2009). Text segmentation via topic modeling: an analytical study. In *Proceedings of the 18th ACM conference on Information and knowledge management*, CIKM '09, pages 1553–1556, New York, NY, USA. ACM.

Mittendorf, E. and Schäuble, P. (1994). Document and passage retrieval based on hidden markov models. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '94, pages 318–327, New York, NY, USA. Springer-Verlag New York, Inc.

Mohri, M., Moreno, P., and Weinstein, E. (2010). Discriminative topic segmentation of text and speech. *Journal of Machine Learning Research - Proceedings Track*, 9:533–540.

Morris, J. and Hirst, G. (1988). Lexical cohesion, the thesaurus, and the structure of text. Technical report, Computer Systems Research Institute, University of Toronto, Toronto. Technical Report CSRI219.

Nguyen, V. C., Nguyen, L. M., and Shimazu, A. (2011). Improving text segmentation with non-systematic semantic relation. In *Proceedings of the 12th international conference on Computational linguistics and intelligent text processing - Volume Part I*, CICLing'11, pages 304–315, Berlin, Heidelberg. Springer-Verlag.

Papka, R. and Allen, J. (1997). Why bigger windows are better than smaller ones. Technical report, Amherst, MA, USA.

Pevzner, L. and Hearst, M. A. (2002). A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28(1):19–36.

Ponte, J. M. and Croft, W. B. (1997). Text segmentation by topic. In Peters, C. and Thanos, C., editors, *ECDL*, volume 1324 of *Lecture Notes in Computer Science*, pages 113–125. Springer.

Pye, D., Hollinghurst, N. J., Mills, T. J., and Wood, K. R. (1998). Audio-visual segmentation for content-based retrieval. In *ICSLP*. ISCA.

Reynar, J. C. (1994). An automatic method of finding topic boundaries. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, ACL '94, pages 331–333, Stroudsburg, PA, USA. Association for Computational Linguistics.

Roberts, I. and Gaizauskas, R. J. (2004). Evaluating passage retrieval approaches for question answering. In McDonald, S. and Tait, J., editors, *ECIR*, volume 2997 of *Lecture Notes in Computer Science*, pages 72–84. Springer.

Rousseau, A., Bougares, F., Deléglise, P., Schwenk, H., and Estève, Y. (2011). LIUM's systems for the IWSLT 2011 Speech Translation Tasks. In *IWSLT*, San Francisco (USA).

Salton, G., Allan, J., and 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*, SIGIR '93, pages 49–58, New York, NY, USA. ACM.

Song, F., Darling, W. M., Duric, A., and Kroon, F. W. (2011). An iterative approach to text segmentation. In *Proceedings of the 33rd European conference on*

Advances in information retrieval, ECIR'11, pages 629–640, Berlin, Heidelberg. Springer-Verlag.

Tellex, S., Katz, B., Lin, J., Fernandes, A., and Marton, G. (2003). Quantitative evaluation of passage retrieval algorithms for question answering. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, SIGIR '03, pages 41–47, New York, NY, USA. ACM.

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)*, Borovets, Bulgaria.

Tiedemann, J. and Mur, J. (2008). Simple is best: experiments with different document segmentation strategies for passage retrieval. In *Coling 2008: Proceedings of the 2nd workshop on Information Retrieval for Question Answering*, IRQA '08, pages 17–25, Stroudsburg, PA, USA. Association for Computational Linguistics.

Tür, G., Stolcke, A., Hakkani-Tür, D., and Shriberg, E. (2001). Integrating prosodic and lexical cues for automatic topic segmentation. *Computational Linguistics*, 27(1):31–57.

Voorhees, E. (1999). The TREC-8 question answering track report. In *Proceedings of the 8th Text REtrieval Conference (*TREC-*8)*, pages 77–82. NIST.

Wartena, C. (2012). Comparing segmentation strategies for efficient video passage retrieval. In Lambert, P., editor, *CBMI*, pages 1–6. IEEE.

Wu, Y.-C. and Yang, J.-C. (2008). A robust passage retrieval algorithm for video question answering. *IEEE Trans. Circuits Syst. Video Techn.*, 18(10):1411–1421.

Xu, J. and Croft, W. B. (1996). Query expansion using local and global document analysis. In *In Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 4–11.

Yun Hsueh, P. and Moore, J. D. (2007). Combining multiple knowledge sources for dialogue segmentation in multimedia archives. In Carroll, J. A., van den

Bosch, A., and Zaenen, A., editors, *ACL*. The Association for Computational Linguistics.