Towards Effective Retrieval of Spontaneous Conversational Spoken Content

Maria Eskevich

Diploma in Linguistics

A dissertation submitted in fulfilment of the requirements for the award of

Doctor of Philosophy (Ph.D.)

to the

Dublin City University
School of Computing

Supervisor:
Dr. Gareth J.F. Jones

January 2014

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Ph.D. is entirely my own work, that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed:

(Candidate) ID No.: 59111445

Date:

Contents

List of Figures

List of Tables

Abstract

The continuing development in the technologies available for recording and storage of multimedia content means that the volume of archived digital material is growing rapidly. While some of it is formally structured and edited, increasing amounts of it are user generated and informal.

We report an extensive investigation into effectiveness of speech search for challenging informally structured spoken content archives and the development of methods that address the identified challenges. We explore the relationship between automatic speech recognition (ASR) accuracy, automated segmentation of the informal content into semantically focused retrieval units and retrieval behaviour. We introduce new evaluation metrics designed to assess retrieval results according to different aspects of the user experience. Our studies concentrate on three types of data that contain natural conversations: lectures, meetings and Internet TV. Our experiments provide a deep understanding of the challenges and issues related to spoken content retrieval (SCR). For all these types of data, effective segmentation of the spoken content is demonstrated to significantly improve search effectiveness.

SCR output consists of audio or video files, even if the system is based on their textual representation. Thus these result lists are difficult to browse through, since the user has to listen to the audio content or watch the video segments. Therefore, it is important to start the playback as close to the beginning of the relevant content (jump-in point) in a segment as possible.

Based on our analysis of the issues relating to retrieval success and failure, we report a study of methods to improve retrieval effectiveness from the perspective of content ranking and access to relevant content in retrieved materials. The methods explored in this thesis examine alternative segmentation strategies, content expansion based on internal and external information sources, and exploration of the utilization of acoustic information corresponding to the ASR transcripts.

To my beloved parents Natalia and Vladimir

Acknowledgments

Work on the PhD-thesis is a life on its own within a human life, and this special path becomes successful and joyful only when great people join it along the way, and it feels like home. I felt the luck of the Irish brought these people into my life, and I owe them all my sincere acknowledgments.

First of all, I would like to express my profound sense of gratitude to my supervisor, Gareth J.F. Jones, for providing me with the opportunity to do this interesting and challenging research, for expanding my scientific horizons and introducing me to the new up and coming fields of studies in multimedia. Thank you, Gareth, for providing me with quick and precise feedback during all stages of my PhD, and in all time zones. It has been a great pleasure to have the opportunity to work with such a true expert in the field.

I would also like to thank my examiners, Alex Hauptmann and Deirdre Hogan, for their interest in my work and so many invaluable suggestions for its improvement and further development. Andy Way gave me the best prep talk before the viva, and he was right, as it was truly one of the most engaging and insightful discussions about my work that I had during these years. My transfer examiners and chair also deserve my gratitude, as John McKenna, Deirdre Hogan, and Josef van Genabith made suggestions on how to visualize the results to make them clearer to the readers.

Special thanks to proof-readers of this thesis. David and Johannes, I appreciate your help, advice, and drawing comments on the margins. Thanks to those in the field, whose names stay unfortunately unknown to me, who reviewed the academic papers this thesis is built upon.

MediaEval community became an important part of my scientific life over these years, and I would like to thank Martha, Robin, Roeland, and all the MediaEval 'family' for broadening my perspective on research, geography of networking, and bringing excitement about benchmarking and brand new tasks creation.

DCU is a great university, and I have had the pleasure of sharing offices and

organizing Christmas party for some fantastic people who became friends of mine. Thank you, Yvette and Agnes, for introducing me to the social life of the lab and the university clubs. Special thanks to my great desk-buddy Teresa, the person who opened to me the cultural heritage of Ireland, and brought very special flavour to my Irish life. I want to especially thank Özlem for being there all the time when I needed it, and especially for our Sunday lunches in the lab; Lamia for lifting our spirits in the lab being our sunshine in a rainy day and on the occasional snowy afternoon, Lorraine for rugby enthusiam and feminism acknowledgement, Jennifer for the longest runs in my life and the great teaching experience. Going along the PhD path, especially close to the end, is so much easier when you share your feelings, fears and hopes with your fellow PhD students at the same stage, thank you Zeliha and Wei for our mutual hugs and support. Thank you, Sara and Mairead for teaching me yoga to calm and balance my body and mind. Thank you all the CNGL gang over the years: Ankit, Antonio, Sergio and Dani, Joachim, Rasoul, Xiafeng, Robert, Rejwanul, Sandipan, Sudip, Hala, John, Ventsi, as well as our short-time visitors: Laurens, Jon, Sara, Mikel. Thank you Debasis, Manisha, and David for inspiring discussions about work and help with capricious software. Thanks to Eithne, Ríona, and Fiona for their kind help since the first day I arrived in Ireland.

I can never thank and acknowledge enough my parents who are there for me, I feel it every step along the way. Thank you, Danya, for being patient in waiting, and inviting me to our next big adventure in life.

Skype and Facebook allowed me to stay in touch with all important people in my life, and to keep track of life that is passing by, while you are working for a new conference deadline.

Last but certainly not the least, I am thankful to Science Foundation of Ireland for the research grants (Research Frontiers Programme 2008 grant 08/RFP/CMS1677, grant 07/CE/I1142, Centre for Next Generation Localisation) that supported this research. Special thanks to DCU, ISCA, and NTCIR for providing student travelling grants that allowed me to present and discuss my work at wider range of venues.

Chapter 1

Introduction

The increasing availability of digital recording and storage technologies is producing very rapid growth in the archives of audio and multimedia data being stored. Realizing the potential of this material and its reuse requires efficient access methods to locate content of interest to the user. While much research has focused and continues to focus on image-based video retrieval, the work described in this thesis targets information in the spoken content stream in search and retrieval tasks.

Being a natural part of everyday life, speech is as diverse as different human activities involving use of spoken communication. Spoken content ranges from being prepared or scripted as in the case of news broadcasts, to partly prepared material as in case of lectures and meetings, to free conversations where speakers can instantly change topics, pass information back and forth, interrupt each other and talk at the same time without following a fixed agenda.

When these diverse communications are recorded in video or audio files, they represent a broad variety of potentially interesting information. Students might want to listen to a discussion that took place during one of their lectures, or to find a more detailed description on a certain lecture topic from within another lecture. People working within the same company might find it useful to be able to search for information from meetings where their managers introduced a new task or described important changes in a product. Content created by semi- or non-professional users

on the Internet might introduce a novel perspective on a topic that might interest the general audience of the web.

The importance of what has been said in the video/audio is the characteristic that brings together these different search tasks. The visual stream might not be of much interest here because the information is primarily within the spoken content, i.e. the audio stream, where the visual stream might simply be a talking head or a group of individuals in conversation.

In this thesis we provide a detailed investigation of these challenging retrieval scenarios where the main focus is on the spoken content. We target various aspects of the SCR: dataset creation, factors influencing the results, and evaluation techniques. In order to verify our arguments, we compare the impact of our approaches using datasets in different languages (English and Japanese), and of different types (meetings, lectures and Internet TV).

1.1 Spoken Content Retrieval system overview

Information Retrieval (IR) systems seek to enable a user to satisfy their information needs by identifying relevant documents, i.e. documents containing information able to satisfy their information need. The complete IR process consists of several steps: indexing of the document collection, entering a search request describing the information need, and retrieving of documents which are potentially relevant to the information need (Büttcher et al., 2010). In addition, a mechanism for the user to recognize potentially relevant documents and then to access the content of individual documents is required.

In the case of spoken content retrieval (SCR) the collection is a set of multimedia files one or more of which the user expects to contain information which satisfies their information need. However, in reality the multimedia collection has to be preprocessed before it can be indexed for search within the retrieval system (Brown et al., 2001). Thus, retrieval of any type of spoken content requires the use

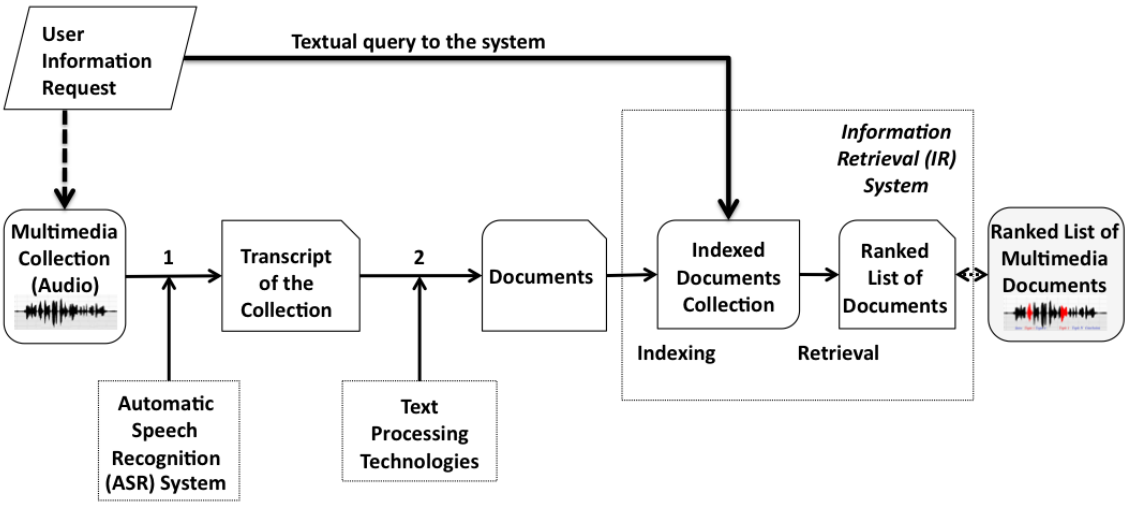


Figure 1.1: General structure of a Spoken Content Retrieval (SCR) system.

of at least two technologies: speech processing (most commonly automatic speech recognition (ASR)) to identify the spoken content and IR. Therefore SCR depends on the availability of effective techniques developed from both the ASR and IR domains, and their successful combination (Garofolo et al., 2000a; Brown et al., 2001; Goldman et al., 2005; Koumpis and Renals, 2005). The general scheme of a SCR system combines text IR and multimedia audio processing as shown in Figure 1.1. Using external technologies of ASR (1) and text processing (2) the multimedia collection of audio/video files is preprocessed into documents that are used within the IR system; the user information request is represented as a textual query to this IR system; and the result list of retrieved documents is given to the user in the form of the original multimedia documents.

Spoken content varies in its level of preparedness and structure: from being prepared, i.e. organized into pre-defined topics and read by a professional speaker, as in case of traditional broadcast news, to different degrees of spontaneity in an academic or working environment, e.g. in case of lectures or meetings (Juang and Furui, 2000; Huang et al., 2001). The quality of recordings depends on the technical parameters of the recording devices. This influences the reliability of the ASR output which represents this data in the retrieval system (Huang et al., 2001; Goldman et al., 2005). Since SCR is the product of the combination of techniques, the impact of

the errors and uncertainty in each of the component systems is not straightforward. Crucial errors from the point of view of one technology by itself, may, within the framework of the SCR system, not be so important and not affect the output result (Garofolo et al., 2000a). Therefore these relations and solutions to the problems raised are the main focus of the work reported in this thesis.

In this thesis we analyze the influence of errors at the level of components on the overall retrieval result, and track how techniques can compensate for each other. This analysis requires specific evaluation methods which form the focus of part of the work described in this thesis. Finally, we use this acquired knowledge to explore techniques to improve SCR effectiveness.

In the following subsections, we overview the principles of textual IR, briefly introduce the potential uncertainty that the use of ASR transcripts may introduce, and discuss the use of textual IR techniques in SCR.

1.1.1 Textual Information Retrieval

As introduced above, an IR system provides the technology to index the documents of a collection, and to perform retrieval of the ones that might be considered relevant to the information need expressed in a query by a user.

Overall textual IR systems typically preprocess the content of both the query and the collection by omitting the words that do not affect the content, e.g. stop words, and by reducing the morphological variety of the same words, e.g. by using suffix stripping (Lovins, 1968; Porter, 1980; Paice, 1990). Then the collection is indexed, and the distance between the query terms and potentially relevant documents from the collection is calculated using one of a number of mathematical functions (Baeza-Yates and Ribeiro-Neto, 1999; Manning et al., 2008). This distance is used to arrange the results in a rank ordered list for return to the searcher.

The vocabulary of the collection consisting of the normalized words, otherwise called *terms*, does not always correspond to the terms used in user queries. This mismatch between expressions in documents and queries represents a significant

challenge for IR systems. In cases where the most terms of the query are present in the target relevant document, the impact of the mismatch of only a part of the query is often very limited, since the other terms are sufficient to retrieve the relevant documents (Allan, 2003). Otherwise it is often possible to expand the query to improve the matching between the query and relevant documents (Allan, 2003; Billerbeck and Zobel, 2005; Manning et al., 2008). This solution needs to have an information source available that is reliable for this purpose, and thus does not distort the document or query, e.g. by introducing irrelevant or too general expansion terms that reduce retrieval effectiveness.

In cases where a document is relevant to only one topic, and the information need corresponds to this topic, a user is satisfied when it is retrieved and they can be expected to be prepared to browse through the whole of the document. However, in reality documents often contain multiple topics and have internal structure of varying complexity. At the same time the queries issued by a user might target only a subset of the information containing in a document, hence only a part of the document might be relevant to the information need. This means that the IR system should take account of the document structure, and to be potentially adjustable to the fact that smaller topically focused content units might correspond to the user information request (Lee and Chen, 2005). Written text often has its structure partially displayed, for example it may be divided into sentences, paragraphs, sections, chapters, etc., with separate headings although a topic may have fuzzy boundaries depending on exactly how the topic is being considered in a specific context (Jurafsky and Martin, 2000). An IR system can often usefully incorporate some segmentation module, or the collection can be preprocessed into segments before the indexing and retrieval are carried out (Hearst, 1993; Choi, 2000). This enables retrieval to be based on topically focused units. Transcripts of spoken content will typically lack these indicative topic boundary markers, meaning that topical segmentation is more challenging in this case. This problem will often be further complicated by the informal structure of the content.

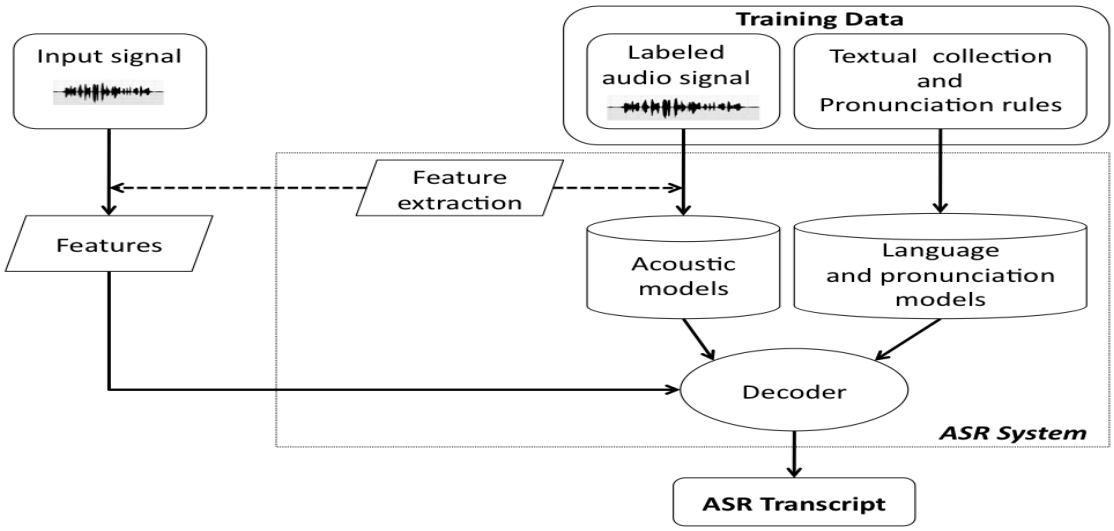


Figure 1.2: General overview of the automatic speech recognition (ASR) system.

1.1.2 ASR transcripts versus human transcripts and written text

The purpose of automatic speech recognition (ASR) is to identify words spoken in the audio stream. Without this transformation the information contained in the audio stream remains so-called "tacit knowledge" that does not easily lend itself to be represented in certain structures and is harder to be transfered from one person to another (Brown et al., 2001). Figure 1.2 shows the main principles of how a transcript is created by a statistical ASR system (Jurafsky and Martin, 2000). Acoustic and language models of the data to be recognised are created based on selected training data (labeled audio signal, collections of texts that use the same language as the targeted spoken content). The same front-end feature extraction module is used to train the acoustic model and to extract the acoustic features from the input signal. The ASR decoder finds the sequence of symbols that is most probable according to the acoustic, language and pronunciation models.

The transcript created by a statistical ASR system always has a certain probability assigned to each consituent unit, e.g. sub-word, phoneme, syllable, word, that reflects how reliable these units in the transcript might be (Jurafsky and Martin, 2000; Huang et al., 2001). As the spoken content might be informal, and the conver-

sation might go back and forth between several topics, even 100% accuracy of the transcription cannot garantee its readability. In general, however it is in practice even harder to read due to errors and absence of such structures as sentences, paragraphs etc. While its conversion into textual format does not necessarily produce a readable transcript, it does mean that it becomes easier to store and process within other applications, e.g. an IR system (Brown et al., 2001; Lee and Chen, 2005; Goldman et al., 2005; Chelba et al., 2008). However, the transcript text has to be treated carefully in the same way as the written one, bearing in mind that potential ASR errors may impact on further processing. The important meaningful words spoken in the audio data might not be recognized correctly by the ASR system, in which cases they will be replaced by other incorrect, meaningful or otherwise common words in the transcript. Since common words are usually removed as stop words by an IR system, these errors in the ASR transcript decrease the chances of the correspondent item being retrieved since important information has been lost or changed, but do not result in false retrieval. Substitution of more meaningful incorrect words in a transcript may be more problematic.

ASR systems vary depending on the task for which they have been created. Early in their development when computational and model training resources were limited, the recognition of individual sounds and isolated words was the target (Rabiner and Levinson, 1981). Currently ASR technologies using much larger training sets and more powerful computers can deal with continuous spoken signal in alternative environments. The actual quality of the ASR result depends on the data the system was built with, the form of speech to be recognized, the acoustic environment and the hardware used for data capture and recognition.

Large vocabulary continuous speech recognition (LVCSR) systems generally attempt to provide a full transcript of the spoken input. They require the collection of data to build detailed acoustic models that correspond to the data to be recognized and large amounts of text to build language models. However, the vocabulary of the system is limited to that chosen when the system was constructed, and any word

outside this selected vocabulary has no possibility to be recognized correctly. Thus this is the first source of potential errors in ASR, even if the rest of the transcript is perfect. As outlined above, errors in the transcript can arise from other sources as well. If the word sequence that was actually spoken has a low probability of occurence, it might be replaced by similar sounding sequence with a higher overall probability in the final ASR output. Alternatively the sounds can be best matched to the wrong models, and further decoding of the sounds into words may result in a high error rate.

The vocabulary of an ASR system may be larger than that which actually appears in the output transcript because certain word combinations are more likely to occur due to the training and language modelling in the ASR system, and appear in the result output instead of potential correct ones. Thus the ASR transcript output is limited not only by the ASR system vocabulary, but also by the word probabilities that are learnt on the training data collection (Jones et al., 2007).

The structure that is usually an intrinsic part of a written text is much harder to deduce from an audio stream (Lee and Chen, 2005). For example, while the ASR transcripts may contain information about the pauses between speech segments, and speaker changes. these units will highly depend on the context, and are hard to generalize upon. The same speaker may talk throughout a long lecture and cover many topics, thus the segmentation on the speaker level will not be very helpful. In case of conversations, a part of the discussion involving several speakers may represent a relevant segment, thus again speaker segmentation might not be useful in this task. The pauses that the speaker naturally takes to take a breath or to make a break in the delivery on purpose cannot be used in combination with the knowledge of the topic because they do not always correspond to a topic break. They might instead signify an important speech segment within the same topic.

While sentence segmentation is currently available for some ASR systems and languages (Gauvain et al., 2002), it does not feature in most ASR systems. Thus segmenting the ASR generated transcripts represents a harder task than in case

of manually written text, because topical segmentation methods usually rely on sentences, e.g. TextTiling and C99 algorithms (Hearst, 1993; Choi, 2000). It is possible to use so-called pseudo sentences, assuming that a full stop should be put in the transcript at each N number of words, however this does not reflect actual semantically coherent sentence units.

Overall, an ASR transcript is not equivalent to a written text for several reasons: the words contain potential errors, the ASR vocabulary and internal weighting poses potential limitations on the words being used in the transcript; and finally the transcript is harder to read and further segment due to lack of structural information even in case of a perfect transcript on the level of words. Also much spoken data (particularly spontaneous content) has different linguistic structure to written text, the language model (LM) of an ASR is often trained on written text, so the LM of an ASR system is often not a good model of the spoken data which is to be recognized. Developments in the ASR field are seeking to close this gap, however it is as yet far from perfect, especially because ASR systems usually need costly adaptation of both their acoustic and language models when there is a change in the content to be recognized.

1.1.3 Overview of SCR main challenges

Following the discussion in the previous sections about IR and ASR principles, in this section we give a brief overiew of the SCR main challenges in context of IR and ASR development and their influence on SCR systems.

ASR errors in transcript: sources and impact on SCR behaviour for different types of content

The effectiveness of an ASR system is usually evaluated via a comparison of the transcript against a perfect manual representation of the spoken content. ASR effectiveness is generally measured in terms of Word Error Rate (WER), the number of individual words in the transcript that were substituted, inserted and deleted as

compared with manual transcript is divided over the overall number of recognized words (Levenshtein, 1966; Jurafsky and Martin, 2000).

ASR systems are generally developed to minimize the WER for all types of spoken content. With the emergence of early LVCSR technologies, initial experiments on SCR retrieval using small private collections of several hours were carried out (James, 1995), as well as projects on the larger scale (more than 1000 of hours) that focused on browsing and access to this vast amount of data (Wactlar et al., 1996). Altogether, once ASR systems had achieved comparatively low WER for well structured data, e.g. broadcast news, the first comparative retrieval experiments on spoken content on a large scale dataset appeared within Text REtrieval Conference (TREC) (Garofolo et al., 1997).

The TREC SDR began with simple known-item search (when there is only one relevant document in the collection) on the basis of the ASR output: the Spoken Document Retrieval (SDR) track at TREC-6 was carried out on broadcast news with known story boundaries (Garofolo et al., 1997). In the next years at TREC-7 and TREC-8 the SDR Track introduced the task of an ad-hoc search, i.e. a scenario when there might be more than one relevant document for a query (Garofolo et al., 1998, 2000a). Over the years the task continued to target retrieval from the broadcast news corpora that grew in size (from circa 40 hours to more than 500) and variety (English Broadcast News Speech corpus (HUB-4), DARPA Topic Detection and Tracking corpus (TDT-2)). In general the WER level on this data is less than 30%. Results of these campaigns led to the overall conclusion that this WER is sufficient for robust retrieval performance and that spoken retrieval can be concluded a "success story" (Garofolo et al., 2000a). However, these experiments only demonstrated that a good level of WER achieves effective search on broadcast news data. In this situation, the target documents have a well defined topically coherent structure, and the queries are posed to these exact units. Thus even if the boundaries are not available, they are easier to define than in case of informal conversational speech material.

Following the progress of ASR technologies application to more challenging types

of spoken content, SCR moved its focus from very formal broadcast news datasets, through lectures (Akiba et al., 2008) and meetings (Carletta, 2007; Morgan et al., 2003), towards totally spontaneous conversational content of personal recordings (Oard et al., 2002; Pecina et al., 2007; Eskevich et al., 2012b). However, traditional IR techniques implemented before for text and audio broadcast retrieval, fail to address the greater challenges of these tasks and require new solutions that take into account the nature of this more informal content. The features to consider are:

- domain specific out-of-vocabulary words (OOV), e.g. in lectures or in meetings;
- higher and more varied levels of WER due to varied recording conditions and pronounciation styles;
- general knowledge shared by meeting participants meaning that significant concepts might not be articulated directly in the discourse;
- the lack of structure in this more informal content.

Document structure and topic boundaries within audio files

The process of going through the result list of potential documents is considerably more time-consuming for spoken content compared to text, because the user has to listen to or to watch the audio/video segment. Depending on the SCR scenario, users may be prepared to spend time listening to the whole audio or watching the whole video. However, when the files become longer and more complicated in structure, i.e. containing conversations covering several topics, the SCR system should support bringing the user to the exact starting point topic of interest expressed in the user's query. This can be achieved when the audio file is segmented into smaller units at some point, and these are presented to the user as a result.

Previous SCR research (Garofolo et al., 1997, 1998, 2000a) has been mostly carried out on data that had the same structure as textual retrieval corpora. These

collections contained transcripts of spoken documents (broadcast news) and sets of queries related to them, where relevance was assumed to pertain to the whole spoken document. It allowed the assessment of SCR applying evaluation metrics developed for text document retrieval, such as Mean Average Precision (MAP), and to compare these with the retrieval performance on textual data (Garofolo et al., 2000a). This approach is suitable for the case of broadcast news where the topics are distinct and for short documents in the form of news stories where the user information need will typically be satisfied by listening to the whole document within a reasonable amount of time.

Broadcast news stories can vary in length, however the content generally has a defined structure, and it is possible to assign binary relevance to a defined news story, i.e. it is either relevant or not. In the case of longer documents with more complicated or more informal structure, such as lectures or meetings, the retrieval system requires more complex content analysis because only part of the document may satisfy the information need of the user.

The problem of document boundary detection for more informal spoken content, i.e. interviews, was introduced at the SDR Track at TREC-8 with the variation of the task where the boundaries were unknown (Garofolo et al., 1998). The effect of the presence or absence of boundaries on the search performance was followed in the investigations at CLEF 2005-2007, where cross-lingual search was carried out for a collection with known boundaries in an English language task, and unknown boundaries in a Czech language task (White et al., 2005; Oard et al., 2006; Pecina et al., 2007).

Segmentation of spoken material into smaller topic-specific documents requires an additional pre-processing stage preceding the retrieval to identify suitable retrieval units. However, segmentation of informally structured materials where the transcripts contain ASR errors is very challenging. Thus there is a need to understand how noisy segmentation of the collection affects retrieval behaviour, and whether it might be improved with new data specific segmentation techniques.

Availability of diverse representations of the multimedia content in form of audio/visual channels and textual representation allows the application of different segmentation techniques. Segmentation of spoken content can be based on the text of the speech transcripts focusing on changes in the vocabulary (Malioutov, 2006), on time information depending on the use of pauses and speech rate (Luz and Su, 2010), on the use of the acoustic features at the level of speaker characteristics and changes in behaviour through a meeting (Hsueh, 2008), and on combinations of all or some of this information, predicting the top-level topic shifts from the transcript, and then defining the sub-level boundaries through use of the acoustic information (Hsueh and Moore, 2006; Malioutov and Barzilay, 2006). Attempts to segment audio content have not generally been followed by exploration of their potential utility for SCR. Therefore even though the topic of the segmentation of the audio material has been studied quite extensively before, so far no substantial investigation of the relation between segmentation methods and retrieval results has been reported.

Where the playback should start?

In practice, lack of well defined boundaries means that the user may need to listen to parts of the audio/video content that are not actually relevant to their information need, i.e. covering another topic. In the case of poorly defined document boundaries the user might be dissatisfied with the results, even if evaluation metrics that assess the presence of documents containing relevant content, assign good scores to the resulting ranked list of documents, based on the fact that the files with relevant content have been retrieved. Focus on user experience led to the introduction of a concept of the so-called "jump-in point", the actual start of the relevant content in the spoken document. SCR systems must favour approaches that let the user browse through documents that start as close as possible to the ideal topical jump-in point. Thus the closeness to this jump-in point makes the overall search procedure easier, more efficient and more pleasant for the user.

1.2 Research questions

The challenges of SCR and the current state-of-the-art lead to the following Research Questions (RQ) which will be explored in this thesis:

- Research Question 1 (RQ1): What is the relationship between ASR errors in the transcript and retrieval behaviour?
- Research Question 2 (RQ2): Do current evaluation techniques sufficiently reflect all the most important aspects of the user experience in SCR?
- Research Question 3a (RQ3a): How does segmentation of spoken data affect retrieval behaviour?
- Research Question 3b (RQ3b): What are the characteristics of a segmentation method that maximizes SCR effectiveness?
- Research Question 4 (RQ4): How can regions of different speech recognition quality be identified and processed in order to improve overall speech retrieval performance (detection, special treatment in the speech retrieval process)? Otherwise, how can the regions of good recognition be used most efficiently in retrieval process?
- Research Question 5 (RQ5): Can we implement a meaningful approach to SCR of conversational content incorporating task specific segmentation?

We note here that addressing these Research Questions robustly is not possible without experiments on different types of datasets and using evaluation metrics that reflect the user experience in accessing relevant spoken content. As there were no datasets available for certain types of informal conversational content, we dedicate substantial attention in this thesis to test set creation. The analysis of system performance also showed that there is potential for new metrics that better reflect user experience in targeted SCR scenarios. Therefore, new evaluation strategies are developed and introduced.

1.3 Thesis structure

This thesis begins with an overview of the development of the field of spoken content retrieval (SCR), and this is followed by details of the experiments carried out in order to define problems that need to be addressed to advance the state-of-the-art in SCR within the thesis. Deep analysis of SCR experiments reveals the main challenges and highlights the investigation perspectives that might solve them. We focus on the shortcomings of the direct application of text retrieval methods on the speech material, and outline the reasons for them.

The chapters of the thesis are structured as follows:

Chapter 2 provides the background of the ASR and IR technologies on which SCR is based. It overviews the structures and the main principles of both, and descibes the data representation within the systems that are used in combination for SCR. As spoken content segmentation represents a challenging task for SCR systems, we also provide a brief overview of relevant work in text segmentation methods.

Chapter 3 overviews the development of the SCR in the context of datasets, retrieval methods and evaluation techniques. We start with carefully prepared broadcast news datasets that were used for experiments in spoken document retrieval tasks at TREC-6–TREC-9, and give details of metrics that were used to evaluate the ranking of the results. We further describe datasets that are more conversational and less structured in style, such as lectures, meetings and Internet television broadcast. We highlight that in lecture retrieval the main focus has been put on interactive browser development, and that there is a lack of results in retrieval performance, except for the recent work in the SpokenDoc task at NTCIR-9/10/11. We cover the topic of meeting search and discuss potential scenarios and interesting use cases for SCR. We finish our overview by giving examples of the least controlled data environments, such as interviews, personal recordings and Internet television channels that are openly available via created commons license, but highly variable

in quality of the recordings content form and topics.

Chapter 4 introduces the datasets used in our investigations within this thesis. We follow the same order of dataset types as in the previous chapter. Thus we give details of lectures taken from the Corpus of Spontaneous Japanese, and the queries and relevance information created within the SpokenDoc task at NTCIR. We then overview the AMI meeting corpus, and describe our work on developing a query set and gathering corresponding relevance assessments for this dataset. We end the chapter with a description of the construction of a test set for the Blip10000 Internet video dataset using crowdsourcing with the Amazon Mechanical Turk platform. Blip10000 consists of semi-professional user generated (SPUG) content, and crowdsourcing enables us to create queries and relevant assessments using real Internet users.

Chapter 5 elaborates the discussion of the evaluation metrics potentially available for use in SCR. Since standard textual IR metrics that have already been used on spoken data are described in details in Chapter 3, in this chapter we begin by overviewing another metric that was previously applied for assessment of textual retrieval of passages, and introduce new metrics that target the ranking of the results, take into consideration the time a user has to spend on listening to relevant and non-relevant content, and use the information about the distance to the jump-in point that affects the user experience.

Chapter 6 analyses the performance of our initial SCR experiments on the meetings and lectures datasets. This helps to better understand the challenges and potential bottlenecks that SCR techniques have to overcome.

Chapter 7 investigates the use of segmentation methods with overlapping segments that undergo further filtering. These filtered results have segment boundaries (jump-in point) further adjusted using knowledge about semantic and acoustic fea-

tures of each segment.

Chapter 8 describes the implementation of traditional document expansion methods, and their further variation to adapt for the needs of SCR that show statistically significant improvement over the baseline.

Chapter 9 concludes the thesis and outlines some avenues for future research.

Appendix A lists the papers published in several peer-reviewed conference proceedings and journals that cover the research presented in this dissertation. There are also other papers which are related to the work in this thesis.

Appendix B contains the list of all NTCIR-9 SpokenDoc queries with their approximate translation into English.

Appendix C lists the queries created on the basis of the AMI Corpus using the procedure described in Chapter 4.

Appendix D contains full text of one of the joint papers that gives further details of the SCR dataset creation process via crowdsourcing that is described in brief in Chapter 4.

Chapter 2

Background Review of Technologies Underlying Spoken Content Retrieval

Research in spoken content retrieval (SCR) requires substantial knowledge of the basic concepts in several fields including information retrieval (IR), speech processing (SP), and content segmentation (CS). Traditionally IR provides a framework for search within a collection of textual documents. Automatic speech recognition (ASR) systems, as part of SP, convert audio content into a textual representation that can be further processed within IR framework. However, as discussed in Chapter 1, spoken content often differs from written texts in many aspects that have to be taken into account by the SCR system to achieve high search effectiveness. Both potential errors in the ASR transcript and uncertain structure of the content can challenge further IR performance. The unstructured nature of spontaneously created content means that it needs to be pre-processed into suitable retrieval passages, this requires segmentation of the speech transcripts into topical units.

In this chapter we present an overview of IR, SP, and content structuring. Understanding these technologies enables us to to appreciate the challenges and potential solutions for SCR systems.

2.1 Information Retrieval

Information retrieval (IR) systems attempt to satisfy a user's information needs by finding potentially relevant documents from within a collection. Prior to the retrieval process such collection needs to be preprocessed and indexed into a suitable form for efficient searching. Once the system has a representation of the documents, an information request expressed in a query can be given to the system, so that potentially relevant documents can be retrieved based on an IR model.

The following sections overview the document preprocessing steps, and introduce popular IR models, highlighting the one used in this thesis. As ASR transcript used in SCR might contain errors leading to potential information loss, we review the principles of content enrichment that generally help to improve overall IR performance, and alleviate this information loss.

2.1.1 Word Preprocessing

Natural language sentences contain very common words that are of little or no value to the IR process, since they do not help to distinguish between relevant and non-relevant documents[1]. Often referred to as stop words, these common words (e.g. "as", "just", "now") can simply be removed in the first preprocessing stage of collection indexing, while not impacting on the retrieval output of the IR system. Lists, although in general improving retrieval efficiency by reducing the required amount of computation, contain stop words that are not specific to any particular domain (e.g. containing pronouns, interjections, modal verbs, etc) and are available for each particular language, and often are incorporated within the IR systems software. The length of these lists can vary, reaching approximately 600-700 words for the English language[2], but typically only about 50 words for the Japanese language[3].

[1] As human languages vary in structure and forms of representations, within this thesis we focus on describing the details that are relevant for those used in the data collections used for the experiments described in this thesis, e.g. mainly English and partially Japanese.

[2] http://snowball.tartarus.org/algorithms/english/stop.txt

[3] http://dnnspeedblog.com/SpeedBlog/PostID/3187/Japanese-Stop-words

Depending on the complexity of the language structure, the meaningful words remaining after stop word removal might be represented by various inflectional forms. These may impact on the retrieval behaviour, since words referring to the same lemma, i.e. having the same meaning, would be considered as different terms within the IR system. Thus stemming algorithms, such as the Porter Stemmer (Porter, 1980), that can be used to remove the word suffixes that are characteristic of the inflectional forms, so that words with the same underlying stem will be indexed as the same term wherever they occur in the document collection, are used.

The relatively simple morphology of the English language means that suffix stripping is relatively straightforward. Stemming of morphologically rich languages is typically much more complex. Other languages can require sentence, phrase or compound splitting, possibly in addition to stemming, to extract suitable indexing units for IR. In this thesis we focus only on preprocssing for the languages of interest to our experimental datasets, namely English and Japanese.

2.1.2 IR Models

Once the documents are preprocessed, they are indexed within an IR system. The component of any IR system that is responsible for retrieval and ranking of the indexed documents for each individual query is the IR model, i.e. a logic or a weighting scheme that defines the likelihood that a document is relevant to a query. This model might be based on the straighforward presence of the query terms in the document, or be based on a more complex calculation that assigns a relevance probablility to each document depending on a weighting scheme or other statistical model. From an operational perspective, the former has the disadvantage that forming effective queries requires a high level of training of the user. There are a number of standard IR models available which all perform this operation, but have different underlying principles and produce slightly different results. An IR model consists of four constituents: document collection representation; query representation; framework for modelling collection representations, queries and their relationships; and

the ranking function which defines the order among the documents with regard to the query (Baeza-Yates and Ribeiro-Neto, 1999).

The *Boolean model* framework, that represents a query as a list of its consituent terms connected with Boolean operators (AND, OR, and NOT), was popular in commercial systems until the mid 90s (Baeza-Yates and Ribeiro-Neto, 1999). However, it has several disadvantages, as it is difficult to control the size of the output set which is not ranked in any order, and there is no possibility of partial match of the document with the query. Thus, queries tend to be very broad, retrieving large numbers of documents for review, or too specific, potentially failing to retrieve all relevant documents.

An alternative to the unranked output of the Boolean model and using simple keyword queries is provided by best-match IR models. The earlies example of which is the *Vector space model*. This represents both documents and queries as vectors within high-dimensional vector space, and the distance between these two vectors defines the similarity between the relevant document and the query (Salton and Buckley, 1988). The degree of this similarity is assumed to be correlated with relevance of the document to the user's query, and is used to arrange the retrieved results in the ranked order. The quality of the ranking can be improved by assigning weights to the search terms. Various term-weighting techniques can be used to assign weights to the terms in the vectors. These are based on the term frequency within the documents and within the collection. The most popular weighting schemes are based on a *tf-idf* approach. *tf-idf* is reflects the composite weight of a term within a document of a certain collection, and takes into account and normalises both the frequency of the document in the document, and overall in the collection (Manning et al., 2008). Equation 2.1 shows calculation of a simple *tf-idf* weight.

$$tf\text{-}idf_{t,d} = tf_{t,d} \times idf_t \tag{2.1}$$

where $tf_{t,d}$ is the term frequency of the term t in the document d; idf_t is the inverse

document frequency that is defined as shown in Equation 2.2:

$$idf_t = \log \frac{N}{df_t} \tag{2.2}$$

where N is the total number of documents in the collection, and df_t is the document frequency, i.e. a number of documents in the collection where the term t occurs.

An alternation to the vector-space model for best-match ranked retrieval is the *Probabilistic model*, introduced in (Robertson and Spärck Jones, 1976), later known as *binary independence retrieval* (BIR) model. It assumes the previous knowledge about relevant and non-relevant documents within a collection, and distinct term distributions in relevant and non-relevant documents. This information is exploited to calculate the likelihood of the relevance of the documents to the issued query, this relevance weight achieves better retrieval performance than simple term matching.

An alternation to these models is provided by the *Language Model (LM)* approach to IR. This is based on a model of language of the documents in the collection, i.e. the information of the words in natural language (Ponte and Croft, 1998; Hiemstra, 2001). In this case, the documents are ranked according to their probability of generating the query based on their individual language model. Equation 2.3 shows the general calculation used within this approach.

$$P(q|d) = \prod_{i=1}^{n} (\lambda_i P(q_i|d) + (1 - \lambda_i) P(q_i)) \tag{2.3}$$

where $q = (q_1, \ldots q_n)$ is the query comprising of n query terms, $P(q_i|d)$ is the probability of generating the i^{th} query term from a given document d being estimated by the maximum likelihood, $P(q_i)$ is the probability of generating it from the collection and is estimated by document frequency, and λ_i is a smoothing parameter that allows to assign more or less value to the query terms value.

We use the language model approach to IR for all experiments described in this thesis because it does not require prior knowledge about the relevant documents in the collection as in case of BIR model, and that represents better real case scenario.

Furthermore, LM method allows the adjustment of terms weights depending on the fact whether they are present in a query or only in a potentially relevant document, thus it gives potential possibility to rely more on the documents when we have prior knowledge of good ASR quality or be more strict and require the presence of the terms in the documents when we know that ASR might contain high level of errors.

2.1.3 Document expansion

As shown in the previous section, effective IR relies on the match between the search terms in the user query and the relevant documents in the collection. However, the query can use words that do not appear in a relevant document or synonyms which refer to the same topic, or the actual relevant content might only refer to the relevant words without explicitly using them, e.g. the nouns can be replaced by the pronouns. In case of conversational spoken content, potential mismatch between a query and a relevant part of the video/audio has higher probability to arise due to the potential errors in the ASR transcripts, limitations of the fixed ASR vocabulary, and to the fact that the speakers might even use gestures to point to objects under discussion instead of naming them.

One approach to addressing this challenge is document expansion (Billerbeck and Zobel, 2005). The expansion means that missing terms that could address the mismatch issue are added either to the query itself or to the documents in the collection. The collection itself or an external data source can act as a source of these expansion terms. The main challenge when using expansion method is to add only the terms that help the retrieval, and not to add terms to the existing collection which are effectively noise.

The main principle of expansion is as follows: the document to be expanded is given as a query to the target collection, then it is assumed that the top N retrieved documents are relevant, and finally additional terms are extracted from these documents and added to the original document with all terms reweighted (Lavrenko and Croft, 2001; Tao et al., 2006).

The main advantage of using the collection itself for expansion is that there is a higher probability of consistency of the language and the topics, and less chance of adding irrelevant terms which are topically unrelated to the document, and most importantly it is always available. However, if some words were not present in the collection at all, it means that no expansion within the collection will find them and they cannot be added to a related document. Nevertheless they may still appear in a query, and the mismatch problem will remain unsolved. In case of spoken content the same words can be consistently misrecognized by an ASR system, if they are not present in the system vocabulary, or if they have lower probablity than other potential transcript words which are more frequent in the training data. However, as conditions of recordings can be diverse across the collection, and the ASR system might vary in recognition results across different speakers, there might be cases when the same words are recognised correctly in part of the audio documents, and thus can be available for use as expansion terms.

In the case when there is an external collection available, and the target collection may lack important terms for reliable relevance weighting of the documents, e.g. when the documents are extremely short, such as in the case of metadata for the video files (Min et al., 2010) or tweets (Efron et al., 2012), expansion based on an external collection can be implemented. The principle of the procedure stays the same, the documents of the target collection are used as queries to the external collection, and the new terms are added to the target collection. The effectiveness of document expansion has been demonstrated for cases where the target collection and the document collection used for expansion represent different types of data, such as the use of newspaper collection to improve retrieval of spoken news broadcasts (Singhal and Pereira, 1999). Thus state-of-the-art IR techniques provide the framework for experiments with more challenging content than is typically represented in textual form, as is spoken content that conveys the information that may be represented in various ways (textual transcript, acoustic highlights, etc).

2.2 Speech Indexing

As introduced in Chapter 1, spoken content requires a preprocessing stage when it is converted into a transcript in order to be indexed and retrieved within an IR system. In this section we overview the main principles of ASR systems, types of transcripts that might be produced, the associated information that reflects the reliability of the transcript content, and measures to evaluate transcript quality.

2.2.1 ASR principles

Most state-of-the-art ASR technologies are fundamentally statistical, being based on hidden Markov models (HMMs) (Jelinek, 1976). An ASR system typically requires two models which need to be suitably trained: for acoustic representation one needs to have a set of labeled speech data, and for the language modelling one needs to have textual data either from manually transcribed spoken content or from other textual collections, see Figure 1.2. The speech signal to be recognized is given to the system, and the ASR decoder calculates the probabilities for all possible paths between the states of data observations according to the acoustic and language models (AM and LM for acoustic model and language model correspondingly). This allows the ASR system to decode the most likely units (syllables, words) present in the data based on the combination of the acoustic and language model scores (Jurafsky and Martin, 2000; Huang et al., 2001). Thus ASR system performance depends significantly on the nature of the data to be processed, and on whether the training data has the same characteristics as the signal to be recognized (for example, signal to noise ratio, diversity of speakers or the same speaker) (Huang et al., 2001).

Poor examples of the speech data labels lead to errors at the acoustic recognition level, whereas a fixed vocabulary of the language model will not be able to recognize words outside the list, these are called out-of-vocabulary (OOV) words. If the textual collection is large, but not representative of an input audio stream, then even the words that are not genuinely OOV may not be recognized correctly, thus becoming

pseudo OOV. This might happen because these pseudo OOV words are present in the texts used for LM training and respectively in the language model, however other words that have higher probability in the models produce higher probability of the paths created by decoder, thus the former are replaced by the latter in the final ASR output. In cases when the quality and the quantity of the training data correspond to the conditions of the actual data it is used to recognize, system performance can be expected to be more reliable.

2.2.2 ASR transcripts in details: 1-best and beyond

As in the ideal case of manual transcription, where there is a perfect textual representation of the spoken content, ASR systems traditionally provide an output in the same format, i.e. a transcript of the spoken content. In ASR, this transcript is referred to as the 1-best, as it is the "best" version of potential transcript available from the ASR system, i.e. it has the highest probability of being the sequence of words spoken (Jurafsky and Martin, 2000). However, it is important to note that while this 1-best path through a network of potential states has the highest overall likelihood according to the models of the ASR system, it might leave out some words that were actually pronounced, but are highly improbable in the context based on the model.

At the decoding stage the system internally calculates different hypotheses and keeps them in various representative structures: lattices or confusion networks (Murveit et al., 1993; Mangu et al., 2000). The former contains potential words that have different start and end times and are decoded by the system with certain probability, while the latter represents a condensed version of the lattice. When the same word is recognized in the area within several paths, it would be pruned into one possible path in the confusion network. Figures 2.1 and 2.2 show examples of lattice and confusion network structures, taken from (Chelba et al., 2008). Within the lattice there are paths of different length that can assign a different number of words between the states, e.g. path 3⇒7 has one word *looking* associated with it, while

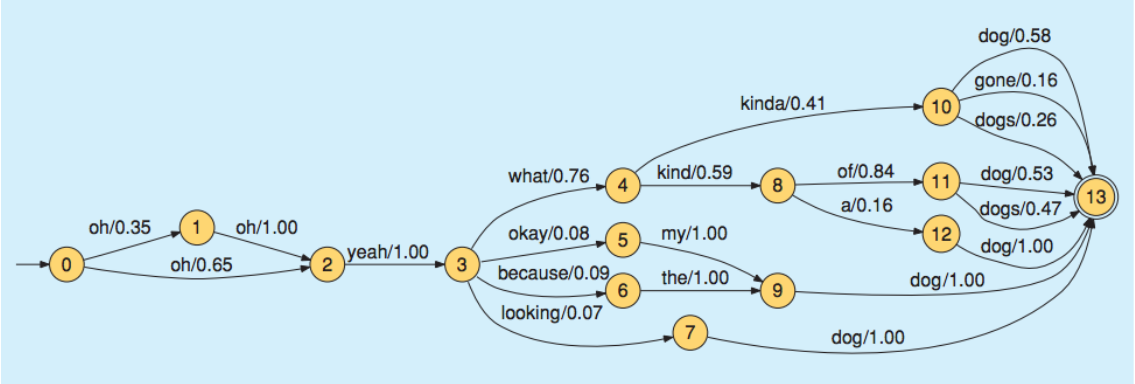


Figure 2.1: Example of an ASR lattice.

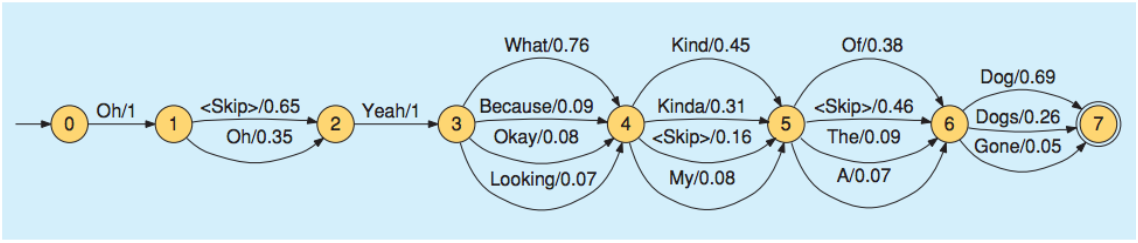


Figure 2.2: Example of an ASR confusion network.

there are also paths 3⇒4⇒8 with *what kind*, 3⇒5⇒9 with *okay my*, 3⇒6⇒9 with *because the*. In the case of confusion networks, there are alternatives only within the same units between two states, which are assigned with different probabilities. The information stored in this form of transcript may be useful for the SCR because it preserves alternative words hypothesis that might be used for the expansion of the 1-best transcript, or its reranking.

The 1-best transcript can be represented not only in words, but also in smaller units that the spoken words consists of sub-word phonemes, or intermediate units that consist of several phonemes or letters of the words - so-called subwords (Ng, 2000).

2.2.3 ASR transcript evaluation: confidence scores, WER

Within each ASR system a probability score that is assigned to the output transcript words by the acoustic and language models is referred to as a confidence score (Wessel et al., 2001; Jiang, 2005). As the training data might vary in quality and

availability, within each individual system it may be possible to adjust the score and assign more weight to one or the other model, thus attributing more reliability to the scores that are based on more reliable data. In the case of the absence of available manual transcripts for the transcribed data, the average confidence score over the ASR transcript or its regions can help to estimate the general quality of the result.

In the case when the proper manual transcript is available for the evaluation, the accuracy of ASR systems is assessed using Word Error Rate (WER), a metric derived from minimum edit distance (Wagner and Fischer, 1974) and the Levenstein Distance (Levenshtein, 1966). This measure reflects how many changes have to be made to the ASR transcript to transform it into the manual correct one. It takes into account 3 types of errors that an ASR system can make, or in other words 3 types of changes which may need to be carried out on the transcript (insertion, deletion, and substitution of the words) to correct it, and is divided over the total number of words in the correct transcript. The higher the WER is, the more difficult it is to browse the transcript and understand what the audio was about. In terms of SCR, higher WER may decrease the retrieval effectiveness of the system, since the relevant terms may be missing from the document transcript, being replaced by irrelevant ones.

Development of ASR systems has been accompanied by the reduction in the WER for many tasks (Fiscus et al., 2007). Current techniques that provide not only the words, but punctuation features as well (Shen et al., 2009; Kolár and Lamel, 2011, 2012) bring the transcript closer to the high level manual transcript quality. Typically reported WERs for the best performing ASR systems for a range of tasks are as follows: broadcast news - 17% (Gauvain et al., 2002); lectures - 15-25% (Cho et al., 2013); meetings - 36% (Hain et al., 2012); multi-genre media archives - 27-31% (Lanchantin et al., 2013).

2.3 Content Segmentation

Documents which cover multiple topics, but do not have pre-segmented boundaries between topical regions represent a challenging task for IR system. This creates problems for a number of reasons. Firstly, terms related to different topics are combined together across topics within a single document. This can make a document containing a relevant region less probable to be retrieved at high rank. Moreover, if a document covers more than one topic, retrieving the whole document is not the best solution for the user, as often only a segment of the main document is relevant to the information need and of interest for a user. This will be especially true for spoken content where the user potentially need to audition large amounts of non-relevant content to find a relevant region.

Content-based document segmentation methods are designed to identify different topical regions within multi-topic documents. Segmentation algorithms have generally been designed for textual data, but these can be applied to spoken data transcripts. However, as the spoken content is more informal in its nature than written text, and the words within ASR transcripts are reliable only with a certain confidence level, methods that rely on these words being all equally correct may produce errors in segmentation.

In the following sections we overview various segmentation methods that can be used for the spoken content: from trivial segmentation methods that are based on fixed length regions to content-based methods that exploit vocabulary changes within text (lexical cohesion based) to determine segmentation boundaries.

2.3.1 Fixed length segmentation methods: time, number of words

The simplest segmentation methods cut the content into chunks of equal predefined lengths (e.g. the same size time units). In the case of spoken content this has an advantage that these units do not depend on the ASR transcript quality, and

therefore are not affected by potential errors or by the nature of spoken content. Segmentation using time dimension allows us to compare the retrieval performance of the same content represented by different ASR systems independent of the behaviour of content-based segmentation, thus it is possible to trace the influence of ASR quality on the ranking of results.

Another simple statistical approach to segmentation is to represent the content as the segments of the same number of words. Moreover, as only non-stop words are actually used within the IR system, it might be reasonable to consider a fixed number of content words only as a parameter for segmentation. These methods depend on the quality of the transcript, however as they use the same number of words being gathered in a segment, they will create the segments of the same length.

Fixed length (in terms of time or words) is an important advantage of these methods, as it allows us to investigate the influence of length variation on the SCR performance. Generally the IR systems tend to be biased towards longer documents, as they contain more information about the topic. This bias is avoided in case of fixed segmentation, thus again the influence of ASR transcript quality can be investigated when compared with the results for the same audio content with manual transcript.

Since fixed length segments are likely to divide semantically coherent regions, they can also be applied in an overlapping manner. Thus some of the segments will truly capture semantically coherent regions while others will be composed of two or more topical fragments.

2.3.2 Lexical cohesion based methods: C99, TextTiling, MCut

Over the last 20 years there has been ongoing interest in the development of methods for automated text segmentation. In principle, all these methods follow a standard approach: information about shifts in lexical content is used to determine whether these shifts signal a change in topic. Two of the best known algorithms are based on lexical cohesion: TextTiling (Hearst, 1993) and C99 (Choi, 2000). Based on this work further methods have been developed which specifically attempt to segment

spoken content, for example LCSeg (Galley et al., 2003), the method of Hsueh and Moore (Hsueh and Moore, 2006; Utiyama and Isahara, 2001; Sharp and Chibelushi, 2008)), and the Minimum Cut (MCut) model described by Malioutov and Barzilay (Malioutov and Barzilay, 2006).

TextTiling computes the cosine similarity between adjacent fixed sized blocks of sentences. The C99 algorithm also calculates the similarity between sentences using a cosine similarity measure to form a similarity matrix. In C99, the cosine scores are then replaced by the rank of the score in the local region and segmentation points assigned using a clustering procedure. MCut regards the segmentation problem as a graph-partioning task that optimizes the normalized cut criterion.

The implementations of Textiling and C99 algorithms are freely available[4] with their default settings as described in (Hearst, 1997) and (Choi, 2000). Since our focus is on understanding the challenges and behaviour of search on unstructured spoken collections, we make no attempt to tune parameters to our datasets within this thesis.

The MCut approach regards the segmentation task as a graph partitioning task that aims to simultaneously optimize the total similarity within each segment and dissimilarity across various segments (Malioutov and Barzilay, 2006). This segmentation was introduced and positioned as a method that targets the challenges of segmenting ASR transcripts, achieving good results for transcript of lectures that contained errors. However, it has the serious disadvantage of not being able to determine the number of segments in the document. This parameter is required to be set explicitly for operation of the algorithm, thus there is a need of a training dataset to be able to predict the expected number of the segments in the document. In case this number varies within the collection, the data needs to be checked manually before the automatic segmentation algorithm is applied.

[4]http://morphadorner.northwestern.edu/morphadorner/textsegmenter/

2.4 Summary

In this chapter we have overviewed the main workflow of IR systems, and highlighted the difference between ASR transcripts and IR for traditional textual document. We have examined the potential need for documents in a collection to be segmented for retrieval in case they cover several topics, and gave an overview of existing segmentation methods that can be applied to address this issue.

The next chapter overviews how IR methods have been implemented for SCR of different data types, and how ASR performance affects retrieval behaviour. Examination of SCR for highly unstructured spoken data will motivate the need for content segmentation.

Chapter 3

Spoken Content Retrieval: Development and Research Questions

Initial experiments in spoken document retrieval (SCR) focused on small collections of several hours of data and restricted tasks to keyword spotting or retrieval of content that has been previously segmented into topically coherent units. These demonstrated the general feasibility of the SCR task with the broadcast news and personal messages recordings as target collections (Rose, 1991; Brown et al., 1994; Foote et al., 1995; James, 1995, 1996).

The results of these early experiments are not directly comparable as the datasets are different, but they laid the foundation to merge the IR and ASR domains into a separate field of SCR. Spoken Document Retrieval (SDR) track at TREC-6, TREC-7, TREC-8, and TREC-9 provided test retrieval collections of broadcast news data that allowed direct comparison of diverse approaches. As audio content of broadcast news used within TREC SDR is of good quality and has low levels of ASR errors, the use of standard IR techniques proved to perform effectively for corresponding transcripts (Garofolo et al., 2000a).

Following the success of TREC SDR, SCR research shifted to more challenging

spoken data sources such as meetings and lectures, with these collections becoming available for potential use in IR experiments. Meeting collections, such as the AMI (Renals et al., 2007) or the ICSI corpora (Morgan et al., 2003), together with lectures archives, such as MIT lectures (Glass et al., 2007), represented a new type of data with less structure in the content, more freedom in pronunciation style and greater variation in recording conditions.

In this chapter we give an overview of the speech search tasks carried out for content with different degrees of formal structure and signal quality, moving on this scale from highest to lowest levels. We describe the datasets, the search tasks and information needs related to them. As the discussion about retrieval from highly unstructured content raises problems of content arrangement into topical units, we introduce different segmentation approaches developed for text and spoken content structure organisation.

3.1 High quality formal speech

In this section we overview SCR experiments based on broadcast news data, including details of the experimental tasks and metrics that were used to evaluate performance. Recording of spoken content with high quality technical characteristics based on a prepared script is possible within the environment of a laboratory, or generally a soundproof studio. The latter is used at radio and television companies, and results in archives that represent valuable data sources which many people may be interested to examine for numerous reasons in the future. Facilitating effective exploration of these resources necessitates the availability of high quality search tools.

Broadcast news is a special type of speech activity because the speaker is generally a professional presenter, usually following the pronunciation norms of the language, while much of the speech is a written linguistically well formed script, and the recording setting is mostly quiet. These aspects of good audio quality and the

vast amount of data that an ASR system can be trained on in the news domain mean that recognition for this type of data is generally reliable. Another particular feature of broadcast news is that these scripted texts have well defined structure of the topics being covered and have distinct borders between discrete news stories, even interviews that have more freedom in a conversational form are generally quite highly structured. Broadcasts are planned beforehand and are carefully edited to be as informative as possible, so that there is little extraneous or redundant information, so listening to the whole section is not wasteful. The nature of the content within a section is also usually self explanatory, i.e. the topic is introduced and developed within the same section, the main subjects of the discussion are named, the script is prepared in a way that assumes only general world knowledge, and all potentially new details are described in detail.

This structure of news programmes and the impartiality of the speaker (presenters only give information about the news and do not express their attitude or opinion) implies a search scenario where the user issues a query about a certain topic which can be anticipated to match with descriptive spoken words in the content, and will typically expect to listen to the whole section of the broadcast programme dedicated to this news story.

Early work on small collections

Early work on the implementation of IR methods in the speech domain was carried out on small private collections. Rose (1991) demonstrated the feasibility of speech-message IR, as distinguished from a more complicated speech-message understanding task that involves semantic and syntactic analysis of the spoken content, by introducing the first end-to-end working SCR system. Prior to the availability of LVCSR transcription systems, which became available for research from the mid 1990s, studies of SCR utilised keyword spotting methods for indexing of spoken content. Keyword spotting was carried out for broadcast news retrieval (Rose and Paul, 1990), and video mail messages (Brown et al., 1994; Foote et al., 1995; Brown

et al., 1996). A keyword spotting system is based on a small predefined vocabulary of typically 50-100 words. These must be chosen in advance of recognition, thus SCR based on KWS is limited to small pre-defined IR tasks. All content outside the keyword vocabulary is recognised as part of a "filler" model designed to model all other acoustic events. While KWS systems are much simpler than an ASR transcription system, they are still prone to recognition errors. While referred to as an IR system, the system described in (Rose and Paul, 1990) could more accurately be described as a message classification tool, since it routed messages (news) into one of a small member of pre-defined classes. To overcome the vocabulary limitation of KWS, the subword phone lattice scanning (PLS) technique was introduced in (James, 1995). This enabled open vocabulary IR by scanning the phone lattice for sequences of phonemes corresponding to words appearing in the user query. A detailed description of the use of PLS for IR is contained in (James, 1995).

Early work in SCR investigated retrieval effectiveness using units of different size and type and in combination to represent the data: Schäuble and Wechsler (1995) used phonetic string representation instead of full words; James (1995, 1996) tried to combine both recognized word and phoneme representations in isolation and Rose and Paul (1990) used subwords; Brown et al. (1994); Foote et al. (1995) and Spärck Jones et al. (1996) combined phone-lattice representation and LVCSR, and Witbrock and Hauptmann (1997) showed that the use of fixed length strings of phonemes performed better than phonetic lattices.

Although the above mentioned experiments showed certain improvements in retrieval effectiveness for the tasks and datasets they used, that grew in size from circa 20 minutes (Rose and Paul, 1990) to several hours (Schäuble and Wechsler, 1995); their results are not directly comparable as the datasets are different, but they laid the foundation to merge the IR and ASR domains into proper SCR. This work led to the conclusion that spoken documents can be retrieved quickly and reliably from within a small collection, and that therefore there was potential for effective SCR on larger scale collections.

The Informedia project

During the 1990s the Informedia project, one of the largest spoken content retrieval systems based on broadcast news was created at Carnegie Mellon University (Wactlar et al., 1996, 1999). The Informedia system included videos of both broadcast news and documentaries, the size of which grew to more than 1000 and 400 hours respectively. Within this framework the retrieval system used both spoken and visual content, however there are no reported results for its retrieval effectiveness for these larger in house collections.

Despite the absence of detailed laboratory of evaluation retrieved effectiveness, the Informedia project is particularly notable as the first practical demonstration of effective retrieval from large-scale archives of digital video recodings.

SDR tasks at TREC Evaluation campaign

The Text REtrieval Conference (TREC) Spoken Document Retrieval (SDR) tracks provided common datasets of broadcast news to run SCR experiments on a larger scale with a controlled comparison environment (Garofolo et al., 1997, 1998, 2000a). The task developed from a known-item task on a small collection, to an ad-hoc retrieval over a large dataset. The main details of the tasks are listed in the Table 3.1.

All versions of the TREC SDR task were based on radio and television news broadcasts. Over the years these were taken from the Linguistic Data Consortium (LDC) English Broadcast News Speech HUB-4 ASR corpus, a subset of the DARPA Topic Detection and Tracking (TDT) corpus. The size of the corpus used for the TREC SDR tasks grew from 50 hours in the first task at TREC-6 to several hundred hours in the last version at TREC-8 and TREC-9. From the beginning of the experiments, the organizers provided the participants with a manual and a baseline ASR transcript. Thus it was possible to have a *Quasi-SDR* run carrying out only the IR experimental part, or a *Full SDR* run doing both the recognition of the corpus and the retrieval within the collection (Garofolo et al., 1997). Since TREC-

		TREC-6	TREC-7	TREC-8	TREC-9
Collection Details	Name	LDC HUB-4		LDC TDT-2	
	Size (hours)	43	87	557	
	Size (documents)	1451	2866	21754	
	Document length (wrds/stry)	276	269	169	169
	Baseline ASR	IBM	NIST/CMU SPHINX	NIST/ BBN Byblos	NIST/ BBN B2
	WER	50%	33.8%/ 46.6%	27.5%/% 26.7%	27.5%/ 26.7%
Retrieval Framework	Paradigm	known	ad-hoc		
	Queries	50	23	49	50
	Unknown story boundaries	No	No	Yes	Yes
	Evaluation metric	MRR	MAP		

Table 3.1: Overview of SDR tracks at TREC 6-9.

7 the participants could not only create their own transcripts and run their IR systems on them, they also had access to other sites transcripts leading to *Cross Recognizer Retrieval* runs (Garofolo et al., 1998). The quality of the transcripts improved over the years. The initial baseline created by IBM using one of their ASR systems had approximately 50% WER (Dharanipragada et al., 1998), though in TREC-7 the University of Cambridge HTK recognition system decreased WER to 24.6% (Johnson et al., 1998). By TREC-8 the size of the corpus (TDT-2) reached circa 557 hours (Cieri et al., 1999). The corpus contained only "close caption" quality transcript, and creation of a proper reference transcript for all this data was beyond the scope and technical resources of the SDR track organisers. Thus the ASR systems used in the task were evaluated only on a small subset of 10 hours of manually transcribed data, and all of the systems managed to get a WER level of less than 30%.

For all versions of the task, the queries, called topics, were created by workers at the National Institute of Standards and Technology (NIST). In TREC-6 there were 2 types of queries targeting the stories with audio quality that is easy or difficult to recognize (Garofolo et al., 1997). With the introduction of the ad-hoc retrieval task at TREC-7, the only requirement for a topic to be selected for the task was

that the number of relevant documents retrieved from within the collection for this query should be: 7 or more in TREC-7 (Garofolo et al., 1998), 1 to 20 for TREC-8 (Garofolo et al., 2000a). In TREC-9 the queries had 2 representations that could be used: short description in 1-2 sentences or shorter keyword only queries (designed to be similar to the description type of queries) (Garofolo et al., 2000b).

Initially all news broadcasts were manually segmented into defined story units prior to retrieval. At TREC-8 a condition of unknown story boundaries became available which was intended to represent a more realistic scenario where the news stories had not been manually segmented. However the task for the participants was not expressed as finding the relevant segments, they were expected to submit "hot spots" or a mid-point of the topical relevant section of the transcript. These hot spots were connected to the known stories, thus allowing the use of traditional document retrieval metrics.

As described in Section 2.1.3, document expansion can potentially improve retrieval performance in cases when there is a mismatch between the query terms and the terms of relevant documents in the collection. In the case of spoken content, the words that are pronounced are not always transcribed correctly by the ASR system, thus the probability of term mismatches is higher than in the case of textual collections. At TREC-7, the AT&T group carried out document expansion using an external data collection containing documents closely related to the target collection (newspapers of the same time as the broadcast recordings) (Singhal and Pereira, 1999), document expansion methods were also shown to improve results for the LIMSI submission at TREC-9 (Gauvain et al., 2000). This can serve as an argument for further document expansion experiments within the SCR framework.

Evaluation

In the case of known-item retrieval task, the single document relevant to a query is known in advance of retrieval. Ad-hoc search on a large collection represents a different situation, since it is not practical to manually assess the relevance of all

documents in collection. Some relevant documents may be available, if they were used to help define the queries. However most relevant documents are identified using a pooling procedure based on the participants submissions. In the pooling procedure the top N ranked documents for each query for a selected number of runs for a task are merged to form a union pool set of unique documents. Manual relevance document assessment is then carried out of the documents in the pool. This type of pooling procedure benefits from varied submissions created using diverse retrieval strategies, since the pool is usually designed to maximize the variety of potentially relevant documents to be assessed (Cormack et al., 1998; Zobel, 1998; Voorhees and Harman, 1999).

Evaluation Metrics

In the case of a known-item search, such as in TREC-6, when there is only one relevant item in the collection for a query, retrieval effectiveness is most often measured in terms of Mean Reciprocal Rank (MRR) metric. The reciprocal rank of a query response is the multiplicative inverse of the rank of the first correct answer. The mean reciprocal rank is the average of the reciprocal ranks of results for a sample of queries Q, as shown in Equation 3.1.

$$MRR = \frac{1}{|Q|}.\sum_{i=1}^{|Q|}\frac{1}{rank_i} \tag{3.1}$$

The main evaluation metric for the TREC SDR experiments targeting ad-hoc search is mean average precision (MAP), one of the most widely used metrics in IR research (Büttcher et al., 2010). MAP combines evaluation of the ranking of all relevant retrieved documents. Equation 3.2 shows the definition of the standard average precision (AP) metric for a single query.

$$AP = \frac{1}{n}.\sum_{r=1}^{N} P[r] \cdot rel(r) \tag{3.2}$$

where n is the number of relevant documents, N is the number of retrieved documents, $P[r]$ is the precision at rank r (the number of relevant retrieved documents divided by the total number of retrieved documents), $rel(r)$ is the relevance of the document ($rel(r) = 1$ if document is relevant, $rel(r) = 0$ if not). MAP is computed by averaging AP across the topic set.

Discussion

Analysis of the TREC SDR results shows that SDR effectiveness, as measured by precision at top ranks and MAP, is largely robust to reasonably high levels of speech recognition errors of around 30% WER, and that with the application of techniques such as document expansion, and the use of external collections to train system parameters, performance comparable to that for accurate manual transcriptions of the speech data can be achieved. At the end of four years of the TREC SDR task, speech search was declared to be a largely solved problem with the remaining challenges being in tasks such as question-answering on speech data and spoken queries (Garofolo et al., 2000a), and broadcast news search moved to the implementation of prototype systems that allow the access to large news collections.

However, this conclusion, though solid enough for the data and challenges investigated in the TREC SDR tasks, is not generally applicable to all SDR tasks. It overlooks fundamental differences in the nature of informal spoken content from written text and scripted speech content such as news data, and the speech recognition challenges of content which is not well matched to the vocabulary of the ASR system and is recorded in challenging environments.

3.2 SCR for Conversational Speech

Natural non-scripted conversations include different types of human communication activities, e.g. lecture presentations, discussions at meetings, and random conversations in daily life or informal voice messaging. These recordings have less structure

Name	eLectures	MIT lectures	TalkMiner
Test collection size (lectures/hours)	6/6	5/6.1	over 10000/NA
Lectures source	University of Exeter	MIT	YouTube, PARC Forum, U.C. Berkeley, blip.tv, coursera.org
Use of ASR transcript	Yes	Yes	No
ASR adaptation		Yes	No
Use of OCR	No	No	Yes
Use of metadata	No	Yes	Yes
Presentation to the user	NA	Video, transcript	Video, OCR transcript, metadata
Slides alignment	Yes	Yes	Yes
Word alignment	NA	Yes	No
Segmentation method	TextTiling	MCut	NA

Table 3.2: Overview of the lecture browser systems examples.

as compared to broadcast news, their content is harder to retrieve as the content itself is less self-explanatory and may not contain all relevant information. In this section we review SCR experiments that have used these more informal data types in retrieval tasks. Specifically we examine search and browsing of lectures and meetings within a collection of recordings. The variability of content, lack of document or topical boundaries, and generally more uncertain structure of conversational speech implies that SCR systems need to address more complicated issues, such as segmentation, definition of regions with more importance, or varying level of transcript quality, than is the case with earlier broadcast news search tasks. We also focus on the identification of suitable jump-in points at which playback should begin within recordings of interviews.

3.2.1 Search of Lectures

The increasing availability of high quality digital recording technologies is promoting recording in various environments, including academic lectures. These recordings potentially enable students to access the contents of lectures that they have attended but have incomplete notes for, enabling them to look up explanations in these recorded lectures, or review specific parts of interest. The growing capacity

of online streaming and emergence of the websites with online courses where the course is structured as a sequence of videos, e.g. `coursera.org` and `udacity.com`, is dramatically increasing the potential audience of students of various ages and background interested in accessing online content for remote study.

Lectures may appear to be similar to broadcast speech because they are generally prepared presentations. However, in reality they more closely resemble conversational dialogues due to their spontaneous conversational style, including features such as hesitations, mispronunciations, and question and answer sessions with audience members (Glass et al., 2004). They are though significantly different to general conversations in that they typically use specialized vocabulary relevant to the specific lecture domain, and also the fact that they are mostly in monologue form with a planned underlying sequential structure. Domain-specific terms represent a challenge for ASR systems because they are often either not part of the recognition vocabulary (out-of-vocabulary words (OOV)) or they have very low probability scores in the language model of the ASR system, due to their low occurence frequency in the training data which is typically taken from diverse not specified textual content. These issues mean that such domain specific words are unlikely to appear in ASR transcripts. Therefore systems developed for browsing lecture datasets make use of any available additional information to adapt to the recognition domain. For example by using the texts on the slides of the actual lecture or textbooks related to the lecture domain the systems can refine the general language model for the lecture specific topic (Lee and Lee, 2008; Glass et al., 2007). Otherwise the SCR system can improve the retrieval effectiveness by expanding the lecture transcript in the index by including search terms from external data sources such as textbooks or knowledge databases (Jones and Edens, 2002).

Lectures browser system examples

In this section we describe a range of examples of lecture browsers to highlight the important features of existing prototype systems. The details of these systems are

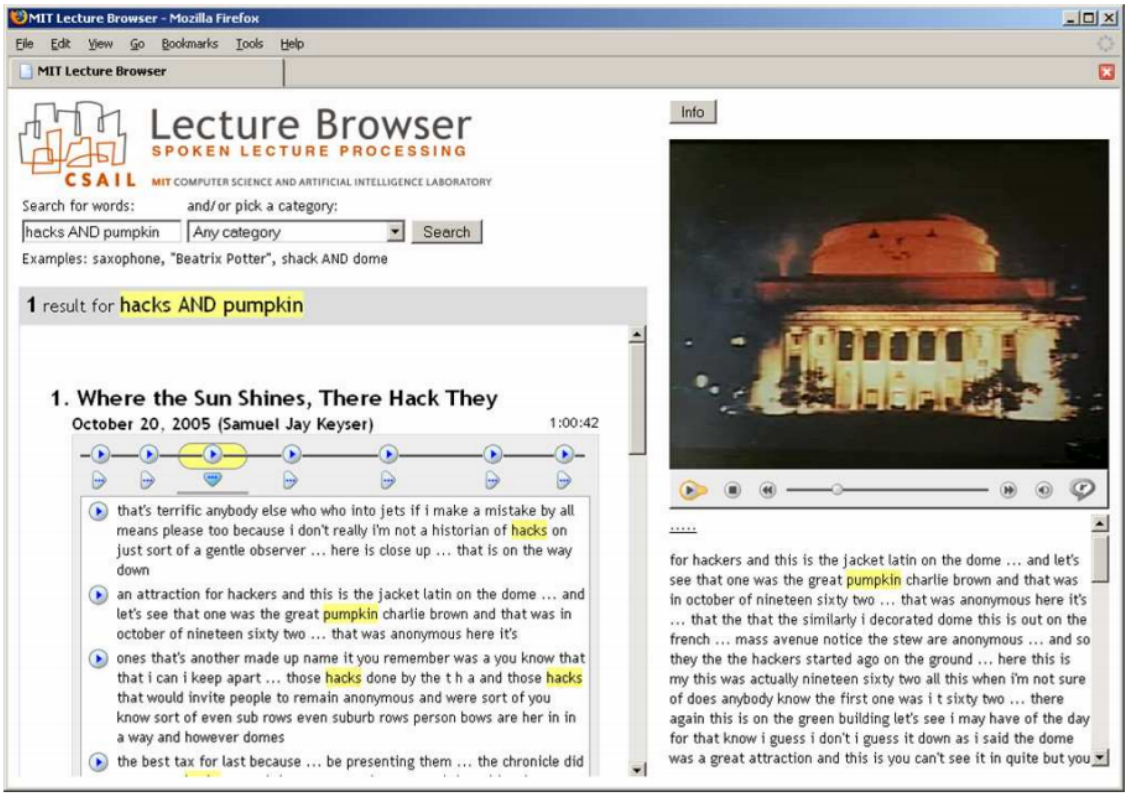


Figure 3.1: Screenshot of a prototype version of the online MIT Lecture Browser.

summarized in Table 3.2.

A method of alignment of audio-video content with presentation slides for eLectures was introduced in Jones and Edens (2002). This represents a combination of ASR technologies and IR methods. Retrieval methods were used to assign regions of a transcript with the audio signal from parts of the lecture to corresponding slides from the lecture presentation. However individual retrieval results on the level of lectures or segments for this corpus were not reported.

The MIT lectures system is based on search of recordings of lectures on specific and general topics (Glass et al., 2007). The language model for the ASR component is tuned for the lecture either using information from slides or web search results for the speaker and the topic of the lecture. This method helps to solve some problems related to OOV words, as in case of the other systems such as TalkMiner (Adcock et al., 2010). The browser interface provides the user with the whole lecture transcript with accompanying video, segmented using the MCut algorithm (Malioutov and Barzilay, 2006). A screenshot of the system search is shown in

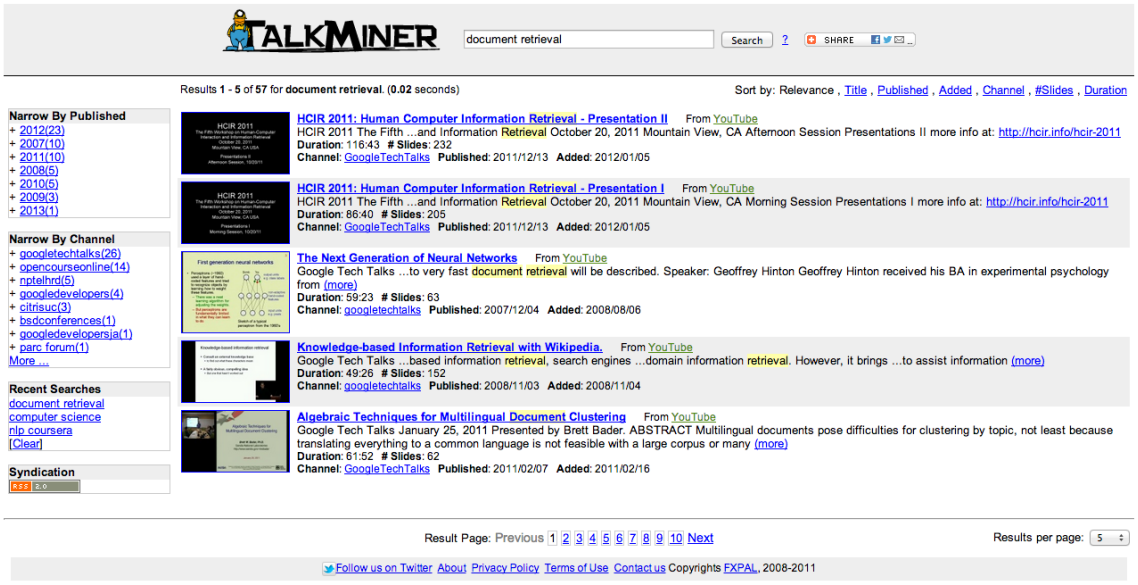


Figure 3.2: Screenshot of a search result for the query document retrieval in TalkMiner.

Figure 3.1 taken from (Malioutov and Barzilay, 2006). In this system the user is required to listen to different parts of the lecture in order to find relevant content, making information access process potentially very unefficient.

A system such as TalkMiner regards the problem of SCR for lectures as the task of detecting distinct slide images. Optical character recognition (OCR) is used to recognize the contents of the slides in order to create metadata at jump-in points for the relevant information assuming that the slides are indicative of relevant parts of the lecture (Adcock et al., 2010). Here the jump-in points are considered to be at the start of the projection of each slide, as the start of slide projection is assumed to indicate a switch to a new topic or a sub-topic, and thus this part of the transcript will be topically coherent. Within this system video lectures originate from video-sharing platforms such as YouTube, Webcast Berkeley, PARC Forum and blip.tv. In processing a video, first the distinct slides are located within the video keyframes using ProjectorBox technology (Denoue et al., 2005), and the OCR of the text on the slide is assigned to the time of the slide keyframe, this text can be indexed for use with traditional text retrieval methods. The TalkMiner search interface is built as a typical search browser with an input field for a query, and an output page with

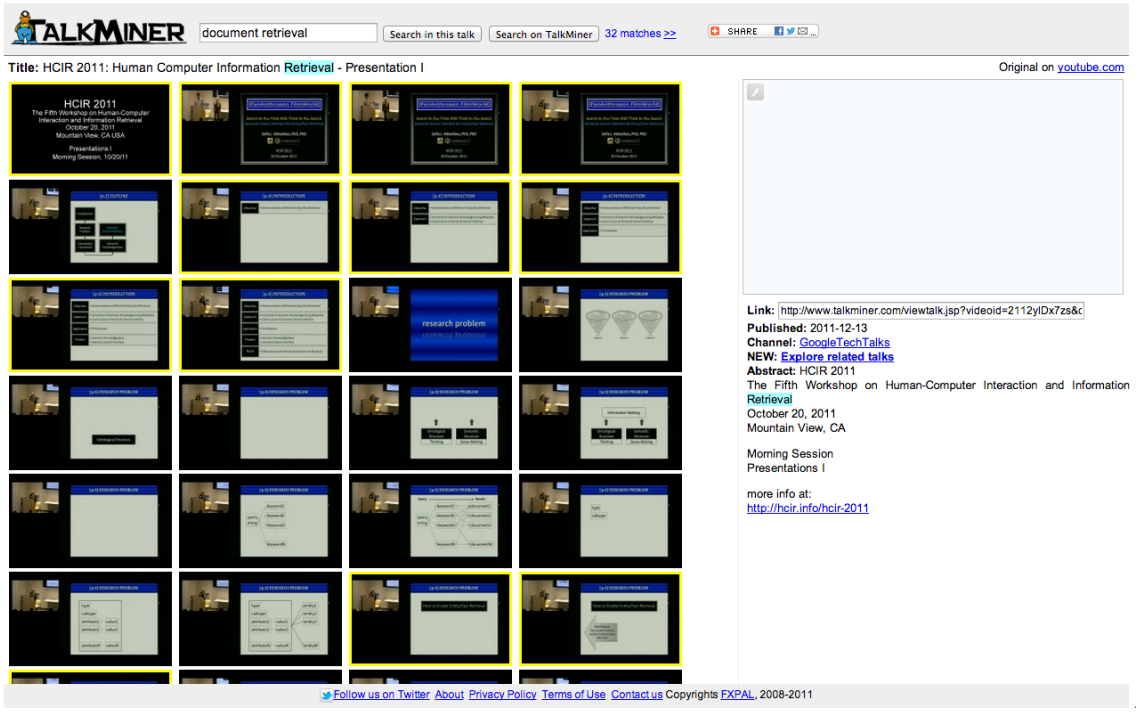


Figure 3.3: Screenshot of slides viewing in TalkMiner.

the list of potential result videos, an example screeshot is shown in Figure 3.2. The user can select one of the videos for playback, and then choose the actual section with the slide of interest, see Figure 3.3 for a screenshot example.

Lectures retrieval in the SpokenDoc and SpokenDoc-2 tasks at NTCIR-9 and NTCIR-10

Lecture retrieval has been formally introduced as a research task at the NTCIR benchmark initiative. Evaluation task was based on the Corpus of Spontaneous Japanese (CSJ) that contains lectures recordings (Maekawa et al., 2000). The task organisers augmented this lecture collection with sets of queries and corresponding relevance assessment within the CSJ lectures for the NTCIR-9 task (Akiba et al., 2011). The task contained options for full lecture or passage retrieval on the same set of lectures, while the NTCIR-10 task provided the Corpus of Spoken Document Processing Workshop (SDPWS), a smaller collection of more informal recording of talks given at the 1-6 annual Spoken Document Processing Workshops for the passage retrieval sub-task (Akiba et al., 2013). For our investigation we use the

dataset of SpokenDoc task at NTCIR-9. Further details of the test set corpus are given in Chapter 4.

The NTCIR SpokenDoc tasks introduced several new metrics inspired by MAP to evaluate search effectiveness of the task, but adjusted for passage retrieval task (Akiba et al., 2011). We give a detailed description of these metrics later in Chapter 5 of this thesis, in order to compare them with our own proposed new SCR evaluation metrics.

Overall the initial results of the participants showed the feasibility and at the same time the complexity of the task, since none of them achieved high results, even with the use of manual transcripts. The use of overlapping fixed length segments scoring with subsegment editing of the result list achieved the best scores amongst the submitted runs (Akiba et al., 2011). The fact that this data contains a set of manually selected ad-hoc queries and relevance assessments, and thus represents a challenging lecture retrieval task is the reason to use it within our experiments reported in this thesis.

Discussion on SCR for lectures

The lecture browsers allow user-friendly interactive exploration of retrieved results, however the retrieval strategies have certain limitations. The segmentation of lectures can be either manual, making it impossible to generalize them over the vast archives of data which are becoming available online (Lee and Lee, 2008). Otherwise the segmentation may rely on the number of segments in the audio data being provided to the system as an input parameter due to the features of the segmentation method being used (Glass et al., 2007), thus making it less flexible for capturing variation in the number of topics in a transcript of a real lecture. When these browsing systems have been developed, the focus has been on providing functionality for content playback rather than measuring and optimising retrieval effectiveness itself, and the efficiency of the interaction and satisfaction of user information needs.

3.2.2 Search of Recorded Meetings

On the scale of naturalness of speech, meetings are an example of more conversationally free content style that combines informally prepared speech and discussions, and as such are often a more challenging SCR than lecture search. Thus, the introduction of meetings as a research topic in SCR poses new questions and challenges compared to broadcast news and lectures. These issues include segmentation of dialogues (Dielmann and Renals, 2007), and use of additional data, e.g. minutes can be added to the sources of metadata together with slides (Jones et al., 1996).

Potential retrieval scenario

Traditionally one participant in a formal meeting is assigned to take minutes which summarize the activities and conclusions of the meeting, in the case of more informal meetings there is often no record of the proceedings. Even when taken, minutes often record only the key elements of the discussions and decisions reached, as understood at the time of the meeting by the person taking the minutes. Thus, minutes may be deficient or inaccurate if the minute taker misunderstands some elements of the discussion or the future significance of some part of the meeting is not apparent to the participants and no record is kept. If in the future someone wants to know the process by which a decision was made or how a particular idea arose, they may find that this information is missing from the minutes. However, if recordings have been made of the meetings, participants and others can potentially play back parts of a meeting to access specific information. But in order for this to be effective, an efficient mechanism for locating relevant sections must exist.

One of the concerns for research in meeting search is identifying a scenario motivated by real-life information needs. Unlike the lecture case it is less clear what exactly the users of such a system might be looking for. Possibilities might include locating jump-in points to places where discussion about a certain topic started or a decision point was reached, the opinion of a certain person or any person assigned with a certain role, or all the discussions on a topic in series of meetings. The first

option involves topic segmentation, discourse analysis, and can be a known-item or ad-hoc task. Search for a spoken snippet of a certain speaker might be more a problem of acoustic search. The situation when all of the mentions of a certain topic are importent and have to be found is a recall-oriented task.

Unfortunately, privacy issues related to the content of meetings generally prevent actual data from real meetings being made available as public datasets for research purposes. Therefore research in meeting search is mostly concentrated around meetings recorded in laboratories, such meetings are typically based around predefined scenarios with actors taking the role of meeting participants (Morgan et al., 2003; Renals et al., 2008). The artificiality of this setting is not reflected in any unnaturalness of the speech itself, because meeting participants get involved in the discussion according to the roles they are assigned in the meeting, and produce utterances on the fly in a linguistically and acoustically natural way (Renals et al., 2008).

As in the case of lectures, the research focus in meeting search has mostly been on developing data, improving performance of speech recognition systems on this type of data (Morgan et al., 2003), segmentation algorithms (Hsueh and Moore, 2006), and browsing software (Renals et al., 2008), and has not really focused on proper retrieval experiments.

Among this existing work, the ICSI project on processing speech from meetings created one of the first widely available meeting speech collections (Morgan et al., 2003). However it included only audio and transcripts of natural meetings, therefore lacking one of the natural assets of meetings - the availability of additional data from slides, minutes etc. Studies using this collection have concentrated on speech segmentation, speaker activity detection, and dialog act annotation.

The AMI and AMIDA projects (Renals et al., 2007, 2008) coordinated creation of the AMI meeting corpus consisting of very carefully collected and documented meeting recordings including slides and minutes, and videos of the meetings. More than 70% of the data is artificial, based on meetings carried out according to a scenario, the rest are recordings of research meetings where the creation of the AMI

corpus is discussed. In the former situation the participants were assigned certain roles and were asked to act accordingly in the recorded discussions. This might have caused certain differences from the way in which real life participants of these conversations might have behaved, although the audio itself was natural and spontaneous as in any other meeting. Since they are role-playing, the participants were not engaging in their real employment roles and there was no history of the issues being discussed or future implications of any decisions made. Also the participants did not know each other in the roles that they are taking. Thus the value and significance that would be associated with decisions in a real working environment were not present, and the participants did not perhaps engage in the same way that individuals working with each other regularly on a long term basis might. However, since the scenario-based meetings were recorded over four sessions (4 separate meetings), each group of participants developed at least some history of interactions in their assigned roles over the course of the meetings. These resulted in natural speech behaviour including monologues, dialogues and multiparty discussions.

Research on the spoken corpora collected in the AMI and AMIDA projects has focused on the development of effective ASR tools for this data, and on means to provide structure to the content by automatically segmenting the content into topical segments, assigning topical labels and summarisation of meetings (Renals et al., 2007).

Retrieval in meetings

While searching these types of corpora is obviously an interesting research topic, no query set describing specific user information needs was provided with these collections, and therefore no retrieval study has appeared for these collections.

In reality companies that record their meetings are likely to be interested in at least browsing facilities, with potential for search in a standard interactive manner by users posing queries to find relevant segmented units. Therefore several systems have been developed that help to structure a meeting recording and allow its further reuse.

The CALO Meeting Assistant (CALO-MA) project is a good example of a browsing system that structures meetings in order to prepare a browser representation of the data with segmentation into topic and dialog acts, and summarization (Tür et al., 2008).

One application of search on meeting data is provided in Chibelushi and Thelwall (2009), which examines the mining of meeting transcripts to identify elements within the meeting associated with points related to the making of key decisions. This application is an example of the situation where high recall is required. In this case in order to ensure that all relevant information pertaining to the taking of a decision has been located, *all* relevant content needs to be identified.

Another alternative to standard search, as introduced by Popescu-Belis et al. (2009), is *query-free* or *just-in-time* retrieval (the AMIDA Automatic Content Linking Device). This system transcribes an ongoing meeting automatically, and uses the transcript to perform searches of previous meetings and additional material at regular intervals. The Ambient Spotlight application follows the same scenario, however it searches for relevant documents not within the meeting dataset, but amongst all possible relevant documents on the user's machine (Kilgour et al., 2010).

Summary on SCR for meetings

As discussed above, recording of meetings is becoming a widespread practice in companies, but content management applications that are becoming available for these collections involve only functionality for browsing or summarisation. There is a lack of investigation of retrieval systems and their effectiveness for this data. For the purposes of research into techniques for retrieval from meeting recordings, retrieving all parts of the meetings where certain topics were discussed is a reasonable case to be analysed.

3.2.3 Search of Interview and Internet Television

Lectures and meetings often contain regions where the conversation is free and unstructured as in casual conversations. The next step on this path towards unstructured data is search over recordings of proper conversations. Datasets that comprise only interviews or ones that are created by different semi- and non-professional users over the Internet are examples of more informal unstructured recordings.

Interview retrieval at CLEF

In 2005-2007 the Cross-Language Evaluation Forum (CLEF) Cross-Language Speech Retrieval (CL-SR) task focused on retrieval of conversational speech which lacked clear topic boundaries (White et al., 2005). This data consists of interviews with survivors and witnesses of the Holocaust from the Shoah Visual History Foundation collection. The MALACH project focused on Multilingual Access to Large Spoken Archives for this collection using large vocabulary continuous speech recognition (LVCSR) systems trained to produce transcripts for this data (Byrne et al., 2004). For the CLEF CL-SR task interviews were manually divided into meaningful segments, and were augmented by including a number of pieces of manually and automatically generated metadata for each "document" unit which provided additional description of the spoken content. This metadata included manually and automatically assigned keywords and short manually prepared expert summaries for each document.

In order to minimize the effort required from task participants, the CL-SR organizers initially adopted a known-boundary item condition in a standard ad-hoc retrieval setting for English interviews (manual segmentation was provided with the data made available to task participants), and MAP was used for evaluation of participants submissions (White et al., 2005). An unknown-boundary condition was later introduced for the subtask on this dataset for Czech interviews (Oard et al., 2006; Pecina et al., 2007). This new unknown-boundary condition task required the introduction of a new metric for result assessment that would reward systems

that identified the closest start point for the playback of relevant content – mean Generalized Average Precision (mGAP) (Liu and Oard, 2006).

The implementation of this score for speech data is described in (Pecina et al., 2007). It measures the error in finding the start point in time of relevant content within items by the retrieval system. The calculation of GAP for a single query is shown in Equation 3.3.

$$GAP = \frac{1}{n}.\sum_{r=1}^{N} P[r] \cdot \left(1 - \frac{Distance}{Granularity} \cdot 0.1\right) \tag{3.3}$$

where N is the number of ranks, $P[r]$ is the precision at rank r, $Distance$ is the distance between the start of the segment and the beginning of relevant part (the limit was set to 150 seconds by task organizers), and $Granularity$ is the step that is used for the penalty function ($Granularity = 15$ seconds at CLEF). Thus segments that make the user wait for longer than 150 seconds are not considered relevant. This metric seemed reasonable for reflecting the use case scenario, however, it does not reflect the time the user will spend on listening to non relevant content in the retrieved list.

Experiments in the CLEF CL-SR tasks showed that even with ongoing improvements in speech recognition accuracy, speech search for more complex and informal speech sources such as this still presents significant challenges. It was found that retrieval for this task using ASR transcripts is enhanced greatly by including manually generated metadata in the search index and improved marginally in comparison to ASR transcript only retrieval by including the automatically generated metadata. In this case the spoken content itself may not be sufficient to enable effective search without augmentation with additional metadata, since the speakers may discuss topics without providing details such as names of people or places required to support effective retrieval.

Internet video content

With the development of Internet sharing and streaming platforms like `YouTube.com` and the growing availability of mobile devices that allow recording of audio and video for further online distribution to a general audience, effective search through these collections becomes harder because of the volume and diversity of the created content.

This informal content that is produced by humans in the freedom of their own recording facilities poses new challenges to SCR technology. For example, the topics in a collection may no longer be coherent, since users on Internet sharing platforms may not share any common interests, and therefore their videos will be dedicated to completely different stories and events. The level of education of the content creators and the quality of their recording devices is not controlled within these new collections that are growing in size exponentially.

Moreover, this type of content will often have varying quality of recording, with variations in topics and styles of video setup which pose further challenges to an ASR system that might be used to attempt to create a transcript for the data. On the other hand, this data is rarely created by itself, as it generally is accompanied by metadata that the users who upload the video assign to it, in addition the audience who interacts with it may add comments and additional information that can potentially be helpful in a retrieval process. However, since there are so many content providers on the Internet, and there is no general agreement between them about features or quality of uploaded video presentations or the style of their content, and additional metadata information, much of this metadata can be considered informational noise in which will have little positive impact on retrieval.

Previously introduced methods for SCR cannot cope with these issues, because these technologies make no attempt to take this diversity of descriptive information into account. Even more importantly, the availability of the techniques to record, upload and store the content raises the expectation of potential SCR users that such systems should go beyond simple factual information needs. Systems may

no longer be limited to factual information of news or interviews about a certain event or place, aspects of emotions of the speakers, their intent, distinct snippets or highlights of their speech may be of interest (Larson et al., 2012). These issues require a deep understanding of the challenges and effective techniques for SCR on informal multimedia.

Currently new benchmark datasets are being collected from the Internet for evaluation of different aspects of multimedia retrieval, involving separate video and speech analysis or their combination[1]. This Internet data is created by semi-professionals and varies in audio and video quality, in organization of the material, in style, and in the amount and quality of the metadata being added. Thus we use it within our research experiments. Further description of this data, as used in our experimental investigation, will be given in Section 4.3.

3.3 SCR: Summary and discussion

In this chapter we have overviewed different types of human interactions with spoken content. The content varies in complexity and becomes more unstructured as it becomes more informal. Initial research in SCR evolved with the availability of improved speech recognition. This enabled research on more complex speech signals as the quality of ASR systems improved further. We examined various SCR use case scenarios associated with different types of data, including search for whole news stories, identification of the start of discussion on a relevant topic, and recall oriented retrieval of all conversations related to a topic.

We observed that data such as meetings, although often recorded, is not widely available to the research community due to privacy issues, and that research meeting collections created in laboratories do not contain the elements needed for SCR research, i.e. queries and relevance data. They thus need to be extended to form retrieval test collections with query sets and relevant information. Therefore devel-

[1] www.multimediaeval.org

opment of test collections has to be part of the work in SCR research.

News articles are designed to be self-describing, so that the listener understands the background of the story. Thus they are generally easy to retrieve, since they contain a lot of domain-specific contextually relevant words. Lectures may be less adequate in this regard, but speakers are generally explaining, and again will tend to include context specific words to ensure that as far as possible their audience follow their arguments. Similar to lectures, meetings can be accompanied with additional sources of information, e.g. slides. Though as meetings happen within a single working environment and might cover one topic over the span of several meetings, important content words may be absent from these conversations. Internet TV videos have less structure, and it is harder to generalize principles for their organisation and pattern of words usage due to their high variability.

Although a large amount of research has been focused on the browsing of collections, formal comparison of retrieval results has rarely been within the scope of the investigation for data types such as meetings or Internet TV. Since this data represents much potential interest for SCR applications, we focus our research on the development of effective retrieval methods for this informal conversational content.

Previous research has highlighted two aspects of SCR that need to be taken into account when working with spoken content: the effect of ASR errors on retrieval behaviour, and the methods available to segment the transcript into suitable retrieval units. The former decreases the possibility of retrieval of relevant documents. The latter directly influences the user experience because the boundaries of the document segments are the positions where they start playback, and decrease the retrieval behaviour as choice or accuracy of poor segmentation methods will give poor retrieval results. In the following sections we summarize the findings on the influence of the ASR errors and transcript segmentation for all types of data, and discuss our research proposals.

3.3.1 ASR WER and SCR

ASR is an important component for most speech retrieval applications, since the cost of manual transcription means that in practice speech retrieval must generally rely on automated indexing methods. An important question with respect to speech retrieval research is the effect on search behaviour of the errors made by the ASR system.

State-of-the-art findings

Using a general purpose ASR system generally produces speech transcripts which are inferior to those that can be created using an ASR which has been trained for the specific recognition task to be undertaken (Byrne et al., 2004; Renals et al., 2008). The basic relationship between average transcription quality and accurate (or near accurate) manual transcriptions is reported in most speech retrieval studies. For example, the TREC SDR track illustrated how the relatively low average recognition error rates on the radio and TV news material used for these studies resulted in little loss in retrieval effectiveness as measured by MAP for a news retrieval task compared to still errorful, but much more accurate manual transcripts (Garofolo et al., 2000a).

An interesting and careful examination of the differences in retrieval behaviour of documents with different speech transcript accuracy levels for the results of the TREC-7 SDR task is described in (Shou et al., 2003; Sanderson and Shou, 2007). The analysis of the distribution of the error rates in the ranked lists retrieved for topics in this task shows a general tendency for documents with low WERs to be retrieved at higher ranks, independent of document relevance to the search query. A natural consequence of this observation is that relevant items with high WERs could be expected to be retrieved at low ranks or not retrieved at all, although the extent to which this occurs and the effect of ASR errors on the recall in ranked lists was not explored in this work, or in any other work of which we are aware.

The impact of the errors according to their types was measured using different quality metrics on the level of the document or on the level of the whole collec-

tion. Such metrics as Named Entity WER and Named Entity Mean Story WER for Cross-Recognizer Results showed the best correlation with retrieval performance (Garofolo et al., 1999). The global semantic distortion metric based on the vector space model and focusing on various types of substitutions (frequent vs infrequent, semantically similar vs dissimilar) revealed a higher impact of infrequent and semantically dissimilar substitution errors on retrieval behaviour (Larson et al., 2009).

New targets

We are interested in further exploring the relationship between ASR accuracy and the retrieval behaviour in speech search for tasks. For example, non-relevant documents will have similar variations in WER to relevant ones, and this factor can be expected to interact with query-document matching scores to affect the rank of both relevant and non-relevant content. Thus, it can be expected that non-relevant content with high transcript WER may be ranked lower than that with a lower WER. Thus, changes in rank position between accurate and noisy transcripts will be subject to a range of interacting factors. This aspect of retrieval has not previously been examined in detail.

In addition, we are not aware of any existing studies which have explored speech retrieval from the perspective of a recall-focused task.

3.3.2 Transcript segmentation and SCR

As discussed in Chapter 2 and the current chapter, informal conversational content represents a challenge for SCR since it does not have prior topical segmentation. Each SCR task requires tests to determine which method achieves better results. segmentation methods for SCR may be borrowed directly from existing text segmentation research or adapted using spoken content features (Tür and De Mori, 2011).

We are interested in comparison of the performance of different techniques on large scale unstructured conversational spoken content, such as lectures, meetings,

and video content uploaded to the Internet. In order to draw meaningful conclusions and talk about potential solutions, we focus not only on the experiments themselves, but on the development of evaluation metrics that allow us to differentiate the effect on the SCR results.

3.3.3 Summary

In the following chapter we introduce different datasets that were available or created as part of work on this thesis. We follow the same logic of increasing conversational nature of the content. The description contains the details of ASR quality, and available relevance assessments. Chapter 5 then introduces novel metrics designed to enable us to evaluate SCR results for more challenging unstructured content and improve our undestanding of difference between achieved results. The rest of the thesis shows extensive analysis of SCR performance and ways to improve its effectiveness.

Chapter 4

Test Collections for SCR Research

In Chapter 3 we overviewed different types of unstructured conversational content that pose challenges to SCR technology. In general people are becoming used to digital recordings as a natural part of their life, for example when studying, in their working environment and as part of their entertainment and social experience. This motivates us to consider all these types of data to fall within the scope of our investigation of SCR technologies. In this chapter we give details of two datasets in English and Japanese that are provided for existing SCR benchmarks, and another SCR research test collection in English created within our research.

First we describe the corpus based on Japanese lectures provided within the NTCIR-9 benchmark (Akiba et al., 2011). We next introduce an SCR test retrieval collection that we created on the basis of the openly available AMI corpus (Carletta, 2007). As part of this description we outline the choice of queries and manual relevance assessment procedure used to complete this test collection. As an example of the least controlled recording environment and with the greatest freedom of topical content, we use data from the Internet sharing platform blipTv[1]. These queries were collected using crowdsourcing procedures for further use at MediaEval benchmark[2]. We overview our experience of retrieval collection creation for this benchmark test collection using real users and content creators via Internet.

[1]http://www.blip.tv
[2]http://multimediaeval.org

Type of content	Speakers	Lectures	Data size (hours)	Number of Inter-Pausal Units (IPUs)
Academic presentations (AP)	819	987	274.4	359 098
Simulated public speech (SPS)	594	1715	529.9	486 430

Table 4.1: Overview of the CSJ Target Document Collection for the SpokenDoc task at NTCIR-9.

4.1 Lectures: Corpus of Spontaneous Japanese (CSJ)

For the exploration of SCR for lectures we used the test collection developed for the SpokenDoc task at NTCIR-9 (Akiba et al., 2011). This is based on the Corpus of Spontaneous Japanese (CSJ) (Maekawa et al., 2000) released by the National Institute for Japanese Language. The CSJ consists of a variety of content sources. The two main sources are academic presentations (AP) and simulated public speech (SPS). The APs are live recordings of academic presentations from 9 different academic societies covering the fields of engineering, social sciences and humanities. The SPS content consists of studio recordings of paid layperson speakers; each of their speeches is about 10-12 minutes in length. These speeches are on everyday topics such as "the most delightful/saddest memory of my life", presented in front of a small audience in a relatively relaxed atmosphere. The AP and SLS content is used in combination as the SpokenDoc test dataset. This results in a total of 2702 'lectures'. Summary details of this data are shown in Table 4.1.

The speech signal was recorded using a head-worn close-talking microphone. Each lecture was segmented by the corpus developers using pauses that are larger than 200 ms. These segments are called Inter-Pausal Units (IPUs). An IPU is short enough to be used as an alternative to a specific position in a lecture. Therefore, the IPUs are used as the basic units to describe the relevant passages in the SpokenDoc retrieval task (Akiba et al., 2011).

Transcrips	W.Corr.	W.Acc.	S.Corr.	S.Acc.
REF-WORD	74.1	69.2	83.0	78.1
REF-SYLLABLE	—	—	80.5	73.3

Table 4.2: ASR performances [%].

In the following subsections we describe the details of manual and ASR transcripts provided by the task organisers, and the search task defined on this dataset. To illustrate the features of this dataset, we give examples of queries and relevant content to be retrieved.

4.1.1 Transcript Types

Manual Transcripts

The lectures in the CSJ are provided with manual 'orthographic' and 'phonetic' transcripts for use as an error free baseline in ASR and other research. The orthographic transcription is a mixture of Kanji (Chinese logograph) and Kana (Japanese syllabary) in the manner of ordinary Japanese writing. Phonetic transcription uses only Kana characters to represent the phonetic details of the speech as accurately as possible within the limits of syllabary. Various tags are included in the transcripts to mark spontaneous speech phenomena such as filled pauses, word fragments, mispronunciation, etc. In addition to these tags non-speech effects, such as laughter and coughing, are also marked.

Automatic Transcripts

A baseline ASR transcript of the spoken data collection is provided to the SpokenDoc task participants. Two types of transcripts were included: using word-level and syllable-level ASR (syllable-level transcripts are appropriate since Japanese is a syllabic language).

The baseline transcripts were created using the Julius ASR system combining acoustic and language models (Lee and Kawahara, 2009). All speech was divided into two groups, two ASR systems were trained on these subsets, and used to recognize

SDPWG-TA1001-01
情報検索性能を評価するにはどのような方法があるか知りたい。
"How can we evaluate the performance of information retrieval?"

Figure 4.1: Example of the natural language query NTCIR-9 task

the other subset. The word-based language models were trigrams with a vocabulary of 27,000 words, while syllable-based trigram models were trained using the syllable sequences of the training group. The Julius decoder (Lee and Kawahara, 2009) with a dictionary containing the above vocabulary was used to perform the word-level ASR. All words registered in the dictionary appeared in both training sets. N-best speech recognition results were obtained for all spoken content.

Table 4.2 shows the word-based correct rate ("W.Corr.") and accuracy ("W.Acc.") and the syllable-based correct rate ("S.Corr.") and accuracy ("S.Acc.") for these reference transcripts.

4.1.2 Task Definition

In order to complete the NTCIR-9 SpokenDoc test collection a set of query topics and corresponding relevance judgments indicating which sections of the lectures were relevant to each query were developed for the CSJ spoken corpus.

Two retrieval subtasks were defined for the NTCIR-9 SpokenDoc task using the same spoken dataset and search query set. The two tasks differed in the unit of the target document to be retrieved, and were defined as follows:

- Lecture retrieval: Find the lectures that include the information relevant to the query topic.
- Passage retrieval: Find the passages within lectures that exactly include the information described by the given query topic. A passage is an IPU sequence of arbitrary length within a lecture.

Within this thesis we are interested in the more challenging passage retrieval task scenario.

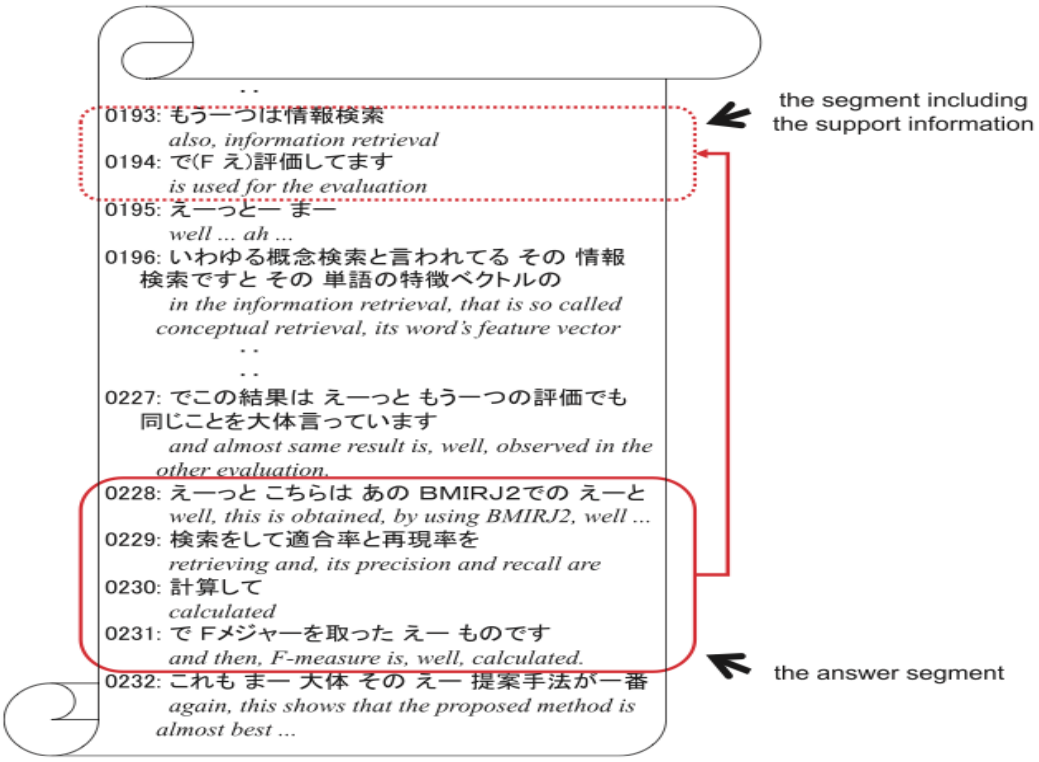


Figure 4.2: Example of the relevant answer to a query and of its supporting segment in NTCIR-9 task

Query Set

Information needs for spoken lectures could potentially relate to the contents of a whole lecture or only part of it. The SpokenDoc task focused on the latter type since it was felt that this will be more likely than the former in terms of practical lecture search scenarios.

The test set consists of 86 queries created by 5 developers. It was observed that the constructed queries tend to be less like a query in a standard ad hoc document retrieval task, and more like a question submitted to a question answering system. A query topic is represented by a natural language written sentence as shown in the example in Figure 4.1, and the list of all the queries is provided in Appendix B.

Relevance Assessment

Relevance judgments for the queries were performed manually after the submission of results by the task participants. As in TREC SDR tasks, the procedure

of pooling was used to carry out the evaluation. One of the difficulties relating to the relevance judgments came from the treatment of the supporting information. A passage was regarded as non-relevant to a given query even if it was a correct answer to the query in itself, if it had no supporting information that would convince the searcher who submitted the query of the correctness of the answer. For example, the query "How can we evaluate the performance of information retrieval?" The answer "F-measure" is not sufficient, because it does not say by itself that it really is an evaluation measure for IR. The relevant passage must also include supporting information indicating that "F-measure" is one of the evaluation metrics used for information retrieval. Figure 4.2 shows an example of an answer and its supporting information for the query "How can we evaluate the performance of information retrieval?" (Akiba et al., 2009).

As shown in Figure 4.2, the supporting information does not always appear together with the relevant passage, but may be somewhere else in the same lecture. Therefore, a passage was regarded as relevant to a given query if it had some supporting information in some segment of the same lecture. If a passage in a lecture was judged relevant, the range of the passage and the ranges of the supporting segments, if any, along with the lecture ID, were recorded in the "golden" file listing relevant content for each query.

For each query, one assessor, its constructor, searched for its relevant passages and judged their degrees of relevance. The assessor classified them according to the degree of their relevance: "Relevant," "Partially relevant," and "Non relevant."

As the retrieval task setup allowed submissions of the runs using word-based or syllable-based transcripts, there were 2 sources of the sets of potentially relevant documents to be assessed. In both cases assessment was conducted against the manual transcript of the target document collection. The relevance assessments were mapped to the IPUs to enable relevance to be measured at the level of IPUs.

4.2 Meetings: Searching the AMI Corpus

Experimental investigation and evaluation of IR applications requires a suitable search collection. For meetings search this must include a rich and well organised dataset of recorded meetings. As we discussed in Section 3.2.2, construction of such a meeting collection is a complex and expensive process requiring planning of the meetings which will be recorded, but also if an analysis is to be made of the impact of speech recognition errors on retrieval effectiveness, a full accurate manual transcript of the meetings. Obtaining such a meeting collection and forming an accurate manual transcript of speech recordings is a very expensive process. Considering these factors the investigation of meeting search for this thesis was carried out using the AMI corpus, collected as part of the AMI project and made publicly available for research purposes (Carletta, 2007). In this section we outline the construction of the AMI corpus itself, and the work we carried out to create a retrieval test collection.

4.2.1 The AMI Corpus Details

The AMI corpus[3] contains 100 hours of annotated recordings of planned meetings (Carletta, 2007). Meetings last about 30 minutes each, 70% of them simulate a project meeting on product design, with the other 30% being meetings that happened in the laboratories working on the ASR for the corpus itself. Meetings usually involve 4 participants, and were recorded using 6 cameras and 12 microphones: 1 headset microphone for each speaker, and an 8-element circular microphone array. For the majority of the meetings, both manual and automatic transcripts are provided, for the latter the developers of the corpus created an ASR system which makes use of a standard ASR framework employing hidden Markov model (HMM) based acoustic modeling and n-gram based language models (LMs) (Renals et al., 2007). The dataset also includes additional materials including slides projected during the meetings. In this study we use the AMI release 1.4 (automatic and manual tran-

[3]http://www.amiproject.org/

scripts) and release 1.5 (automatic segmentation results for ASR transcripts (Hain et al., 2012)).

The meeting transcripts in the AMI corpus are published separately for each speaker participating in the meeting. Since we were interested to explore search of the meeting, rather than information relating to each speaker, we first merged the per speaker transcripts using the time marking data provided in the corpus to form a single transcript file for each meeting. We omitted incompletely transcribed meetings, and used only the fully transcribed ones in this study, since we wished to work with only complete meeting transcripts. This gave us a total of 160 meetings for our experiments.

For the purpose of comparison between manual and ASR transcripts, word recognition rate (WRR) was calculated as the proportion of non stopwords in the manual transcripts recognised correctly in the ASR transcripts. As retrieval systems usually index stemmed versions of words, we also calculated the WRR after running a publicly available implementation of the Porter stemming algorithm[4] on the transcript (Porter, 1980). Comparison of the transcripts showed that the automatic transcripts have accuracy ranging between 45–85%, with an average of 72% of the manual transcripts. The stemming of the words did not significantly affect the average WRR across the corpus (73%), however it increased the WRR for some of the files making the range between 47–92%. It can be observed that the recognition rate often varies across the length of meeting transcripts, sometimes dropping much below 45% for some reasonably long parts of a transcript. While MAP has generally been shown to be robust with WRRs as low as 70%, WRRs lower than that have been shown to impact retrieval accuracy (Garofolo et al., 2000a). Thus we can note that while sections of the ASR transcripts are likely to be retrieved with good reliability, assuming that they match the content of search queries well and are sufficiently selective to distinguish them from other ones, other regions of the transcript are likely to present significant challenges for retrieval.

[4]http://tartarus.org/ martin/PorterStemmer/python.txt

<top>
<num> Number: 21
<title>
Evaluation
Criteria rating with seven-point scale
Shape biomorphic 1
Size small 2
Color bright warm 1
Feel as soft as possible 3
Functionality
people won't use it before they buy it paradoxically other features will be main selling points 5
<top>

Figure 4.3: Example of information present in a slide, AMI Corpus (query 21).

4.2.2 Retrieval Task from within the meetings data

For our investigation of meeting search we assume the scenario of a meeting participant who wants to find all locations in meetings where the topic of a PowerPoint slide was discussed regardless of whether it was projected at that time or not.

The PowerPoint slides provided with the AMI corpus were used in multiple meetings thus creating different instances of conversations by different participants about the same topic. The topic on the projected slide could also have been discussed at other points in this or other meetings while the slide was not actually being projected, leading to other conversations about the same topics possibly by the same people.

Query set

For our investigation, we took a subset of 35 of the PowerPoint slides provided with the AMI corpus as a topic set based on the following criteria: the text of the slide should be reasonably detailed (more than 15 content words), diverse in structure (lists of actions, sentences describing work to be done), diverse in situation of use (beginning of the meeting or closing), have a different number of possible relevant documents (from uniquely used ones as in a known-item search, to slides that were

used in almost every set of 4 meetings). Figure 4.3 shows an example of the contents of a slide used as search topic. We split this set into 10 and 25 queries for the development and test set respectively, they are all listed in Appendix C.

The search task was to retrieve *all* segments relevant to the topic being discussed in the slide. It thus represents a recall-focused search task which aims to support meeting participants looking to find all discussed material relevant to each query slide, i.e. missing even one instance of the relevant content is considered as failure to provide the user with the requested information, because any individual relevant segment may be the one that the user is looking for, and finding only some instances of relevant information being mentioned at points in meetings may not be enough to fulfill the task goal if the target information is not among that retrieved. This target relevant material may be taken from discussions by the same participants, or by participants in another discussion examining topically related issues.

Relevance Assessment

In order to carry out our search experiments, corresponding manual relevance assessments identifying the relevant content for each slide topic was generated using a pooling procedure. As it would have been impractical and overly time consuming to create the pooling union based on the documents containing the whole meetings and to look for parts of the relevant documents in them, we started with segmentation of the content using varying strategies, and then carried out the retrieval on those versions of test collection. These initial runs that were created to collect the union for the pooling assessment had to be representative to our further experiments, therefore we used varying segmentation methods that correspond to the approaches that we are interested to explore. In next subsections we describe how we ran various segmentation methods on the collection, carried out the retrieval procedure, and these retrieval results were used within our pooling basis. A special tool was written in Java that mapped retrieval results from different runs back to the manual transcript of the initial document and highlighted the portions that needed to be

checked manually for relevance. All transcript words have time stamps, therefore once the relevant segment is defined it can be used for the assessment of any other run with potentially different segment boundaries, as long as the time information is preserved for the new runs. In the rest of this section we outline the segmentation methods used for creation of initial results that can be further assessed using pooling procedure, previously used for broadcast news SCR results assessment, as mentioned in Section 3.1.

Segmentation The AMI collection as provided already contains manually created topic segmentations of the transcript. Topics and subtopics form a hierarchical structure, where labels have been assigned by annotators choosing tags from a list of suggestions. This topic segmentation was made based on the manual transcripts, but does not cover all of the meetings in the dataset. These segments are provided for only for a subset of 139 out of the total of 173 meetings. Since our goal is to investigate the impact of the segmentation of spoken material on retrieval results where no manual segmentations are provided, we decided to automatically segment the AMI meeting transcripts ourselves. We segmented the manual transcript using simple time- or length-based methods, and content-based algorithms. For the content-based segmentation we used Choi's popular C99 algorithm (Choi, 2000) and Hearst's TextTiling algorithm (Hearst, 1997), Minimum Cut (Malioutov and Barzilay, 2006), and the method of Hsueh and Moore (Hsueh and Moore, 2006). All these methods work on the level of sentences. However current ASR systems do not by default provide punctuation in their transcript output. Thus we needed to use pseudo-sentences of reasonable length for this collection. The length value was calculated as the average length of the sentences in the manual transcript.

Retrieval The segments obtained using each segmentation technique from the manual transcripts were indexed for search using a version of SMART information retrieval system[5] extended to use language modelling (a multinomial model

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

with Jelinek-Mercer smoothing) with a uniform document[6] prior probability (Hiemstra, 2001), as introduced in equation 2.3 in Section 2.1.2. The retrieval model used $\lambda_i = 0.3$ for all q_i, the value being optimized on the TREC-8 ad-hoc retrieval dataset. Stopwords were removed using the standard SMART stopword list, and the remaining content words stemmed using a variant of the Lovins stemmer (Lovins, 1968) which is packaged in SMART by default.

Pooling procedure To carry out the relevance assessments, the following pooling procedure was adopted.

1. Retrieval runs were carried out for each topic using segments created using the different segmentation schemes.
2. The top 50 retrieved results for each run were collected and compiled into a pool for each of the topics. (We chose the number of top ranked documents to be assessed empirically, assuming that it is reasonable to expect the user in a real case scenario to try to browse through this number of retrieved documents.)
3. An interactive application was developed which highlighted the union of the retrieved segments in the original documents they belong to, i.e. if one meeting had several different segments in the pool, the whole area between the beginning of the first of the segments and the end of the last of the segments was highlighted for assessment.
4. Relevant regions between the beginning of the first segment of each segment group and the end of the last one were marked manually by an assessor.
5. After defining the relevant region for the manual transcript for each topic, this information was projected onto each segment unit based on time correspondence in order to create individual relevance files for each of the segmentation techniques. Thus for each segment in each segment set, we know the beginning and end points of all assessed relevant content.

[6]In this section documents refer to indexed segments.

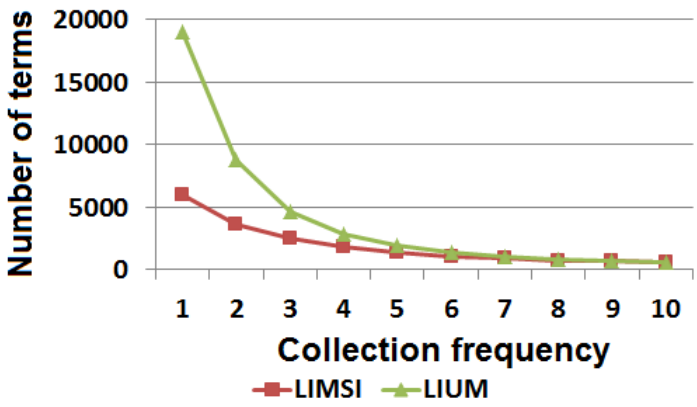


Figure 4.4: Number of terms with collection frequency equal to 1-10.

While this pooling procedure only uses results generated using one retrieval model, the diversity of the segmentation schemes used for the content means that we get a wide variety of content originating in segments generated using different methods. Additionally since the relevance labelling tool requires the assessor to examine all data from the beginning point of the first retrieved content for a meeting to the end of the last one, the assessor actually looks at content not retrieved by any schemes giving a greater coverage of the relevance assessment than would otherwise be the case.

4.3 Semi-professional user generated (SPUG) content

Development of compact and affordable recording devices is leading to an explosive growth in the amount of the content created by semi- or non-professionals. Internet sharing portals such as YouTube[7] or Vimeo[8] enable videos to be stored and shared with other users. This new type of multimedia collection presents new challenges for SCR technologies and potentially more complicated and creative types of user requests. In order to investigate SCR for this type of multimedia content, we built a retrieval collection for a SPUG content collection crawled from the Internet using the

[7]http://youtube.com/
[8]http://vimeo.com

technology of crowdsourcing to define the queries and relevance data. The retrieval task for these experiments aimed to reflect the behaviour of potential real users of this type of content and their search interests. The creation of the corpora was carried out in collaboration with Rich Speech Retrieval and Search and Hyperlinking tasks organising team at MediaEval Evaluation Campaign.

4.3.1 Blip10000 dataset

Our SPUG retrieval task was based on the Blip10000 dataset created by the Peta-Media Network of Excellence (Larson et al., 2012). This contains 14,838 Creative Commons videos from blip.tv[9], and corresponding user provided metadata. The data comprises a total of ca. 3,260 hours of data and is divided into development and test sets, of 5,288 and 9,550 videos respectively. Additionally, two transcripts were provided for all the videos in the collection, by LIMSI/Vocapia (Lamel and Gauvain, 2008) and LIUM (Rousseau et al., 2011). Since the main focus of our current study is to define suitable techniques and evaluation methods for a search task, we wished to work only with a monolingual English language dataset. To this end, we used a subset of the test set classified as English language by LIMSI/Vocapia, and transcribed by English language versions of both ASR systems. This resulted in a corpus of 4,890 video files. The audio transcription sytems were treated as black boxes, since we did not have access to the internal details of the ASR systems. Thus our comparison of the LIMSI/Vocapia versus LIUM transcripts quality relies only on the analysis of the transcripts themselves. Figure 4.4 shows a comparison of the number of terms with a total collection frequency equal to 1-3 for the two transcripts. It can be seen that this is considerably higher for the LIUM transcript, suggesting that these transcripts contain more unique terms in the documents than the LIMSI/Vocapia ones. This may be caused by the fact that the LIMSI/Vocapia system has a smaller vocabulary in the language modelling component of the ASR system; otherwise the LIMSI/Vocapia system may be more reliable in recognition

[9]http://blip.tv

of rare words. The exact explanation of this difference is not the focus of this work, as for the SCR experiments we just need to be aware of different vocabulary sizes between the transcripts of the collections.

4.3.2 MediaEval Retrieval Task

Query set and relevance data collection procedure

The collection of queries and relevant segments was carried out using crowdsourcing on the Amazon Mechanical Turk (AMT[10]) platform and was originally carried out for the MediaEval 2012 Search and Hyperlinking task (Eskevich et al., 2012a). We chose this method because the workers on such platforms are doing work online, and most probably are quite active on the Internet as well, thus performing video browsing on the web is likely to be quite a regular activity for them.

In the Mechanical Turk (MTurk) setting tasks are referred to as 'Human Intelligence Tasks' or HITs. To initiate a task, the requester uploads a HIT consisting of relevant instructions, questions, files, etc. to be used by the workers while completing the HIT. When the workers have carried out the HIT, the requester reviews the completed work and confirms payment to the worker with a previously set remuneration. If the requester is not satisfied with the work carried out, they can opt not to pay the worker.

The questions in the HIT were formulated using a concept with which general workers will often be familiar when working with the videos – sharing. The concept of sharing seemed to be part of the everyday experience of people who work with the Internet and provokes a natural human response setting. An example screenshot of our HIT is given in Figure 4.5. The workers were asked to watch a video, and to chose a segment within it that they would like to share with their friends on a social network, and to identify its start and end time. The status update that they would put on the social network represents a query to the selected segment in the form of

[10]www.mturk.com

natural language sentence (NLS). We explicitly asked the workers to be as neutral as possible when creating the status text, so that this query would not refer to any information personal to themselves. We also asked the workers to explicitly give us a query in the form of a search engine request (SER) that they would use to refind this segment.

Additionally we asked the workers to write the manual transcript of the segment of interest. This information allowed us to further analyse the ASR perfomance on these segments to be retrieved. More details about the use of crowdsourcing for collection of queries and relevance data can be found in Appendix D.

Query set and relevance segments details

The MediaEval 2012 test query set collected using this HIT consists of 30 textual queries of both NSL and SED styles. Figure 4.6 shows an example of a NSL and SER style query collected via crowdsourcing, as well as the corresponding relevant passage.

The queries were created for videos selected at random from the top 10 genre categories in the document collection. The genres were assigned to the videos by their uploaders and are available within the Blip10000 dataset release. The average frequency of the query terms in the collection lexicon is relatively high (3015 and 2897 for LIMSI and LIUM respectively), although this was lower than those of the transcripts of the relevant segments (6753 and 6342 for LIMSI and LIUM respectively). 20 queries out of 30 do not contain any out-of-vocabulary (OOV) terms, defined here as terms which do not appear in the transcript; whereas for the remaining 10, the ratio of the OOV terms over all the terms in the query is not higher than 0.25, see Table 4.3. With these statistics in mind, the feature of the transcripts that has most potential to influence the retrieval performance is ASR errors. Table 4.3 shows the correct word recognition rate for words in relevant segments for the target segment (word recognition rate (WRR)) for each query.

query ID	WRR		OOV		query ID	WRR		OOV	
	LIMSI	LIUM	LIMSI	LIUM		LIMSI	LIUM	LIMSI	LIUM
1	0.90	0.70	0.13	0.13	16	0.92	0.77	-	-
2	0.76	0.71	-	-	17	0.71	0.59	-	-
3	0.83	0.70	-	-	18	0.67	0.67	-	-
4	0.84	0.74	-	-	19	0.70	0.74	-	-
5	0.86	0.72	0.13	0.13	20	0.50	0.48	-	-
6	0.44	0.33	-	-	21	0.53	0.67	0.22	0.22
7	0.67	0.63	0.14	0.14	22	0.86	0.89	-	-
8	0.88	0.74	0.15	0.10	23	0.47	0.40	-	-
9	0.69	0.62	-	-	24	0.72	0.56	-	-
10	0.77	0.68	-	-	25	0.38	0.08	-	-
11	0.65	0.74	0.07	0.13	26	0.52	0.48	0.25	0.25
12	0.42	0.09	-	-	27	0.65	0.57	-	-
13	0.59	0.64	-	-	28	0.73	0.50	0.14	0.14
14	0.75	0.60	0.17	0.17	29	0.54	0.42	-	-
15	0.83	0.72	0.09	0.09	30	0.71	0.71	-	-

Table 4.3: Overview of the MediaEval 2012 search sub-task query set characteristics.

4.4 Summary

In this chapter we overviewed the datasets used for the experiments reported in this thesis. Since there is a lack of publicly available test collections for the challenging types of SCR task explored in this thesis, we created our own test sets for meeting search and Internet television. While the queries and relevance assessment for the meetings retrieval test collection were acquired using traditional IR method of pooling across a number of retrieval runs of the results, for the case of Internet SPUG collection we explored the potential that crowdsourcing platforms can bring to this procedure. As the Blip10000 collection already contains SPUG content shared on the Internet, an appeal to an Internet audience for query creation and relevance assessment proved to be useful, and allowed to collect the data that is representative of the use case scenario, and can be used to its future adjustement.

The next chapter examines evaluation metrics for these SCR tasks that reflect user experience when searching for relevant content within this challenging informal data.

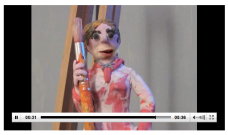
Find interesting things people say in videos

Imagine that you are going to share this video on a social network such as Facebook, Twitter, etc. And you can choose a part of this video that the other users can preview while reading your sharing comment. This part of the video, which we call a **"segment"**, will give them an idea of what the video is about and give them a chance to watch the most interesting part of it.

This segment should be between 10 and 60 seconds long.

Some videos were segmented beforehand within a larger video, so do not be alarmed if the video doesn't start at the beginning (and also don't scroll back beyond the beginning of video segment).

When you are finished with answering the questions, don't forget to click the "Submit" button at the bottom of the page. Thank you very much for your help!



We understand that work with videos is complicated, so we ask you to describe briefly what the video is about and explain what were the difficulties working with it in the text box below.

For some videos it might be hard to define a segment that can be a good highlight of the video, so if you fail to find a good segment, please skip the Questions 1-5.

1) For **your selected segment**, what is the **start time** (please specify exactly in minutes and seconds)? Please pay attention to the time shown in the **left** corner of the bottom line of the video player.

Minute Second

2) For **your selected segment**, what is the **end time** (please specify exactly in minutes and seconds)? Please pay attention to the time shown in the **left** corner of the bottom line of the video player.

Minute Second

3) What was said during **your selected segment**? Please write down the **exact words** the speaker is saying (please transcribe precisely). If you are not sure what the exact word was, please write down what your think the word was and mark it with a star (for example, 'French president *Sarkosie was saying ...' if you are not sure how to spell the name 'Sarkozy' properly)

4) When sharing this particular part of the video (your selected segment) on a social network, what **comment** would you add to the video to make sure that other users have an idea what the video segment is about?
Please do not use informal internet language (such as '4 u' instead of 'for you').
Be as objective as possible when describing the video segment and do not express your personal opinion/attitude, either positive or negative.

5) Imagine you would like to search for **similar video segments** using a search engine (such as Google, Bing, Yahoo) what would you put in the search box?

Figure 4.5: Example of the HIT used to collect a set of queries and relevant video segments for the Search and Hyperlinking task at MediaEval 2012.

Natural Language Sentence Query Example: Curtis Baylor of Allstate gives a small piece of planning advice for small business using his basic three factors.

Question for a Search engine Request Query Example: interviews with business professionals

Transcript of the relevant passage: I do what I call say needs-based, planning, dreams. So the first thing you need to come with is your vision, and then what are the needs that you have and that your business has. And then, bring me the what is your current situation. And then my job is to say, Okay, given your vision, which sets your priorities, what do you need to do in order to be able to survive and make it both in the business and individually, and the where are you right now, you know, we need to collectively do what? Tailor the plan to take into account those three areas.

Figure 4.6: Example of a query in the form of a natural language sentence (NSL), and a search engine request (SER), and the relevant passage for these queries in the Search sub-task of the Search and Hyperlinking task at MediaEval 2012.

Chapter 5

Evaluation of Spoken Content Retrieval

Since initial research on SCR focused on retrieval of spoken documents based on ASR transcripts using text IR methods, initial research on SCR was evaluated using standard text IR evaluation metrics. Further development of the SCR field beyond retrieval of distinct documents introduced a requirement for evaluation metrics designed to evaluate effectiveness of aspects of SCR not relevant to text IR tasks. In this chapter we review the traditional metrics of IR, and their variations adjusted for the specific needs of the SCR tasks, and then introduce new metrics designed to capture the user experience in SCR in forms of the efficiency of accessing relevant information and allow more detailed evaluation of SCR systems.

Text, written according to standard punctuation rules, already has structural units, typically in the form of sentences and paragraphs, and often sections and chapters. Overall, such documents are often structured into topically distinct regions. This means that it is often quite easy to identify meaningful retrieval units and build a retrieval collection based on well-defined documents that have binary relevance to the query. In this situation it is reasonable to apply standard document based evaluation metrics, since the searcher will generally be interested to listen to the retrieved document from its beginning. For more spontaneous mate-

rial, such as the spoken material based on lectures and meetings, and other informal dialogue based material, introduced in previous chapters, the situation is more complex. Since the material generally lacks formal structure it is not possible to extract unambiguous topically focused retrieval units, even manually, since people will not agree on segmentation points, which may in any case be dependent on the search query.

However, in order to support efficient user access to relevant content, an SCR system must: i) seek to retrieve relevant content at high ranks, ii) minimize the amount of non-relevant content in the retrieved unit. These points are particularly important for SCR, since audition of spoken content is very time consuming.

These factors mean that standard document based evaluation metrics are not sufficient for evaluation of SCR on informal spoken content. Evaluation needs to reflect the user experience of information access, but also to enable researchers to investigate and understand the factors which impact on retrieval effectiveness, including WER, content segmentation, and IR ranking models.

In this chapter we briefly overview relevant evaluation metrics used for the standard document based retrieval, then follow up with a discussion of metrics that have already been introduced for segmented text content and SCR in other work. Finally we introduce new evaluation measures that are inspired by the metrics already in use, but help to take into account the specificity of working with audio content.

5.1 Textual Content Retrieval Evaluation Metrics

Document level evaluation metrics developed for textual IR show the quality of the documents ranking. Metrics that target the evaluation of passage element retrieval can potentially be used for SCR tasks, that need to retrieve spoken segments, and reflect the experience of reading the transcript. In this section we describe these metrics in overview.

5.1.1 Mean Average Precision (MAP) and Mean Reciprocal Rank (MRR)

One of the most widely used document level retrieval metrics is mean average precision (MAP) (Büttcher et al., 2010). MAP was widely used in evaluation of SCR in early work such as TREC SDR task, and was included in our review of early SCR research in Section 3.1. In order to make our further metrics investigation clearer we cite here the equation for Average Precision (AP) calculation introduced earlier:

$$AP = \frac{1}{n}.\sum_{r=1}^{N} P[r] \cdot rel(r) \tag{5.1}$$

In the case of known-item search mean reciprocal rank (MRR) is used to evaluate the retrieval results, as introduced in Section 3.1:

$$MRR = \frac{1}{|Q|}.\sum_{i=1}^{|Q|} \frac{1}{rank_i} \tag{5.2}$$

5.1.2 Mean Average interpolated Precision (MAiP)

Since MAP has a binary score relevance, it generally assumes that if the retrieval units are taken from within larger documents, they have been perfectly segmented into coherent topical units. Even for search with multi-topical retrieval units it is assumed that any relevant content is contained completely in the retrieval unit. Therefore MAP is not a good measure when relevant content may have been split between multiple segments.

In order to measure the amount of relevant content contained within a passage the *Mean Average interpolated Precision (MAiP)* metric was introduced for the text passage retrieval task at INEX (Kamps et al., 2008). Document relevance was not counted in a binary way, but rather it was assumed that the amount of relevant information retrieved should be reflected in the metric. This metric is based on the mean generalised average precision (mGAP) that was introduced to deal with

human assessment of partial relevance (Kekalainen and Jarvelin, 2002).

In MAiP, precision at rank r is defined as the fraction of retrieved text that is relevant,

$$P[r] = \left(\sum_{i=1}^{r} rsize(s_i) \right) \quad / \quad \sum_{i=1}^{r} size(s_i) \tag{5.3}$$

where r is the rank of the document, s_i is the document at rank r, $rsize(s_i)$ is the length of relevant text contained in s_i in characters (if there is no relevant text in s_i, $rsize(s_i)$=0), $size(s_i)$ is the total number of characters in s_i.

Recall at rank r is defined as the fraction of relevant text that is retrieved,

$$R[r] = \left(\sum_{i=1}^{r} rsize(s_i) \right) \quad / \quad Trel(q) \tag{5.4}$$

where $Trel(q)$ is the total number of relevant characters across all segments, i.e. the sum of the lengths of the (non-overlapping) relevant regions.

The INEX organizers were afraid that $P[r]$ can be biased towards systems returning several shorter segments rather than returning one longer segment that contains them all. This prompted INEX to define MAiP in terms of precision at fixed recall levels rather than ranks. Thus, the measure interpolated precision $iP[x]$ is defined as the maximum precision at the selected recall level x. Retrieval effectiveness is calculated using average interpolated precision A_iP calculated by averaging the interpolated precision scores calculated at 101 recall levels (0.00, 0.01, ..., 1.00),

$$AiP = \frac{1}{101}. \sum_{x=0.00,0.01,...,1.00} iP[x] \tag{5.5}$$

MAiP is calculated by computing the mean of the AiP values across the topic set. Although MAiP looks to be a suitable metric for evaluating speech search using segments, the way of averaging is inconvenient for speech tasks as discussed later in Section 5.2.3.

5.1.3 Discussion

As discussed in Section 3.1, MAP and its variation for the case of only one relevant document MRR, though suitable for cases of documents with defined boundaries and binary relevance, do not allow thorough analysis of more challenging SCR cases. On the other hand, the MAiP measure targets passage retrieval, though it does not take into account time information of the audio content that plays a crucial role when the user is to listen to the SCR results.

Only some of the state-of-the-art transcript systems contain punctuation symbols (Gauvain et al., 2002), therefore a metric that relies on punctuation will have limited usefullness since it cannot be generally used for spoken data ASR transcripts. Speech naturally contains pauses (periods of certain length of time when there is no spoken content, because speakers make a pause on purpose or take a breath), therefore regions between pauses can be used as potential unit separators (Akiba et al., 2011). However, they do not always correspond to actual sentence boundaries.

5.2 Spoken Content Retrieval Evaluation Metrics

Mean Generalized Average Precision (mGAP), introduced in Section 3.2.3, was one of the first attempts to create a metric that incorporates the calculations for the textual content but also takes into account audio specific features. In this section we give an overview of measures that treat audio content as documents consisting of sentences that are separated by pauses. We further discuss new metrics that are based on precision/recall of the relevant content within the segment, and those that are time precision oriented. This set of metrics will be used in evaluation of our experiments to describe the results from different perspectives in the rest of this thesis.

In general the dimension of time is important for speech because the same words can be pronounced with different speeds, thus the amount of sentences that are of same or comparable length in a textual version may take different times for the

user to listen to. Thus listening to a passage spoken by different speakers, though corresponding to a transcript of the same length in terms of text (i.e. number of symbols, words, sentences), can require a different amount of time. This aspect of user experience has to be taken into account by the metrics used to assess SCR performance.

5.2.1 Metrics using Inter-Pausal Units (uMAP, pwMAP, fMAP)

The NTCIR SpokenDoc tasks introduced a set of metrics inspired by MAP, but which are adjusted for passage retrieval (Akiba et al., 2011). The transcript is divided into units between the pauses in the audio (so-called Inter-Pausal Units (IPU)). IPUs only partially reflect the user experience in terms of time spent on auditionning them, because the IPUs themselves can vary in length, and this is not taken in to account by the metrics. The relevance for the passages was manually created in terms of number of relevant IPUs that constitute a relevant passage.

Utterance-based MAP (uMAP)

The uMAP metric calculates MAP on the level of IPUs after each retrieved passage has been expanded into its constituent IPUs and they have been rearranged so that the relevant IPUs are at the beginning of the sequence.

Suppose the ordered list of passages $P_q = p_1 p_2 \cdots p_{|P_q|}$ is submitted as the retrieval result for a query q, and further that we have a mapping function $O(p)$ from a (retrieved) passage p to an ordered list of utterances $u_{p,1} u_{p,2} \cdots u_{p,|p|}$. We can obtain the ordered list of utterances $U = u_{p_1,1} u_{p_1,2} \cdots u_{p_1,|p_1|} u_{p_2,1} \cdots u_{p_{|P_q|},1} \cdots u_{p_{|P_q|},|p_{|P_q|}|}$, then $uAveP_q$ is calculated as shown in Equation 5.6.

$$uAveP_q = \frac{1}{|\tilde{R}_q|} \sum_{i=1}^{|U|} \delta(u_i \in \tilde{R}_q) \frac{\sum_{j=1}^{i} \delta(u_j \in \tilde{R}_q)}{i}, \tag{5.6}$$

where $U = u_1 \cdots u_{|U|} (|U| = \sum_{p \in P} |p|)$ is the renumbered ordered list of U and

$\tilde{R}_q = \bigcup_{r \in R_q} \{u | u \in r\}$ is the set of relevant utterances extracted from the set of relevant passages R_q.

For the mapping function $O(p)$, an oracle ordering mapping function is used to reorder the utterances in the given passage p so that the relevant utterances come first. For example, given a passage $p = u_1 u_2 u_3 u_4 u_5$ where the relevant utterances are u_3 and u_4, the mapping function will return the passage as $u_3 u_4 u_1 u_2 u_5$.

uMAP is defined as the mean of the $uAveP$ over all query topics Q.

$$uMAP = \frac{1}{|Q|} \sum_{q \in Q} uAveP_q. \tag{5.7}$$

Point-wise MAP (pwMAP)

The pwMAP metric counts as relevant only segments for which the IPU in the centre of the segment is relevant. If the center utterance is included in a relevant passage found in the golden relevance file, then the returned passage is deemed relevant and the relevant passage is considered to be retrieved correctly. However, if there is at least one passage earlier in the retrieved list that is also deemed relevant with respect to the same relevant passage, the returned passage is deemed non relevant since this relevant passage has been retrieved already. In this way, all the passages in the result list are labeled by their relevance. Thus any conventional evaluation metric designed for document retrieval can be applied to the returned list.

For an ordered list of correctly retrieved passages $r_1 r_2 \cdots r_M (M \leq |R_q|)$, where the relevance of each passage is judged according to the process described above. $pwAveP_q$ is calculated as shown in Equation 5.8.

$$pwAveP_q = \frac{1}{|R_q|} \sum_{k=1}^{M} \frac{k}{rank(r_k)}, \tag{5.8}$$

where $rank(r)$ is the rank of passage r in the original ordered list of retrieved passages.

pwMAP is defined as the mean of the $pwAveP$ over all query topics Q.

$$pwMAP = \frac{1}{|Q|} \sum_{q \in Q} pwAveP_q. \tag{5.9}$$

Fractional MAP

The fMAP metric is designed to capture the relevancy of the segments. This measure evaluates the relevance of a retrieved passage fractionally against the relevant passage in the golden relevance files. Given a retrieved passage $p \in P_q$ for a given query q, its relevance level $rel(p, R_q)$ is defined as the fraction of the relevant passage that it covers, as shown in Equation 5.10.

$$rel(p, R_q) = \max_{r \in R_q} \frac{|r \cap p|}{|r|}. \tag{5.10}$$

Here r and p are sets of utterances. rel can be seen as measuring the recall of p at the utterance level. Accordingly, we can define the precision of p as shown in Equation 5.11.

$$prec(p, R_q) = \max_{r \in R_q} \frac{|p \cap r|}{|p|}. \tag{5.11}$$

Then, $fAveP_q$ is calculated as shown in Equation 5.12.

$$fAveP_q = \frac{1}{|R_q|} \sum_{i=1}^{|P_q|} rel(p_i, R_q) \frac{\sum_{j=1}^{i} prec(p_j, R_q)}{i}. \tag{5.12}$$

fMAP is defined as the mean of the $fAveP_q$ over all query topics Q.

$$fMAP = \frac{1}{|Q|} \sum_{q \in Q} fAveP_q \tag{5.13}$$

5.2.2 Precision/Recall of the relevant content within the segment

The NTCIR official metrics described in previous section allow to check how many of the relevant IPUs have been retrieved. However, even though we process the transcripts and return as output a list of segments with information about the start and end times (e.g. in terms of time or IPUs), our ultimate goal is to provide the user with audio or video segments to listen to. Therefore the actual timing information of the beginning and end points of relevant data are important for the analysis of results and is not provided by the uMAP, pwMAP, and fMAP. Thus, in order to better understand the relationship between retrieved segments and the amount of relevant content that has actually been retrieved, we calculate the within segment precision/recall for each retrieved segment which contains some relevant content. We then calculate the average of these precision/recall values for each query, and then average these values across the complete query set.

5.2.3 Novel Time Precision Oriented Metrics for SCR (MASP, MASDwP)

In this section we describe two new metrics for evaluating retrieval effectiveness for searching informally structured spoken content taking into account time information in terms of both precision of the retrieved segments and the distance of the beginning of the retrieved segment to the real start of the relevant content.

Mean Average Segment Precision (MASP)

MASP is a modification of MAP, inspired by MAiP, but specifically adapted to speech search when no pre-defined segmentation of search units exists. The motivation for MASP is to create a metric that combines both the ranking quality and the segmentation quality with respect to precision of relevant content in a segment in a single score. Thus, the ideal state for MASP is not only to retrieve the relevant

speech segments at the top of the ranked results list, but also to have precision of 100% for each segment over relevant speech data. Unlike MAP, relevance for MASP varies from 0 to 1 according to the amount of relevant content present in the segment. This is similar to the measurement of relevance in MAiP, but there are two fundamental differences: the amount of relevant content is measured over time instead of text; and the average segment precision (ASP) is calculated at the ranks of segments containing relevant content rather than fixed recall points as in MAiP.

Segment precision ($SP[r]$) at rank r in MASP is calculated as shown in Equation 5.14.

$$SP[r] = \sum_{i=1}^{r} rperiod(s_i) \quad / \quad \sum_{i=1}^{r} length(s_i) \tag{5.14}$$

where $length(s_i)$ is the length of segment s_i in time units (minutes or seconds), and $rperiod(s_i)$ is the length of the relevant period in the segment s_i. Unlike MAiP, the average segment precision (ASP) is calculated at the ranks where relevant content is found as shown in Equation 5.15:

$$ASP = \frac{1}{n}.\sum_{r=1}^{N} SP[r] \cdot rel(s_r) \tag{5.15}$$

where n is the number of segments that contain relevant content, and $rel(s_r)$ is equal to 1 if s_r contains any relevant content, and 0 otherwise. MASP is defined as the mean of ASP across a query set Q.

$$MASP = \frac{1}{|Q|}.\sum_{q \epsilon Q} ASP_q \tag{5.16}$$

The motivation behind taking the average of $SP[r]$ over the ranks of relevant content is the same as that for MAP. The assumption is that the position where a user stops checking the ranked list is usually a relevant item, which varies for different users. The stopping position at a relevant rank is assumed to be uniformly distributed, which is why the AP is calculated in this way.

The claim for applying the averaging of MAiP at fixed recall points, as described

Rank	**1**	2	**3**	**4**	5	**6**	Avg
rperiod/length	2/3	0/5	3/4	6/6	0/2	5/10	Value
Prec.[r]	1	1/2	2/3	3/4	3/5	4/6	0.771
SP[r]	2/3	2/8	5/12	11/18	11/20	16/30	0.557
SDWP[r]	2/3 * 1.0	2/8	5/12 * 0.9	11/18 * 0.0	11/20	16/30 * 0.0	0.260

Table 5.1: Example comparing AP, ASP, and ASDWP. The average values are calculated at ranks in bold, the segment at the first rank starts with the relevant information, the relevant content at the third rank position starts only later within the segment, the relevant content starts long before the segments found at ranks 4 and 6.

by the INEX organizers (Kamps et al., 2008), is that the score can be biased towards retrieving shorter segments. However, we hypothesize that this issue is automatically resolved within the implementation of MASP. In MASP retrieving shorter segments of relevant content will increase the number of segments with relevant content (n), therefore the averaging process will be applied on a larger number of ranks and ASP will thus not be biased to the length. MASP is low when the percentage of relevant parts in segments is consistently low, which indicates bad segmentation, or when the ranks of the relevant contents are deep in the results list.

Table 5.1 shows a simple illustrative example of how ASP is measured and compared to AP. The example topic has 4 relevant segments appearing in the top 6 rank positions. As shown in Table 5.1, ASP takes into consideration the length of each segment as well as the percentage of relevant content in each one. It can be seen that long and short segments are not treated the same; the score gets lower values when long periods of irrelevant speech are returned on the top of the list. This factor is not measured when using standard AP.

Mean Average Segment Distance-weighted Precision (MASDWP)

Since it takes a lot of time to listen to the spoken content, it is important for the user to be able to start playback as close to the beginning of the relevant content, i.e. jump-in point, as possible. Therefore evaluation metrics for SCR should take into account the distance from the beginning of the retrieval unit to the jump-in point.

The ASP metric reflects the amount of relevant content present at different ranks. However it does not show how far the user has to listen into the segment at a certain rank until the relevant part actually begins, or whether the segment starts after the beginning of the relevant part and the user will have to rewind in the recorded audio signal beyond the beginning of the segment in order to get to the starting point of the relevant content. In order to take this information into account we introduce the same style of penalty function that was used in CLEF CL-SR task evaluation (Equation 3.3). The new ASDWP metric is shown in Equation 5.17. this can be seen to combine MASP with mGAP.

$$ASDWP = \frac{1}{n}.\sum_{r=1}^{N} SP[r] \cdot rel(s_r) \cdot \left(1 - \frac{Distance}{Granularity} \cdot 0.1\right) \tag{5.17}$$

As for the other metrics, MASDWP is calculated by averaging the ASDWP across the topic set.

In the illustrative example in Table 5.1, if we suppose that the first segment starts at the relevant point, that the third has a playback drop-in point inside the segment at one step from the start of the segment and that the fourth and sixth segments are actually far beyond the limit set for the distance to be of practical relevance to the user, only the first and third results are included in the calculation of ASDWP, and thus the metric reflects whether while listening to the results in the ranked list the user will start the playback close to the beginning of the actual relevant data. The MASDWP score reflects that the example ranking of the results is far from being efficient for the user, as s/he would have to listen to a lot of non relevant information and may even skip some passages that contain the information too far away from the beginning of the segment.

5.3 Summary

In this chapter we overviewed a number of metrics that have been used for evaluation of effectiveness of SCR for informally structured data. The metrics reviewed included the standard document based retrieval metrics (MRR, MAP), and metrics introduced for SCR in benchmark evaluation campaigns including mGAP, uMAP, pwMAP, fMAP. Following this we introduced two new measures: one that focuses on the precision of the relevant content in the list of retrieved documents (MASP) and another that combines this calculation with the information about the jump-in point distance (MASDWP) to fully understand the retrieval behaviour and its impact on the user experience. We argue that SCR results should be evaluated using a set of measures, since although MAP reflects overall ranking performance (even though for segments with unknown boundaries, this metric requires certain adjustement); mGAP helps to relate ranking and closeness to the jump-in point; while MASP and MASDWP introduce factor of the user experience, i.e. how much of the relevant versus irrelevant content the user will have to listen to. The calculation of the precision and recall of the relevant content within the segment describes the segments quality.

In the following chapter we describe our experiments on the datasets introduced in Chapter 4, and carry out an extensive analysis of retrieval behaviour in context of varying ASR transcript quality and different segmentation methods used. Further on in the thesis we use evaluation metrics examined and introduced in the current chapter to trace the impact of applied methods to improve SCR.

Chapter 6

Exploring the impact of ASR errors and alternative segmentation methods on SCR behaviour

In the previous chapters we gave an overview of the IR and ASR fields that are combined within the SCR process, and demonstrated the SCR development with the increasing availability of diverse data collection. In this chapter we use three of the described datasets (AMI, NTCIR-9 SpokenDoc, and BlipTV) to carry out an analysis of the impact of ASR errors and different segmentation techniques on SCR behaviour. In order to be able to do a fair comparison, we use one segmentation method for both ASR and manual transcripts for our experiments. Therefore, we first describe the principle that enables this cross comparison, and then discuss the influence of different factors on the results.

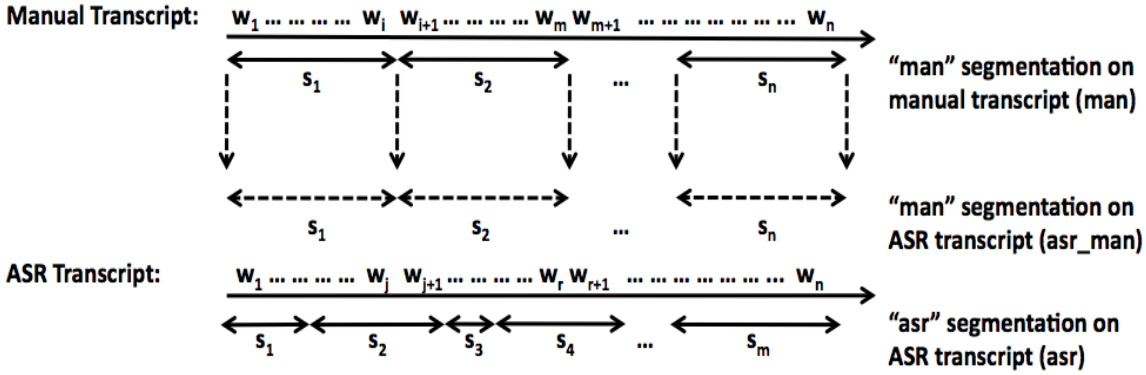


Figure 6.1: Example of "man" segmentation projection on ASR transcript ("asr_man")

6.1 Using consistent segment boundaries for both manual and ASR transcripts

Corpora that contain both ASR and manual transcripts provide a significant opportunity for direct comparison of the influence of ASR errors on all SCR behaviour, since all manipulations can be carried out on both sets of transcripts (manual and ASR).

In this thesis we aim at implementing automatic methods for data processing in SCR, thus we chose to use automatic segmentation methods. However it is extremely time consuming and expensive to obtain manual assessment of the segmentation results for different segmentation techniques over a large corpora. Moreover, the details of topical segmentation may vary depending on the specifity of the request for information in a potential user's query, which makes it impractical to attempt to judge only one segmentation result as correct for all future applications. At the same time, the manual transcript represents an ideal case of input to the segmentation algorithm since it corresponds to the perfect automatic recognition transcript. Therefore we consider the output of all segmentation approaches carried out on the manual transcript to be a reliable baseline. We understand that for SCR experiments it cannot be claimed to be the best possible segmentation of the data; and the reference to it as manual in our experimental study refers only to the source

transcript input, not the segmentation method. Thus, although we consider the manual transcript to be the ideal version of the speech recognition transcript, we do not suppose that it forms a gold standard for our segmentation runs.

While we can of course run the segmentation methods on the ASR transcript separately, tracking the influence of the ASR accuracy on SCR for the segments will not be possible if the two segmented collections (manual and ASR) have different segment boundary points. Thus we project the segment borders of the manual transcript onto the ASR transcript by using the word timing information of the transcripts, see Figure 6.1 for an illustration. This results in a third collection of segments (labelled *asr_man* in our experiments) where the only difference between this and the manual transcript segment collection is that the content of the segments is formed from the ASR transcripts.

Sometimes manual transcripts do not cover the whole region of the ASR transcripts since they do not include areas regarded as not relevant to the meetings by the manual transcribers, in these cases the additional words in the ASR transcript are placed in the adjoining manual segment.

6.2 Analysis of SCR for meeting search using the AMI corpus

As introduced in Section 3.2.2, we can consider many meeting retrieval tasks to be recall focused activities where the user needs to find all relevant items. In recall focused search, the user typically goes much deeper into the ranked retrieval list than in standard precision focused search tasks. For example, patent examiners may look more than 100 items deep into the returned list. From a practical perspective due to the temporal nature of speech content and the time taken to review content of this type, we assume that the users can reasonably be expected not to look for relevant items as deep into the list as for a text-based patent examiner. For this study we thus chose the rank 50 as the cut off for retrieved results that a user will

be ready to listen to in order to attempt to find the relevant content. We thus take retrieval at rank 50 as the cut off point for our experimental analysis. Clearly in practice if the user is looking for a known relevant item and has not found it after inspecting the first 50 items they may be willing to continue searching. Similarly if the information that they are looking for is particularly valuable they may continue to keep searching. It should be noted that while we have chosen a cut off of 50 retrieved items for this study, similar trends to those observed and examined here in the ranking of retrieved items would be observed if a different cut off point in the list were to be used.

6.2.1 Segmentation of the AMI Corpus

The meetings in the AMI corpus have an average duration of approximately 30-minutes. This is too long to expect the user to listen to them in their entirety or even to read through a transcript in their search for relevant information. Additionally since they cover multiple topics, it is likely to be difficult to distinguish between similar multi-topic meeting transcripts in the retrieval process to identify those containing relevant information. Thus we need to segment the meetings to identify suitable more topically focused retrieval units. This segmentation is motivated both by the need for segment cohesion to promote the selectivity of relevant content in retrieval, and also for efficient location of specific information of interest within a relevant segment.

As discussed in Section 4.2.1, since the topic segmenation provided with the AMI collection does not cover all 160 meetings that have both ASR and manual transcripts, and thus are taken into consideration, we decided to automatically segment the AMI meeting transcripts ourselves. We used implementations of several statistical algorithms based on lexical cohesion (C99, TextTiling, Minimum Cut) and fixed length based segmentations that take into account number of words or time units. We gave a short description of each of these algorithms in Section 2.3, and in the next section we introduce the details of their application to the AMI corpus.

Lexical Cohesion Based Algorithms

The C99, TextTiling and Minimum Cut algorithms work with the fundamental unit of the sentence, placing segment boundaries between the end of one sentence and the start of the next one. Thus since we do not have punctuation in the ASR transcripts, we perform segmentation for both manual and ASR transcripts using pseudo "sentences." We found the average length of real sentences in the manual transcripts to be 15 words, and we thus used this as the length of our pseudo-sentences.

After application of the C99 algorithm, the total number of segments for the 160 meetings dataset was found to be 2462 for the ASR transcripts, and very similar, 2476, for the manual ones. This yielded an average word count per segment of approximately 351 and 347 respectively for the two transcripts sets. In terms of the absolute number of segments, it appears that the algorithm behaves in the same way for the different transcripts. However when the actual segmented files are compared, many instances of, sometimes large, variations in the number of segments are revealed between different transcripts of the same meeting. Less than 10% of the meetings were found to have the same number of segments for manual and ASR transcripts, some have significantly more segments for the ASR transcript or for the manual one. Even where the same overall total number of segments are found for a meeting, this does not mean that the borders are positioned in similar locations. Manual analysis of segmented files showed that within an individual file while a number of the segments may line up well between the transcripts, in other localised regions there are significant variations in the number of segments. We examined a number of factors to explain this effect, the one that appears to correlate in many cases is the word recognition rate of content words (WRRC), i.e. words that were taken into account by the segmentation algorithm. This revealed that consistent low WRRC across a region usually corresponds to the undersegmentation of the ASR transcript, i.e. there are more segments in the manual transcript for this region; while significant local variations in WRRC for short regions between average

Segmentation Algorithm	Average Segment Length (words)	
	MANUAL	ASR
C99	346.6	351.3
TextTiling	363.6	374.5
Time_60	160.5	159.4
Time_90	240.2	236.4
Time_120	313.0	312.7
Time_150	389.5	387.8
Time_180	464.1	461.3
Length_non_stop_words_100	275.1	292.5
Length_non_stop_words_200	560.7	597.3
Length_non_stop_words_300	855.7	916.5
Length_non_stop_words_400	1157.8	1242.3
Length_non_stop_words_500	1448.9	1556.0
Length_non_stop_words_600	1752.2	1864.5
Length_non_stop_words_700	1995.6	2130.9
Length_non_stop_words_800	2292.7	2430.1
Length_non_stop_words_900	2456.0	2629.5
Length_non_stop_words_1000	2685.5	2772.7

Table 6.1: Average segment length (words)

WRRC and low WRRC are generally associated with greater numbers of segments in the ASR transcript for this region.

The MCut algorithm requires as one of its input parameters the number of the segments into which the input document is to be divided. The authors of the algorithm set this based on manual analysis of examples of their input data (Malioutov, 2006). Since we assume that there is no manual segmentation available, and that the desired number of segments will differ for each input file, we used the number of segments generated by one of the other algorithms for each transcript file as the input parameter for the MCut algorithm. This resulted in runs named mcut_C99 for C99, and mcut_tt for TextTiling.

Length-based Segmentation

In order to examine the value of lexical cohesion based segmentation in our speech retrieval task, we also carried out segmentation of the manual and ASR transcripts based only on timing information and the number of words in a segment. For the

time-based segmentation, segment boundaries were placed at regular intervals of 60, 90, 120, 150, 180 seconds. These time intervals were chosen as the range from minimal to average length of automatically created segments. The time boundary points were applied with flexibility to prevent words at the boundaries being split between segments. Segmentation based on the number of words had two variations: all the words were taken into account and the segment point placed after every 100, 200, 300, 400, 500, 600, 700, 800, 900 and 1000 words; and alternatively only non-stop words that were actually used by the segmentation and retrieval system afterwards were considered when counting the segment length (100-1000) and marking the boundary (length_non_stop_words or length_nsw).[1]

Table 6.1 summarizes the average length of segments created using the lexical cohesion based algorithms, fixed time intervals and the number of non stop words. As would be expected, the average number of words increases for longer time segments and with the number of non stop words counted.

6.2.2 Metrics scores comparison

In this section we report retrieval results for the test set of the search collection. Initially we ran the experiments with varying segment length on the development set in order to narrow the list of the length based segmentation methods to be used for the main runs on the test query set. The results for the development set are consistent with those of the larger number of the queries in the test set (25 in the test set against 10 in the development). Results on this larger query set can though be expected to be more reliable. Therefore we describe only the results for the test set in this thesis.

All evaluation metrics are calculated for alternative segmentation methods that produce segments of comparable average length (tt, mcut_tt, C99, mcut_C99, time_120, time_150, time_180, len_300, len_400, len_nsw_100, and len_nsw_200), we include two

[1]Stop words were taken from the list at `http://snowball.tartarus.org/algorithms/english/stop.txt`

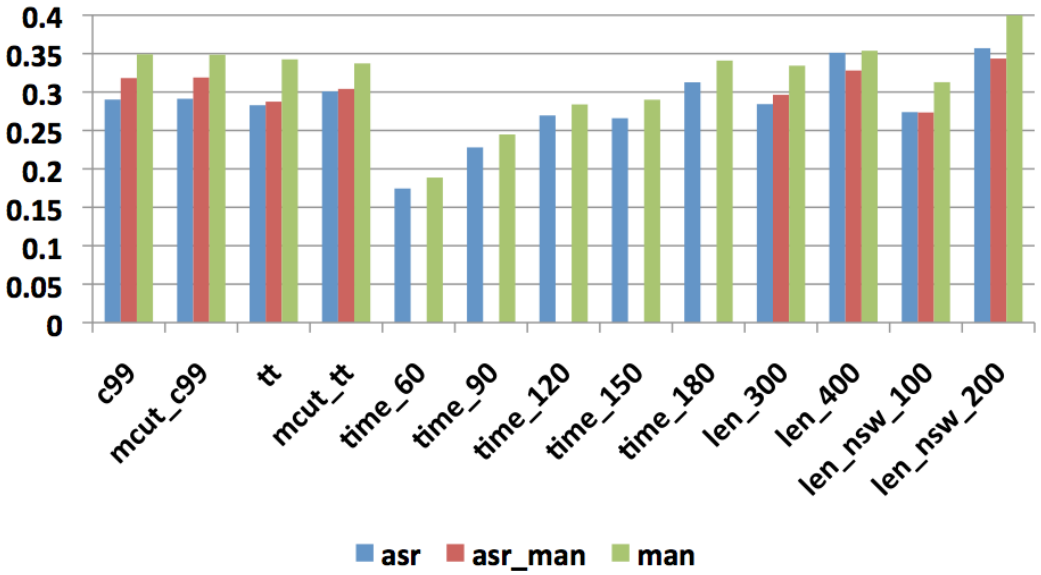


Figure 6.2: Mean Average Precision (MAP) for three types of content (ASR transcript (asr), ASR transcript with manual segmentation boundaries (asr_man), manual transcript (man)) for different segmentation schemes (cut off at rank 50)

Figure 6.3: Mean Generalized Average Precision (mGAP) for three types of content (ASR transcript (asr), ASR transcript with manual segmentation boundaries (asr_man), manual transcript (man)) for different segmentation schemes (cut off at rank 50)

Figure 6.4: Mean Average Segment Precision (MASP) for three types of content (ASR transcript (asr), ASR transcript with manual segmentation boundaries (asr_man), manual transcript (man)) for different segmentation schemes (cutoff at rank 50)

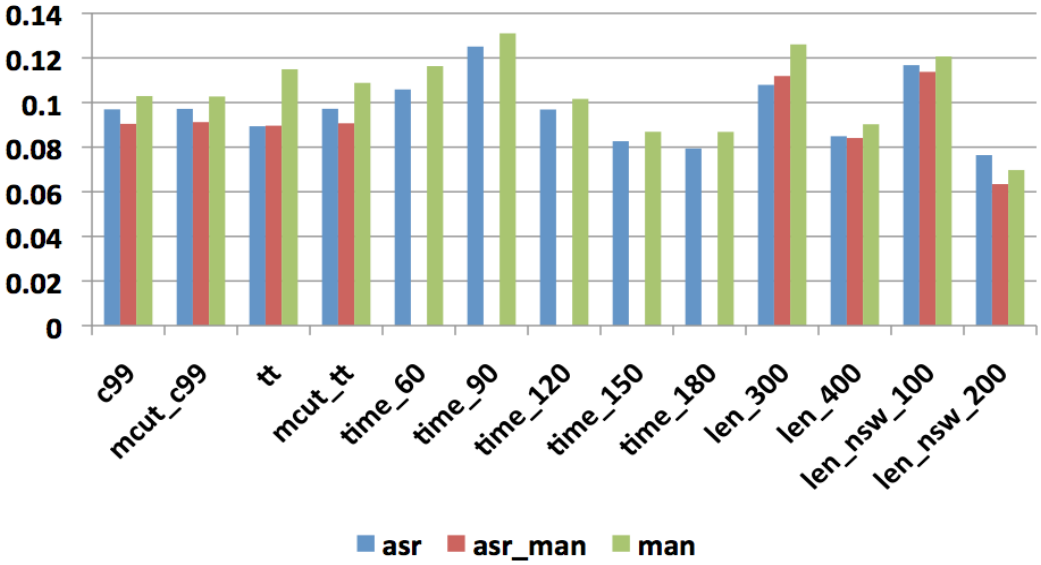


Figure 6.5: Mean Average Segment Distance-weighted Segment Precision (MAS-DwP) for three types of content (ASR transcript (asr), ASR transcript with manual segmentation boundaries (asr_man), manual transcript (man)) for different segmentation schemes (cutoff at rank 50)

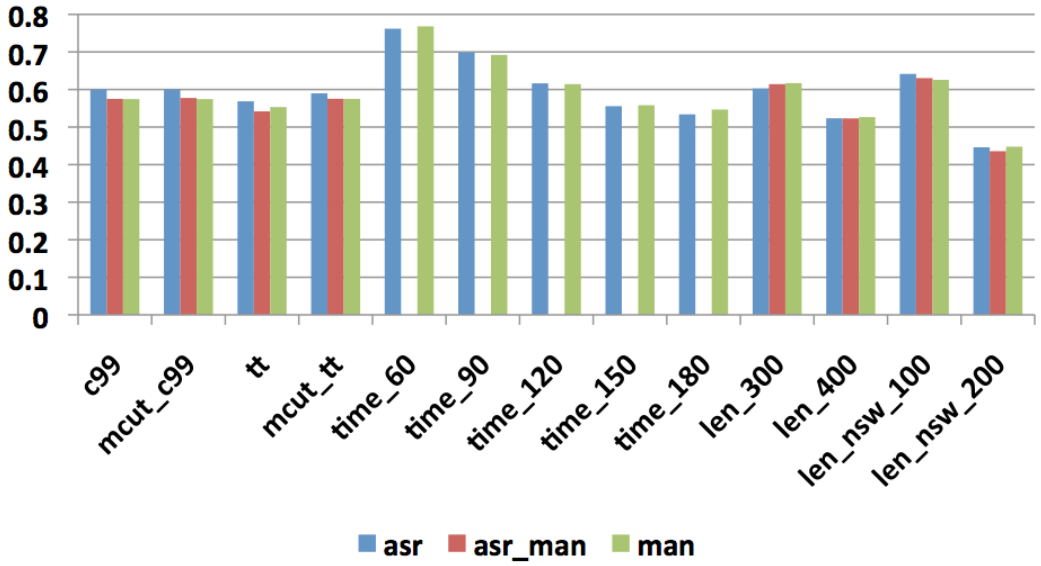


Figure 6.6: Precision of the relevant content within segments containing relevant content for three types of content (ASR transcript (asr), ASR transcript with manual segmentation boundaries (asr_man), manual transcript (man)) for different segmentation schemes (cutoff at rank 50)

shorter versions of time segmentation (time_60 and time_90) for additional analysis of retrieval performance on the shorter segments.

Figures 6.2 and 6.3 show MAP and mGAP respectively calculated for the different segmentation schemes with the three segment types. These figures allow evaluation of the retrieval results from the perspective of ranking of segments containing relevant content, and ranking of the segments containing the jump-in point or being within the allowed window distance from the jump-in point. Figures 6.4 and 6.5 show MASP and MASDwP for the same retrieval runs. These figures show the difference between retrieval performance for the different segmentation methods where the retrieval performance metrics are based on the length of the segments in time combined with rank, and rank and distance to jump-in point. Overall, as would be expected the manual content with automatic segmentation shows better results than ASR content with both forms of segment boundaries. We calculate a Wilcoxon signed-rank test with significance level equal to 0.05 (Wilcoxon, 1945) for these MAP values, and further for mGAP, MASP, MASDWP values. The results for MAP show that higher MAP values for the runs based on manual transcripts

are statistically significant for most of the segmentations present (all except len_400 and len_nsw_100). The difference in results between two segmentations based on the same segments containing different transcripts (asr_man vs man) is statistically significant only for tt, ncut_tt, len_300, len_400, len_nsw_100, and len_nsw_200 runs.

For Figures 6.2 and 6.3 it can be seen that for segments based on time segmentation MAP is higher for the longer segment units. However, according to the mGAP metric the time_90 segmentation retrieves more segments which are close to the jump-in point. We can see from Figure 6.4 that there is no significant difference between time_90, time_120, time_150 and time_180 runs in terms of MASP according to Wilcoxon signed-rank test with significance level equal to 0.05 (Wilcoxon, 1945). This arises since they retrieve approximately the same proportion of relevant content across the length of the segments in the top 50. In general, the precision of the relevant content within the segments containing relevant content is lower for longer segments, cf Figure 6.6. Of course, it should be noted that for longer segments there are less relevant and irrelevant segments within collection. This leads to better MAP, but due to the lower precision of relevant content within the retrieved segments, MASP remains at the same level. In the case of time_60 segmentation, the segments have the highest precision of relevant content. However, these segments are harder to retrieve because they contain less potentially relevant content, and there are more relevant segments to be accounted in the collection which can match with the query. Overall this means that retrieval is more difficult and that relevant segments are spread out over a longer list of retrieved items.

Looking further at these figures, retrieval of length-based segments follows the same pattern for all cases (len_300 vs. len_400, and len_nsw_100 vs. len_nsw_200): longer segments have better MAP, but the precision of the relevant content is lower, and the retrieved segments begin further from the jump-in points. This will degrade the efficiency of user access to relevant information, since they have to rewind or listen to long parts of irrelevant content.

The lexical cohesion based methods show similar levels of performance for all

four metrics that are not statistically significant accroding to Wilcoxon signed-rank test with significance level equal to 0.05 (Wilcoxon, 1945). Comparing the MAP results for asr, asr_man and man runs in Figure 6.2, it can be seen that as we might expect, man runs are better than asr_man runs which are better than asr runs. This corresponds to the average ranking of relevant content. While this trend is consistent for MAP, except for very long fixed length segments, looking at the mGAP results in Figure 6.3 shows different behaviour where in many cases the asr ranking results are better. This can be explained by the fact that for the lexical cohesion methods, the difference in segmentation between the asr and asr_man runs is much greater than for the fixed length segments. This arises due to the difference between the segmentations for the manual and ASR content. Thus the segments that might be lower in the result list for the asr runs, as compared to the asr_man runs as shown by the lower MAP, in a number of cases have a better starting time which is closer to the ideal jump-in point. This is reflected in the higher mGAP values for the asr results in these cases.

The results in Figure 6.4 for MASP are quite similar across all segmentation schemes. These results are a combination of rank and precision of relevant content in the segment at each relevant rank. The results show that the precision of the content in the individual segments can often be higher where the MAP is lower. This is most noticeable for time_60, where the MAP result in Figure 6.2 is much lower, but the MASP value in Figure 6.4 is comparable to that of other segmentation schemes. These relative differences are most noticeable for shorter segmentation schemes, indicating that while the query-document match may be less strong due to lower numbers of terms in short segments, the absence of non-relevant content in them, leads to better segment precision.

Finally, Figure 6.5 showing MASDwP records much greater variation when combining rank of relevant items, distance to jump-in point and segment rank precision. The fact that time_90 outperforms the other methods according to this metric shows that retrieval using these segmentation units has the best balance of the ranking of

Table 6.2: Average number of changes in rank for man (m) and asr_man (a_m) runs within the top 50 retrieved results

Segmentation type	Type of change				
	a_m<m	a_m≤50<m	a_m=m	m<a_m≤50	m≤50<a_m
C99	9.8	1.6	2.9	6.1	3.8
MCut_C99	9.8	1.6	2.9	6.1	3.8
TextTiling	9.5	2.2	2.2	6.8	3.4
MCut_tt	8.9	1.5	3.0	7.3	3.1
Time_120	10.2	2.0	2.2	6.5	3.5
Time_150	9.9	1.9	2.2	7.2	3.4
Time_180	9.4	1.5	2.2	7.5	3.2
Len_300	10.4	8.9	0.3	1.9	0.8
Len_400	9.8	2.2	2.3	6.2	3.5
Len_nsw_100	10.0	2.5	2.2	6.5	4.1
Len_nsw_200	10.0	1.6	1.9	6.6	3.2

the segments that start close to the actual jump-in point, while minimizing the amount of irrelevant content in each segment. This result opens another direction of discussion which is out of the scope of this research: whether users would prefer to examine a greater number of shorter segments while auditioning less non-relevant material, or to examine less segments but to spend more time auditioning non-relevant content.

Detailed example comparing results for single segmentation boundary condition with different transcript content

It is interesting to take a more detailed look into the changes that happen in the ranks of retrieved segments when the only difference underlying the ranking is the transcript input. We compare the asr_man and man results across all the segmentation methods. In order to compare these, we calculate the average number of changes in the rank of retrieved segments containing relevant content between the asr_man and the man runs for the 25 test queries. These results are presented in Table 6.2. Items retrieved in top 50 for the manual segments are divided into four groups: asr_man segment receives improved rank (a_m<m), it remains at the same level (a_m=m), it achieves a degraded rank (m<a_m≤50), it disappears from the top 50 items in asr_man (m≤50<a_m). In the last column we show the number

of segments that are retrieved in top 50 for asr_man transcript, but fall below the cutoff of 50 in the manual transcript (a_m≤50<m).

As only a small number of relevant segments stay at the same rank (0.3-3.0 segments on average across different segmentation methods), the vast majority of the relevant data changes its position. Since the errors in the ASR transcripts may cause the loss of content words important for the topic of the query, it is not surprising that half of the segments are at reduced rank in the retrieved list. However, it is important to note that most of the remainder of the items move up the list. If we choose a certain threshold (as we do with the rank 50 throughout this investigation), certain parts of the relevant information retrieved in the top 50 for the man run drop below this threshold for the asr_man run and are lost to the user, based on our assumption that they will not look beyond the top 50 ranked items. On average for all types of segmentation the number of asr_man segments that move up the list (a_m<m) is higher than the number at reduced rank (m<a_m≤50) or are lost after the set cutoff point (m≤50<a_m). Also, it should be noted that there is a certain amount of content that appears in the top 50 ranks for asr_man while not being retrieved in top 50 man at all (a_m≤50<m). This partially compensates for the amount of relevant content present in the man ranked lists that is lost in asr_man ranked lists (m≤50<a_m) when the ranks at which relevant content is retrieved are averaged.

Overall except for len_300, the average number of segments containing relevant material present in man (m<a_m≤50 and man<50<a_m) and asr_man (a_m<m and a_m≤50<m) are roughly comparable or higher, showing that as we would anticipate, on average man runs retrieve more segments with relevant content. However in the case of len_300 there are many more relevant segments retrieved in the asr_man list than lost ('a_m≤50<m' = 8.9 versus 'm≤50<a_m' = 0.8)). Also it should be noted that the ranks of relevant segments that move up the list for asr_man are usually lower than the ones for which the rank is reduced or which are lost. Overall this provides a clear explanation for the small reduction in MAP observed between

man and asr_man. We are not aware that this effect has been observed or properly examined in any previous SCR studies.

Examination of changes in rank of relevant content between man and asr_man transcripts

In this section we first examine the ranked list result for a single search request taken as an example, looking at changes in ranks between man and asr_man runs, and when different segmentation methods are used to preprocess the collection. We then show that the effects for this query can be seen over all the segmentation methods averaged across the whole test query set.

Detailed analysis of example query While Table 6.2 presents information about rank changes on average, we next show the specific changes in rank position of relevant content for our single search request. Figure 6.7 illustrates the connection between the number of changes in the rank of relevant content and the amount of relevant content which changes place in the example case for one query (query 21) and our overall preferred segmentation method (C99). The three parts of the figure depict the changes that happen within the top 50 results for man and asr_man runs. The left side (a) shows movements up and down the list for relevant items retrieved in the top 50 for both man and asr_man, the middle section (b) shows relevant items retrieved in the top 50 for man transcripts that drop out of the top 50 for asr_man, and the left side (c) shows relevant items promoted to the top 50 for asr_man which are below the top 50 man transcripts. In all three sections, next to the rank we provide information characterizing the segment: length in seconds (e.g., 105/125 if there is 105 seconds of relevant content in a segment that is 125 seconds long, or simply 105 if the segment contains only relevant material); WRR on this segment (e.g. 0.67); the sign "JP" means that the segment contains the jump-in point for the relevant content associated with this relevant region of a meeting.

Although the amount of relevant content with higher ranks in the asr_man run

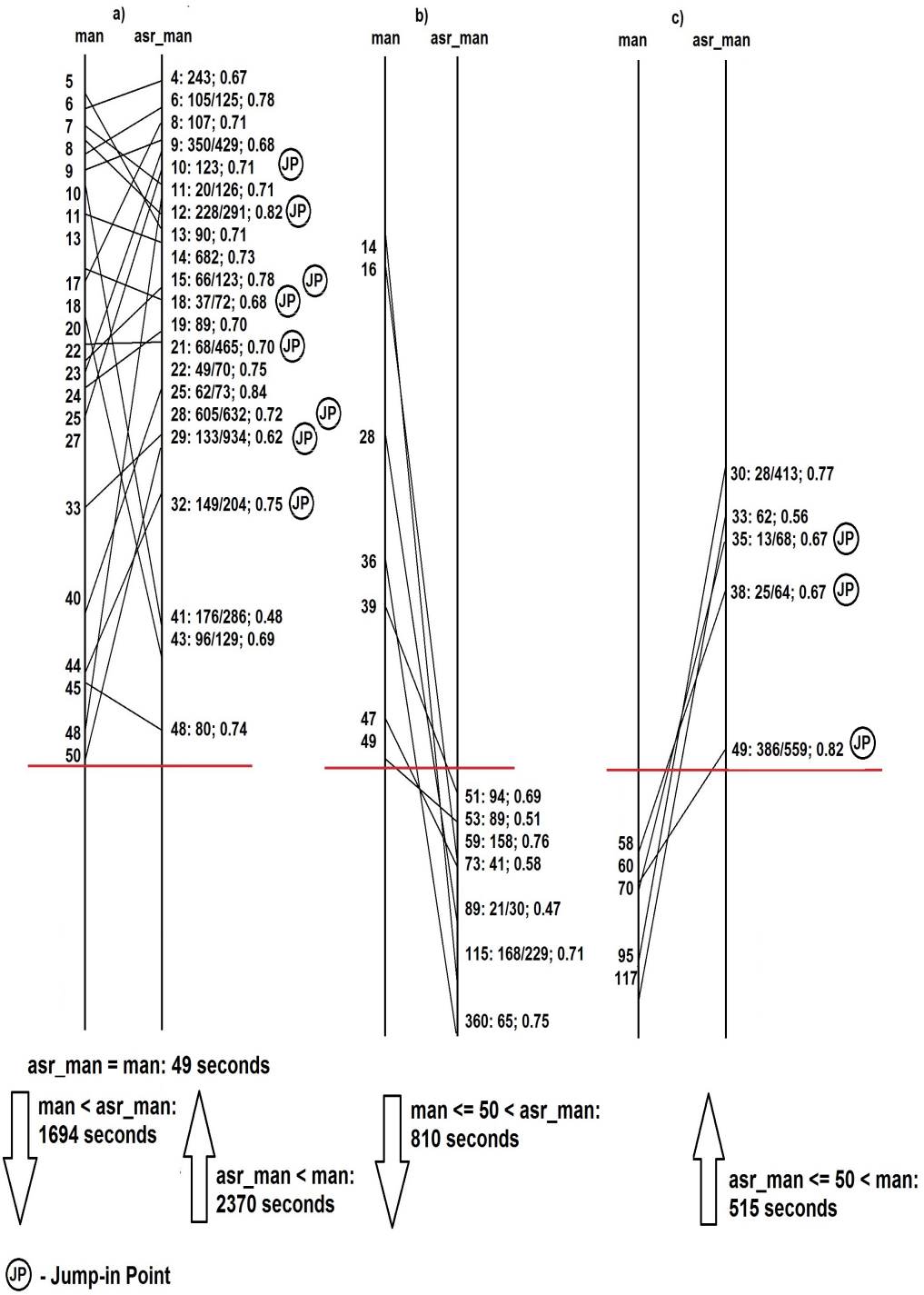


Figure 6.7: Rank changes within top 50 man and asr_man runs for C99 segmentation for query 21: a) changes within the top 50 (a_m < m, a_m =m, m < a_m < 50); b) segments present in the top 50 for the man run and falling below the top 50 in the asr_man (m < 50 < a_m) run; c) segments present in the top 50 of the asr_man run and falling below the top 50 in the man run.

within the top 50 is higher in absolute terms (2370 seconds versus 1694, as shown in Figure 6.7), the corresponding ranks in the manual run are higher in the list. Looking at columns (b) and (c) of Figure 6.7, we can see that the loss of a number of relevant segments below rank 50 between man and asr_man is to some extent compensated for by segments that were not found within the top 50 for the man run, but are promoted for the asr_man run. However both the amount of relevant content for asr_man that moves into the top 50 is lower than the amount lost (515 seconds versus 810 seconds) and the ranks at which relevant content is found are on average lower in the list.

For this example, especially in columns (b) and (c) we can see that the segments that move down the list have on average lower WRR (e.g., 0.51, 0.58, 0.47) than the WRR of the segments promoted in the top 50 of the asr_man run (0.67, 0.77, 0.82), the same trend can be distinguished in the changes that happen within column (a) for relevant segments present in both lists.

The segments that move up the list in columns (a) and (c) often contain the jump-in point for the relevant content. In the cases of the top ranks within asr_man (ranks 4, 6, 8, 9, 12, 14) the retrieved segments contain the longest part of the relevant content which is spread across several segments. Within the top 50 retrieved results of the manual run there are two cases where the relevant content is split into two parts by the segmentation methods. Both of these adjacent segments are present in the retrieved list (case 1: rank 8 (segment 2), rank 24 (segment 1), these change to the following positions in the asr_man rank 13 (segment 2), rank 9 (segment 1); case 2: rank 6 (segment 2), rank 45 (segment 1) that move to rank 4 (segment 2) and rank 48 (segment 1)). In both examples, the segments that have more relevant content and are longer (segment 1 in first case and segment 2 in the second case, 350 and 243 seconds of relevant content respectively) move up the list in the asr_man run, whereas the shorter adjacent segments (90 seconds and 80 seconds respectively) move down the list, even when the WRR is relatively high (0.71 ad 0.74 respectively).

Figure 6.8 allows us to compare the changes illustrated in Figure 6.7 for the seg-

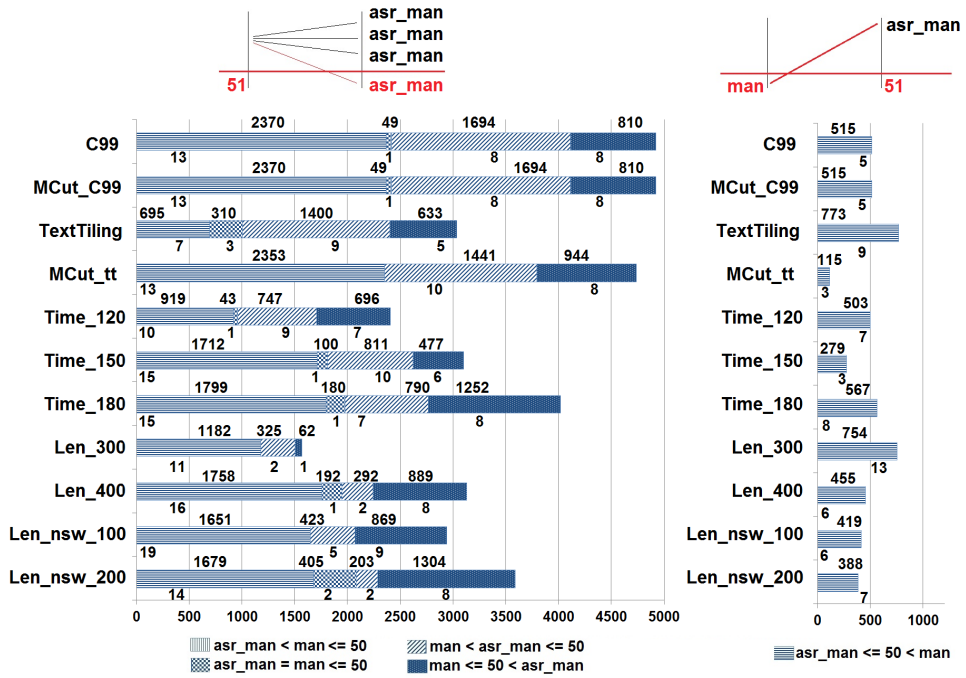


Figure 6.8: Length of the relevant content within different types of changes in ranks for man and asr_man runs, query 21

mentation methods exhibiting the most interesting behaviour in the earlier analysis on the development set. The left part of the figure represents the amount of relevant content that was present in the top 50 and which changes its position or stayed the same, including changes where the segment falls below rank 50. The right hand side shows the amount of relevant data that appears in the top 50 for the asr_man run which was not present in top 50 results for the man run. The numbers above each bar represent the length in seconds, and the ones below the number of segments that this content is present in for each type of rank position change.

Comparison between length-based algorithms and lexical coherence based algorithms shows that the latter outperform the former (the amount of relevant content present in the top 50 manual and not lost in the top 50 asr_man is higher for C99, mcut_C99 and mcut_tt than for the results of all length and time based runs).

In the case of length-based segmentations, runs with longer segments based runs contain more relevant content within the top 50 than ones with shorter segments of

Table 6.3: Total amount of relevant content and number of ranks with relevant content in top 50 for asr_man and man runs, query 21

Segmentation type	Relevant content (seconds)		Number of ranks	
	man	asr_man	man	asr_man
C99	4922.47	4627.42	30	27
MinimumCut_C99	4922.47	4627.42	30	27
TextTiling	3038.12	3177.8	24	28
MinimumCut_tt	4738.83	3910.02	31	26
Time_120	2405.34	2212.18	26	26
Time_150	3099.87	2901.78	32	29
Time_180	4020.13	3334.07	31	31
Len_nsw_100	2942.74	2493.10	33	30
Len_nsw_200	3591.35	2675.34	26	25
Len_300	1569.06	2260.88	14	26
Len_400	3131.47	2697.09	27	25

the same type: (len_nsw_200 (2287 sec) vs len_nsw_100 (2074 sec), and len_400 (2242 sec) vs len_300 (1507 sec). Although if we take all the content present in the top 50 of the man runs as 100% and look at the position changes of this content, then the shorter segments have better statistics: 56% and 75% of the content moving up the list in len_nsw_100 and len_300, against 47% and 56% for len_nsw_200 and len_400, 30% and 4% lost against 36% and 29% respectively. Only the amount of relevant content going down the list is better for longer segments: 6% and 9% for len_nsw_200 and len_400 against 14% and 21% for len_nsw_100 and len_300.

As we assume the scenario of a user looking for as much relevant information in the list as possible and not being willing to listen beyond a certain number of items in the list, we can see that segmentation runs that find more information in absolute length in time should be considered better than those that have better scores in terms of the percentage of position changes of improved rank or smaller reduction in the ranks of relevant content present in top 50 of the manual run against the asr_man run. Comparing len_300 to len_400: 1182 seconds of relevant content in the segments move up the list in the asr_man run vs. 1758 seconds, corresponding to 75 % vs 56% in percentage terms.

We can see that in general segments where rank is promoted in the case of

Table 6.4: Word recognition rate (WRR) using Porter stemming for segments with relevant content in the top 50 retrieved results, query 21

	a_m<m≤50	a_m≤50<m	a_m=m≤50	m<a_m≤50	m≤50<a_m
C99	0.74	0.70	0.75	0.70	0.64
MCut_C99	0.74	0.70	0.75	0.70	0.64
TextTiling	0.74	0.72	0.72	0.72	0.69
MCut_tt	0.73	0.67	–	0.71	0.67
Time_120	0.75	0.74	0.78	0.72	0.64
Time_150	0.75	0.76	0.69	0.70	0.64
Time_180	0.72	0.74	0.72	0.69	0.69
Len_300	0.73	0.76	–	0.70	0.49
Len_400	0.73	0.70	0.86	0.79	0.63
Len_nsw_100	0.74	0.75	–	0.72	0.69
Len_nsw_200	0.74	0.74	0.71	0.61	0.70

asr_man are in general longer, have high segment precision and high WRR. These features of such segments are to be expected. However, this does highlight the challenges of retrieving items whose rank is reduced in the asr_man run, which are typically shorter, and have lower segment precision and WRR.

Table 6.3 shows the total amount of relevant content present in man and asr_man top 50 ranked segments for each segmentation type, and the number of retrieval ranks where this relevant content is found. Only two segmentations (tt, len_300) have more relevant content in the asr_man than in the man runs in absolute numbers. This is explained by Figure 6.8, where we can see that only these two runs have a larger amount of relevant content moving up from below the list for the asr_man run relative to the man run, than that which falls below rank 50 in the man run (773 seconds versus 633 seconds for tt and 754 seconds versus 62 for len_300). This trend is consistent for most of the queries for len_300, i.e. more content moves up the list in the asr_man run than falls below rank 50 in the manual run. This is the reason for the unusual behaviour of the manual run of this segmentation when evaluated using all 4 metrics in Figures 6.2, 6.3, 6.4 and 6.5.

The analysis so far has examined the length of the relevant content which changes position, but another feature that is important to look at when analyzing the changing of ranks is the WRR of these segments. Table 6.4 shows the average WRR for

Table 6.5: Word recognition rate (WRR) using Porter stemming for segments with relevant content in the top 50 retrieved results, average across test set

type	a_m<m≤50	a_m≤50<m	a_m=m≤50	m<a_m≤50	m≤50<a_m
C99	0.76	0.65	0.75	0.75	0.68
MCut_C99	0.76	0.65	0.75	0.75	0.68
TextTiling	0.76	0.61	0.74	0.72	0.65
MCut_tt	0.76	0.60	0.75	0.71	0.65
Time_120	0.77	0.60	0.75	0.75	0.68
Time_150	0.75	0.63	0.73	0.75	0.68
Time_180	0.75	0.63	0.75	0.74	0.69
Len_300	0.76	0.65	0.73	0.75	0.67
Len_400	0.76	0.63	0.76	0.75	0.68
Len_nsw_100	0.76	0.63	0.77	0.75	0.66
Len_nsw_200	0.75	0.64	0.75	0.73	0.70

Table 6.6: Word error rate (WER) using Porter stemming for segments with relevant content in the top 50 retrieved results, average across test set

	a_m<m≤50	a_m≤50<m	a_m=m≤50	m<a≤50	m≤50<a_m
C99	0.26	0.26	0.32	0.30	0.36
MCut_C99	0.26	0.26	0.32	0.30	0.36
TextTiling	0.32	0.25	0.26	0.32	0.37
MCut_tt	0.28	0.24	0.27	0.29	0.31
Time_120	0.28	0.24	0.21	0.32	0.32
Time_150	0.29	0.27	0.23	0.33	0.34
Time_180	0.30	0.17	0.27	0.35	0.31
Len_nsw_100	0.32	0.28	0.21	0.32	0.36
Len_nsw_200	0.33	0.22	0.25	0.35	0.33
Len_300	0.32	0.30	0.12	0.29	0.24
Len_400	0.30	0.29	0.31	0.35	0.35

relevant segments according to the changes of ranks in asr_man compared to the man runs for the same example query. The WRR of the segments that fall in rank below top 50 in the asr_man (m≤50<a_m) is lower than the WRR of the relevant segments that compensate for this loss (a_m≤50<m) for all the segmentation methods. For all the segmentation methods except len_400, the average WRR of the segments that move up the list is higher than the average WRR of the segments that get lower ranks within top 50 of asr_man. Examining the reasons for the differences in the lists we find that segments retrieved at high rank for the asr_man segments generally have high WRR, this observation agrees with the findings reported in (Sanderson

and Shou, 2007). Consistent with this, segments which experience reduced rank generally have lower WRR. This effect is particularly noticeable when examining relevant segments which disappear from the top 50 ranked items. This poses a major challenge to achieving high recall in meeting search when some regions of the ASR transcripts of the meetings experience low WRRs.

WRR and WER of segments containing relevant content across the test set In the previous section we focused on the details of one query taken as an example. Table 6.5 shows that the behaviour for this query is consistent across the test query set. This table shows the overall average values that correspond to the transcript quality of the segments that change ranks between asr_man and man runs within the same segmentation method. The asr_man segments present within the top 50 ranks have higher WRR values than those that fall below the cut off.

Thoughout the investigation so far we have used the WRR metric because it better reflects the behaviour of the indexing and retrieval systems (the order of the words in the transcript is not important; only the stems of the non-stop words corresponding to the query that are recognized by the ASR system influence the results). However we also calculated word error rate (WER), the standard metric that is used to define the quality of ASR system output in speech recognition. Tables 6.5 and 6.6 show the WRR and WER averaged across the set of test queries. WRR of segments that stay within the top 50 rank positions (a_m<m$\leq$50, a_m=m$\leq$50, m<a_m$\leq$50) is higher than that of the ones which fall below the top 50 cut off rank for one of the transcript types (a_m$\leq$50<m, m$\leq$50<a_m). In the case of m$\leq$50<a_m, asr_man segments that fall below the top 50, do so because important content words have been misrecognized. The case of a_m$\leq$50<m illustrates the situation where non relevant content has low WRR, thus it falls lower in the ranked list allowing the relevant asr_man content to move up the list. WER values have a more straightforward relation with the rank changes: the asr_man segments that are higher than man segments in the list (a_m<m$\leq$50, a_m$\leq$50<m) have lower WER

than the asr_man segments that are lower in rank than their manual counterpart (m<a≤50, m≤50<asr_man). This trend illustrates that the cut off value of 50 does not affect the trend observed in the results.

6.2.3 AMI test example: summary and discussion

In previous sections we have described our initial extensive investigation into retrieval of transcripts of multi-party meetings based on the AMI corpus. We carried out an investigation into the segmentation of transcripts using the C99 algorithm and other methods to provide suitable retrieval units. Our examination of results of the outputs of these segmentation methods shows significant variations between segmentation behaviour for manual and ASR transcripts of the meetings. Initial investigation beyond this study reveals that these differences arise from the impact of word recognition errors on topical matching between related sentences and the resulting topic boundary decisions of the segmentation algorithms, further examination of this issue is beyond the scope of this thesis.

As we discussed in Chapter 5, SCR results require an evaluation based on a combination of different metrics to explore the retrieval behaviour of SCR systems. Use of multiple metrics in this way enabled us to address different facets of system behaviour in terms of efficiency in locating relevant content in temporal spoken content.

Experiments with this retrieval collection show that the lexical cohesion based segmentation methods perform consistently better compared to length- and time-based segmentation methods when the set of diverse metrics are used to evaluate the retrieved results. C99 outperforms both TextTiling and the Minimum Cut method that uses the number of segments from TextTiling as input. However detailed analysis of the behaviour of the segments shows that retrieval with the C99 segmentation method would need to be augmented to take into account word recognition errors in order achieve more reliable retrieval behaviour.

Overall we can see that longer segments achieve superior MAP at the expense

of time consuming browsing of non-relevant content. By contrast, shorter segments reduce the browsing time, but more segments must be explored to find the relevant content. ASR errors make the ranking of short documents less reliable, while content redundancy can compensate for this in longer segments.

Thus, ideally segments should have the highest possible precision within the segment while being as long as possible to help compensate for word recognition errors. As our results illustrate, extracting such segments is a very challenging task, results of our investigations and other studies examining topical segmentation of standard text, show that current segmentation methods are often inaccurate.

We continue our investigation into the use of lexical cohesion based segmentation methods for SCR in the next section, as we are interested to trace whether there is a consistency in SCR system behaviour across different types of data and languages.

6.3 Analysis of SCR for lecture retrieval using the NTCIR-9 SpokenDoc collection

The NTCIR-9 SpokenDoc test collection represents another type of data, i.e. lectures in contrast to the meetings as in case of the AMI corpus, as well as different language, i.e. Japanese versus English. In the following sections we describe our experiments when using the lexical cohesion based segmentation methods on this different type of data, and the retrieval results in context of official NTCIR metrics and precision/recall information.

6.3.1 Preprocessing and segmentation of the NTCIR corpus and queries

The NTCIR-9 Spoken-Doc dataset, as described in overview in Section 4.1, is provided with n-best word-based and syllable-based automatic recognition transcriptions of the lectures. For our experiemnts, we use only the 1-best word-based tran-

scripts and manual transcript of the corresponding lectures taken from the Corpus of Spontaneous Japanese (Maekawa et al., 2000).

Transcript and queries preprocessing details

Since in Japanese the individual morphemes of the sentences need to be recognized for further processing. We used the ChaSen system, version 2.4.0[2], based on the Japanese morphological analyzer JUMAN, version 2.0, with ipadic grammar, version 2.7.0, to extract the words from the sentences in ASR and manual transcripts. ChaSen provides both conjugated and base forms of the word, for later processing we used the latter since it avoids the need for stemming of different words forms. In order to have a consistent representation of all the data, we follow the same preprocessing procedure for the query set.

Text Segmentation details

Following our investigation on the AMI corpus example described in the previous sections, for the NTCIR corpus we focus on the segmentation of the transcripts into topically coherent passages to be used as retrieval units. Our objective here is to explore the use of segment units to retrieve relevant content on the assumption that these units will capture relevant passages. As C99 and TextTiling do not depend additional input parameters as Minimum Cut, we explored the application of these two segmentation algorithms.

Both algorithms work with the fundamental unit of the sentence placing segment boundaries between the end of one sentence and the start of the next one. Since the ASR transcripts did not contain punctuation, we considered each Inter-Pausal Unit (IPU) to be a sentence on its own. We ran the segmentation algorithms on both ASR and manual transcripts, and on the ASR transcript when stop words had been removed from the text[3] (asr_nsw).

[2]http://chasen-legacy.sourceforge.jp

[3]Stop words were taken from SpeedBlog Japanese Stop-words: `dnnspeedblog.com/SpeedBlog/PostID/3187/Japanese-Stop-words`

Transcript type	Segmentation type	uMAP	pwMAP	fMAP
BASELINE		0.0670	0.0520	0.0536
manual	tt	0.0859	0.0429	0.0500
manual	C99	0.0713	0.0209	0.0168
ASR	tt	0.0490	0.0329	0.0308
ASR	C99	0.0469	0.0166	0.0123
ASR_nsw	tt	0.0312	0.0141	0.0174
ASR_nsw	C99	0.0316	0.0138	0.0120

Table 6.7: Scores for official metrics

6.3.2 Experimental Results and Analysis

The segments obtained using each segmentation technique from the manual transcripts were indexed for search using a version of the SMART information retrieval system[4] extended to use language modelling (a multinomial model with Jelinek-Mercer smoothing) with a uniform document prior probability (Hiemstra, 2001), as described in Section 2.1.2. Separate retrieval runs were carried out for each topic for each segmentation scheme for segments created from the manual and ASR transcripts. In this section we present results and analyse for these experiments.

Analysis of experimental results

The official evaluation metrics for this task are variations of the standard MAP. As described in Section 5.2.1, these are applied to the list of the retrieved items after expanding the retrieved passages into IPUs. In case of the uMAP metric, relevance is assigned to individual IPUs in a relevant region of the lecture, where uMAP is calculated for relevant segments at the level of IPUs. For pwMAP, relevance is assigned to the whole passage retrieved at a certain rank if its centre IPU is part of the relevant content, the score is then calculated for retrieved passages classified as relevant according to this criteria. Recall of a passage and precision up to its rank at IPU level are taken into consideration in the fMAP calculation.

Table 6.7 shows our experimental results for these metrics along with the base-

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

line scores provided by the task organisers. It can be seen that, as would be expected, runs using manual transcripts show better results than those based on ASR trascripts. However manual runs outperform the baseline only for one metric (uMAP): 0.0859 and 0.0713 for TextTiling and C99 respectively versus 0.0670. It can be seen that transcript segmentation using TextTiling consistently achieves higher scores than segmentation using the C99 algorithm for all types of transcript.

The remainder of this section provides a more detailed analysis of our results for each of the evaluation metrics.

uMAP Results The uMAP metric calculates MAP on the level of IPUs after each retrieved passage has been expanded into its constituent IPUs and they have been rearranged so that the relevant IPUs are at the beginning of the sequence.

In order to better understand the relationship between our retrieved segments and the amount of relevant content that we had actually retrieved, we calculated the precision of the content for each retrieved segment which contained at least one relevant IPU, as described in Section 5.2.2. We then calculated the average of these precision values for each topic and then the average of these values across the completed topic set. Although we process the transcripts and return as output the numbers of start and end IPUs (passages), our ultimate goal is to provide the user with segments to listen to. Therefore the actual timing information of the beginning and end points of relevant data are important for the analysis of results. This is especially true since IPUs may differ considerably in time length and this is not included by any of the metrics. Thus the precision value of each segment was calculated using the length in time for each IPU unit provided with the ASR transcript.

Figure 6.9 shows these averaged values for both TextTiling and C99 for manual, ASR and ASR with stop words removed transcripts. From these results it can be seen that, similar to the official results in Table 6.7, TextTiling outperforms C99 in all cases. Comparing the results for the three different transcripts in each case, no

clear trend emerges in terms of precision of the contents of the individual segments, which is perhaps a little surprising, since the results in Table 6.7 show a clear trend that manual transcripts outperform ASR with respect to uMAP which outperforms ASR without stop words.

pwMAP Results pwMAP metric counts as relevant only segments for which the IPU in the centre of the segment is relevant. The results in Table 6.7 show that none of our methods was competitive with the provided baseline result with respect to pwMAP. This contrasts with the uMAP results, and indicates that although we are able to retrieve similar amounts of relevant content at similar ranks, the content segmentation methods that we are applying do not reliably place relevant content at the centre of the retrieved segments.

In order to analyze the scores further, we calculated the number of the segments in each run that were counted by the metric as relevant and the ones that had relevant content, but where it was not located in the centre of the retrieved segment and was therefore overlooked by the pwMAP metric. Figure 6.10 shows the average numbers of these relevant captured and relevant non-captured retrieved segments. From the figure, it can be seen that the runs on the manual transcript (manual_tt and manual_c99) contain more segments with relevant content. All of the runs using TextTiling segmentation (manual_tt, ASR_tt, ASR_tt_nsw) have more retrieved segments with relevant content that are included in the pwMAP score than C99 segmentation runs. This means that in general TextTiling segmentation is more likely to have the relevant content in the centre of the retrieved segment than C99 segmentation, and that thus the boundaries formed using TextTiling are not just more effective for retrieval of relevant content, but are more likely to place the relevant content towards the centre of the segment. However, it should be noted that in all cases the proportion of segments containing some relevant content, but where it is not in the centre of the segment is very high.

Since the pwMAP metric is based on standard MAP, it gives higher scores to

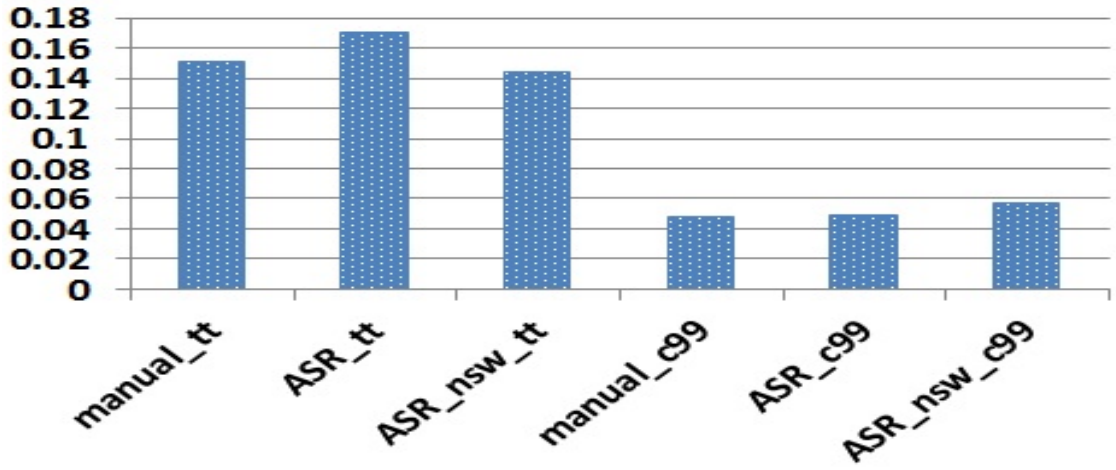


Figure 6.9: Average of Precision for all passages with relevant content.

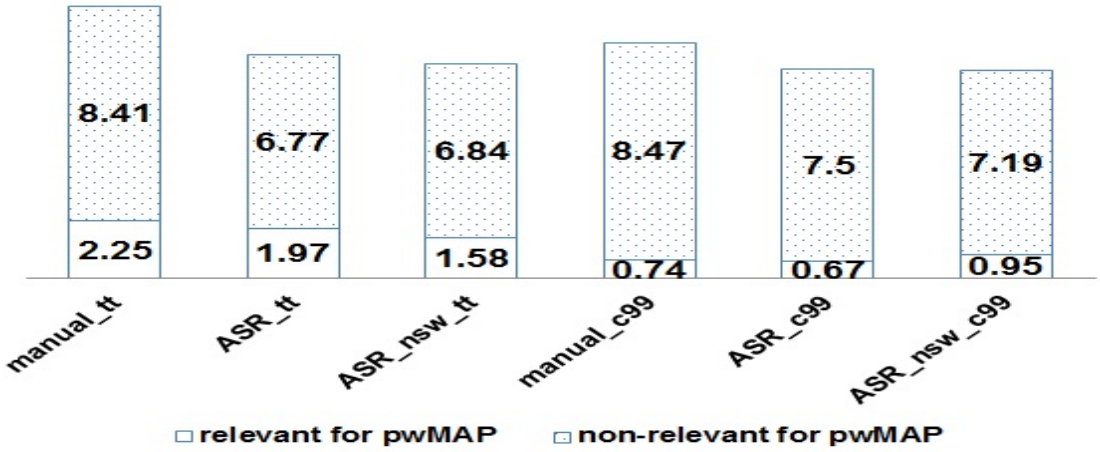


Figure 6.10: Number of ranks with relevant content that are taken or not taken into account for calculation pwMAP

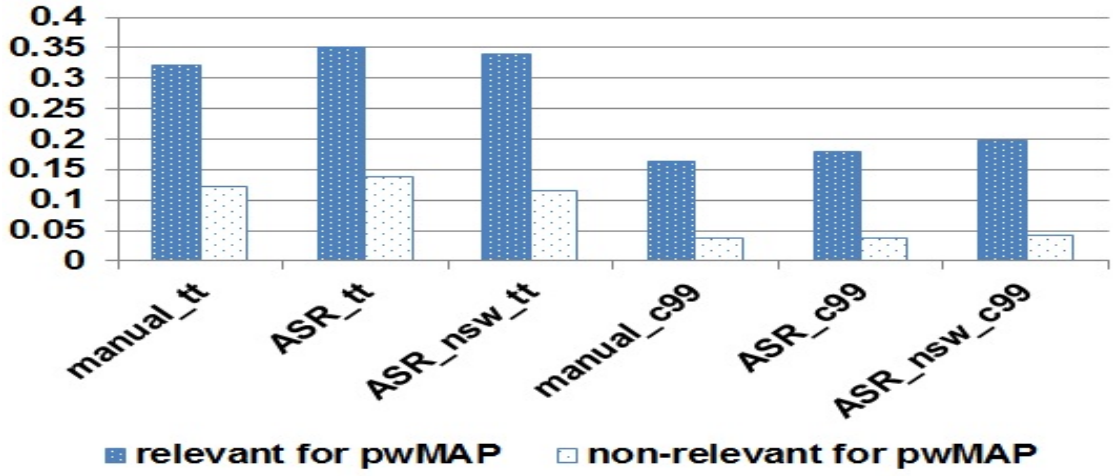


Figure 6.11: Average of Precision for the passages with relevant content that are taken or not taken into account for calculation pwMAP

techniques that place relevant documents higher in the ranked list. Therefore a larger number of retrieved segments containing relevant content does not automatically imply that the run will be scored better. The pwMAP scores of the runs using TextTiling segmentation on manual and ASR transcripts have better rankings than all the other methods, including C99 segmentation of the manual transcript. The same trend exists between the C99 runs: the average number of retrieved segments considered relevant for each topic using C99 segmentation is the highest for ASR_nsw, but apparently the rank of the relevant passages is better for both manual and standard ASR transcripts, since their pwMAP values are higher, suggesting that ASR_nsw is the worst one in terms of content ranking.

Comparing the numbers of retrieved segments containing relevant information and the breakdown by content included and not included in pwMAP calculations in Figure 6.10, it can be seen that while TextTiling and C99 segmentation retrieve similar numbers of segments containing relevant content, that the number of included segments is much lower in the case of C99. This indicates that the balance of many of these segments is poor, i.e. that they are not centred on relevant material. Looking at this finding in the context of Figure 6.9, we can see that poor segmentation in this way correlates with the rankings of relevant segments even where all available segments containing relevant content are taken into account when calculating uMAP.

Figure 6.11 shows the average of the precision of segment content for segments averaged across the topic set, counted as relevant for the pwMAP calculation and those not included by pwMAP. It can be seen that on average the precision is much higher in all cases for segments which are included in the pwMAP calculation than those which are not. This could be expected since segments for which the central IPU is not relevant are likely to have lower precision on average than those for which the central IPU is relevant. It can further be noted that all results for segmentation using TextTiling are superior to the corresponding results for C99 segmentation. Precision for the passages that have the relevant segment in the middle is always more than twice as high as that for passages that do not. Again these results indicate

Run	Rel Length		Total Length	
	centre	non centre	centre	non centre
manual_tt	83.65	142.06	260.73	1153.38
ASR_tt	73.88	110.53	210.01	805.26
ASR_nsw_tt	71.99	117.59	212.19	1026.69
manual_c99	60.45	171.32	372.34	4710.93
ASR_c99	59.57	154.46	332.99	4203.04
ASR_nsw_c99	65.36	144.87	332.48	3496.67

Table 6.8: Average relevant and total length of segments with relevant central IPU and segments with non-centered relevant content (in seconds)

that these segments are associated with segments which are topically consistent, as measured against their relevance to the topics. Looking again at Figure 6.9, this further emphasizes the role of good segmentation in superior ranking in retrieval as measured by uMAP.

fMAP Results The fMAP metric is designed to capture the relevancy of the segments. In this evaluation none of our segmentation methods outperformed the baseline, shown in Table 6.7. This result is probably caused by the low precision of the segments containing relevant content, as observed in Figure 6.9 (low average of precision) and in Figures 6.10 and 6.11 (where the number of the segments having lower precision due to the fact that the relevant content is not located in the centre of the segment, is considerably higher than the number of the segments with centered relevant content). It is interesting to note that for this metric TextTiling segmentation not only shows better results for each of the same transcript types as C99, but even its ASR transcript outperforms C99 scores for the manual transcript.

To calculate the fMAP score, precision and relevance is counted in IPU units. Following the same reasoning as in Section 6.3.2 when calculating the average of precision from the user perspective, the actual length of the segments which must be auditioned is important, we decided to look at the precision in terms of the length in time (in seconds). Table 6.8 shows the average lengths of relevant content retrieved per topic in each run and the average of the total length of the passages containing the relevant content per topic. We keep the distinction between the segments with

a relevant central IPU and with non centred relevant content. The average lengths of the relevant content for segments with relevant central IPU are figures of the same order for both segmentation schemes, with TextTiling segmentation runs being slightly higher. In the case of non-centred relevant IPU, C99 segmentation runs have longer relevant content than TextTiling ones. The total average lengths of the relevant content retrieved in the list is higher for all C99 runs. Unfortunately due to less accurate segmentation, retrieving more relevant content is correlated with having much longer segments: the total lengths of C99 segmentation runs are considerably higher than the TextTiling ones and therefore a metric focused on precision gets lower scores for the C99 segmentation runs. Also they contain more non-relevant content and are thus likely to be ranked more unreliably as observed in the uMAP results in Figure 6.9.

6.3.3 Summary of NTCIR-9 SpokenDoc Findings

Our experiments show that for the task of retrieving passages from the Japanese lecture archive, TextTiling segmentation is a more suitable algorithm than C99 for preprocessing the data collection in order to obtain retrieval units better corresponding to the actual relevant content, confirming our assumption that there is no single segmentation method that addresses all possible challenges of the SCR task. However, these results are consistent with results and analysis for the AMI SCR collection reported in the previous section, and confirms that better precision of the content within the retrieval unit is important for SCR effectiveness. In the next section we report a simple analysis for the MediaEval 2011 Rich Speech Retrieval task.

6.4 Analysis of SCR for Semi-Professional User generated content in the MediaEval Rich Speech Retrieval task

Following the exploration of the SCR behaviour on varying type of conversational spoken content, we next examine the less controlled and more free example of SCR for Internet TV data. The Rich Speech Retrieval (RSR) task at MediaEval 2011 (Larson et al., 2011) used a subset of the BlipTV corpus, described in Section 4.3, and a different set of queries, though created in the same way, through the crowdsourcing procedure (Eskevich et al., 2012b). As organisers of the task we carried out extensive analysis of the submissions following similar procedures to those used for our previous investigations described earlier in this chapter. In this section we overview our findings on the MediaEval 2011 RSR task that highlight its consistency with our hypothesis of the importance of precision and recall of the relevant content within retrieval unit.[5]

6.4.1 Details of the ME10WWW corpus

The RSR dataset includes 1727 videos that are predominantly in English. These were segmented using the following approaches: lexical cohesion based approaches (C99 and TextTiling) (Eskevich and Jones, 2011b), speech segments provided by the transcripts (Aly et al., 2011), and a sliding window of a number of words (20-40 words) (Wartena and Larson, 2011). These runs are further referred to as LC_asr_{segmentation method (c99 or tt)}, Sp_asr, and SW_asr respectively, with {meta} standing for additional use of metadata.

Within these experiments we also addressed the question of potential for using the metadata provided with the videos as source for additional information to help improve SCR effectiveness. This metadata as assigned to a video by its uploader.

[5]This section draws on joint work reported in (Eskevich et al., 2012c)

RunName	WindowSize					
	60		30		10	
	MRR	mGAP	MRR	mGAP	MRR	mGAP
LC_asr_c99	0.28	0.19	0.21	0.15	0.07	0.07
LC_asr_meta_c99	0.33	0.21	0.23	0.16	0.08	0.08
LC_asr_tt	0.36	0.25	0.29	0.18	0.09	0.09
LC_asr_meta_tt	0.39	0.28	0.30	0.20	0.14	0.14
Sp_asr	0.34	0.27	0.27	0.22	0.16	0.16
Sp_asr_meta	0.34	0.25	0.26	0.21	0.15	0.15
SW_asr	0.38	0.30	0.34	0.22	0.10	0.10
SW_asr_meta	0.39	0.33	0.39	0.28	0.15	0.15
SW_asr_sh	0.37	0.32	0.32	0.27	0.19	0.19

Table 6.9: Mean Reciprocal Rank (MRR) and mean Generalized Average Precision (mGAP)

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

Table 6.10: Average Precision and Recall, Window size = 60 sec

We observed that on this dataset the difference between runs using ASR transcript only and combining it with metadata use is not statistically significant for any of the retrieval frameworks.

The Blip10000 collection does not provide manual transcripts. However, as the workers who chose the query and assigned its relevant segment, were also asked to type in what was pronounced within the relevant segment, as described in Section 4.3, we have the manual transcript for these segments. This allows us to calculate and use the information about segment WER and OOV in our results analysis.

6.4.2 Metrics scores comparison

Table 6.9 shows MRR and mGAP scores for all the runs with different window size (60, 30, and 10 seconds). As expected the smaller window size decreases the scores, however the trend of difference between approaches stays the same. Runs that have a larger drop between MRR and mGAP have the start of the segment with relevant content further from the jump-in point. For example MRR for LC_asr_tt and SW_asr_sh is 0.36 and 0.37 (window size = 60s), but mGAP is 0.25 and 0.32, meaning that the second run has segments closer to the jump-in points.

Table 6.10 presents average precision and recall for the runs. The mGAP metric was designed to reward approaches that retrieve the beginning of the relevant segment better, the runs that have higher mGAP values have higher precision and lower recall values. However the analysis of the results per query shows that the recall within the segment is likewise important for the ranking of the segment containing relevant content.

Topic segmentation effects

The words present in the relevant content or metadata do not always overlap with the query terms. On average the overlap with the manual transcripts is 0.30, with ASR transcript - 0.25, and with metadata - 0.22 (after standard stopword removal). 19 queries out of 30 were found not to have any overlap at all with the ASR transcripts, and 15 had no overlap with the manual transcripts, while most of the queries (27 out of 30) that have an overlap with metadata attached to the document containing the relevant passage.

Non-zero overlap of query vocabulary of content words with ASR transcript In case of runs with non-zero overlap of the query vocabulary and relevant content vocabulary, there is no direct correlation between ASR WER and retrieval results, because the good ranking of the segment containing relevant content depends on the good topic segmentation around relevant area: if the non-relevant

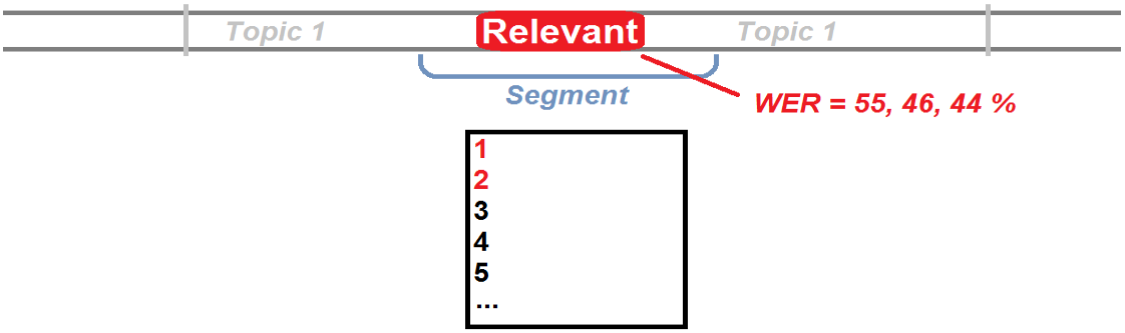


Figure 6.12: Illustrative example of the relationship between retrieval effectiveness and segmentation methods: all relevant content within one segmental unit with high WER.

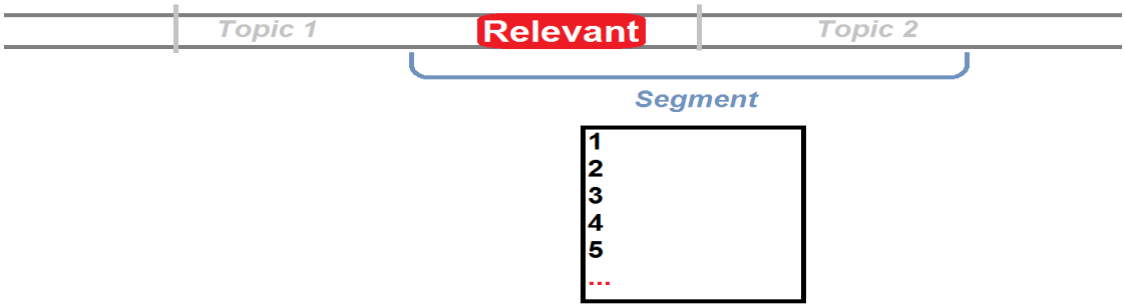


Figure 6.13: Illustrative example of the relationship between retrieval effectiveness and segmentation methods: all relevant content within one segmental unit that contains also content on another topic.

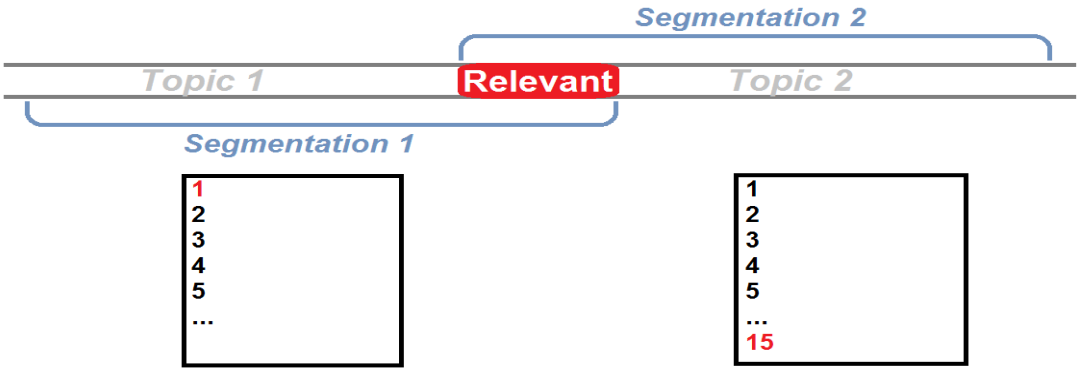


Figure 6.14: Illustrative example of the relationship between retrieval effectiveness and segmentation methods: all relevant content within different segmental unit in two segmentation approaches.

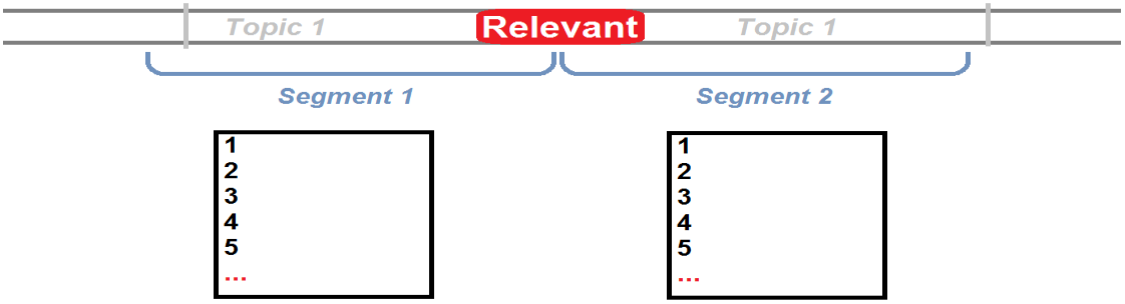


Figure 6.15: Illustrative example of the relationship between retrieval effectiveness and segmentation methods: relevant content is split into two adjacent segments.

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

Table 6.11: Example of MRR, Precision, Recall results for queries with different ASR WER

content present in the retrieved segment belongs to the same topic, even segments with ASR WER = 55, 46, 44 % are retrieved at the 1st or 2nd rank, cf. Figure 6.12, whereas when the same relevant content is contained in a segment that also contains a part of the transcript on another topic, its retrieval rank is much lower, cf. Figure 6.13. For example, cf. Table 6.11, for query 36 with WER = 46 %, LC_asr_tt and SW_asr runs have 100 % recall, a high level of precision (30 and 56 %), and retrieve the relevant content at the 1st rank, whereas segments created for LC_asr_c99 and Sp_asr, though containing all the relevant content, have lower precision and expand

to cover a different topic, thus the segment is lower down in the list. The case of query 6, WER = 62 %, is even more representative because SW_asr, LC_asr_c99, and LC_asr_tt runs have the same level of precision (23 %), although the non-relevant content is after the relevant segment in the case of SW_asr run, and before it in the case of the LC_asr_c99 and LC_asr_tt runs, meaning that segments of the SW and LC runs overlap only within the relevant content, cf. Figure 6.14. For the SW_asr segment, the non-relevant part contains a change of the topic, and this causes the drop in retrieval rank to 15th position. The same trend is observed for the queries with much lower ASR WER (even when it is equal to 0.0. as in case of query 24).

Query 24 shows another effect of transcript segmentation on the ranking of the results. When the relevant content is divided between two segments, it influences the retrieval process, cf. Figure 6.15. LC_asr_c99 segment will be counted in the mGAP metric (because the segment starts within a window of 30 seconds), however its ranking is affected by non-100% recall of the segment.

Non-overlap of query vocabulary of content words with ASR transcript There are 10 queries that do not have any overlap of stemmed content words with both metadata and ASR transcript, and 9 queries that have an overlap with the metadata and not the ASR transcript. For the first type of queries retrieval results highly depend on the segmentation of the area surrounding the relevant content. If the topic of the discussion stays similar or the same, these segments are usually retrieved at the top of the list, and then if these segment are within the window used for mGAP metric, they are taken into account, although precision and recall are 0 %. Results of the runs for the queries of the second type are not affected by the use of metadata when the surrounding segments have another topic or have different vocabulary from the query. In cases when the surrounding queries are on the same topic, and fit within the mGAP window, adding metadata decreases the results.

Non-zero overlap of query vocabulary of content words with both ASR transcript and metadata In these cases runs that have high recall of relevant

content in broader segments that have a topic within them, are ranked better when metadata is added to the index, see for example run SW_asr_meta versus SW_asr for query 6.

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

6.4.3 Summary of MediaEval 2011 RSR Results and Analysis

In this section we described a comparative investigation of the retrieval behaviour of different systems for an Internet video dataset. These experiments add to the evidence that segmentation of the content plays the main role in retrieving the relevant content. When the segments have high recall and precision, and the rest of the segment belongs to the same topic, all ranking methods tend to rank relevant items on the top of the list. Also, textual metadata can be useful when the segment (with high recall and non relevant content, or with low recall) would otherwise be ranked low in the retrieved list.

6.5 Summary

In this chapter we overviewed SCR experiments that serve as examples of trends across multiple types of content and retrieval tasks: precision and recall of the relevant content within the segmentation unit play an essential role in its efficient retrieval, no matter what ranking scheme is used as the IR model.

A detailed example on the AMI data that compares the changes in the ranking of a number of relevant segments showed that the ASR errors exerts more complicated influence on averaged retrieval behaviour than only decreasing the ranking of the relevant segments. As the segments containing nonrelevant content may be affected

by ASR errors, they may be moved lower in the result list, thus allowing segments with relevant content that would otherwise be lower to move up within the list.

These findings define the strategies of the retrieval experiments that are introduced in the following chapters: to target higher ranking of relevant content through segmentation adjustement in terms of relevant content precision, to explore the segment combination to improve ranking for each individual query, and to investigate the possibilities of document expansion that can help to deal with ASR transcript errors.

Chapter 7

Filtering of Overlapping Ranked Results and Segment Boundary Adjustment

Following on from the findings and analysis in the previous chapter, in this chapter we report experiments that aim to examine whether the use of overlapping segments combined with subsequent filtering approaches adjusted to each query, can improve overall ranking of relevant segments and proximity to their ideal jump-in points. We also investigate whether these boundaries can be further adjusted at a more detailed level using additional features such as the position of pauses or certain words within the sentences.

We run SCR experiments on two datasets: the NTCIR-9 SpokenDoc collection and our AMI corpus collection, since they both have ASR and manual transcripts that enable us to compare evaluation of the real case (ASR transcripts) and ideal case (manual transcripts). Both these datasets are structured in a way that enables us to carry out segmentation in units of fixed length with an overlap step.

The NTCIR collection is split into interpausal units (IPUs) (Akiba et al., 2011) that can be regarded as sentences for dataset segmentation, and we evaluate the results using IPU-based metrics (uMAP, pwMAP, fMAP), introduced in Section

5.2.1.

Since the AMI corpus collection is provided with a transcript that includes time stamps for words, but lacks sentence punctuation, we carry out the dataset segmentation in terms of temporal units. The presence of word boundary information permits further experiments that evaluate closeness to the actual jump-in point in the retrieved segment with imperfect start time. Thus we explore the use of pauses and energy peaks in the audio to help to identify potential ideal jump-in points for the relevant content. As lexical cohesion based methods are based on the actual content of the transcripts, we investigate another approach to adjust the boundaries of segments of fixed length by using the boundaries for the segments in this region that are created using the lexical cohesion based segmentation methods. Since we base segmentation and boundary adjustment on the time information, at the evaluation stage we use not only MAP, Section 3.1, but also such time-based metrics as mGAP, introduced in Section 3.2.3, and MASP and MASDWP introduced in Section 5.2.3.

An overview of filtering methods and boundary adjustment, that will be further described and discussed in this chapter, is shown in Table 7.1. The main idea is to start with different types of initial segmentation and to carry out post-processing of the retrieval results by: i) filtering the result list, in cases when initially the collection is represented with overlapping segments, and ii) adjusting the result segment boundaries using various sources of information (other types of segmentation, acoustic information about pauses or loudness). Each line in Table 7.1 represents potential segment ranking and boundary adjustments that are implemented for the initial segments, e.g. 'time_overlap' means that all the content is segmented in units of the same length in seconds, and '+' in each column means that the results of the run are modified first by one of the filtering approaches using removal or combination, and second using boundary adjustment.

Initial segmentation type	Target: to improve ranking		Target: to improve closeness to JP (Boundary adjustment for JP)			
	Filtering		Lexical cohesion based	Use of pauses		Use of loudness (energy peaks)
	overlap removal "RemSeg"	overlap combined "CombSeg"		first	longest	
lexical cohesion based	–	–	–	+	+	+
len	–	–	+	+	+	+
len_nsw	–	–	+	+	+	+
time_overlap	+	+	+	+	+	+

Table 7.1: General framework of the experiments that target ranking improvement via filtering and jump-in point (JP) closeness via boundary adjustment.

7.1 Filtering approaches for segmentation with sliding window

Previous research has shown that the use of sliding windows to segment a speech transcript into segments that are close to the length of the target segments with further filtering of the overlap in the result list can achieve higher evaluation scores (MAP, mGAP, MASP, IPU-based metrics) as compared to runs based on lexical cohesion or turn of speakers (Wartena and Larson, 2011; Wartena, 2012; Akiba et al., 2011). This result is achieved when the target units are of approximate known length (e.g. where Rich Speech Retrieval relevant items were known to be no longer than 60 seconds, methods could be tuned to this value (Larson et al., 2011; Wartena and Larson, 2011)), therefore segmentation with overlapping windows into units of this tuned length enables creation of segments with higher internal precision, and recall of relevant content which increases the possibility of these segments being retrieved higher in the result list in a subsequent retrieval phase, as discussed in Chapter 6. However, in the case where the queries require varying amounts of information to be retrieved, as they differ in the level of specificity, this approach appears less useful, since it is impossible to define what size the segments and sliding window should be to achieve the same level of retrieval effectiveness across a set of queries.

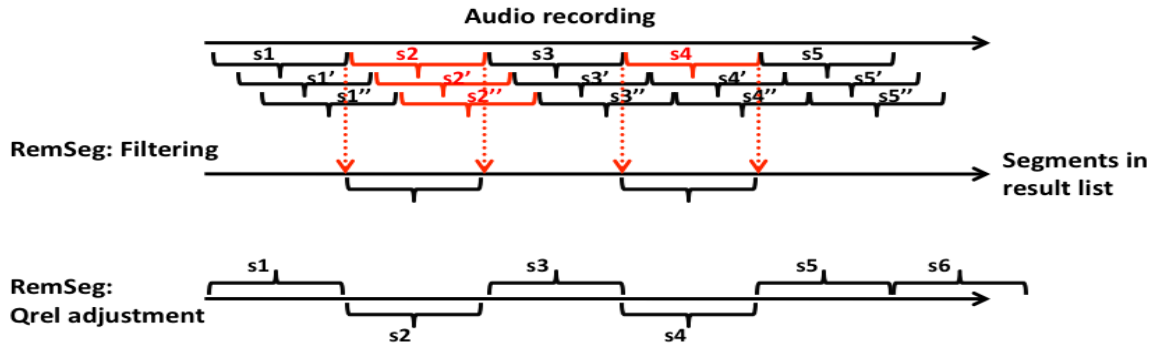


Figure 7.1: Segmentation adjustment per query within qrel after type RemSeg filtering approach (removal of overlapping segments at the lower ranks)

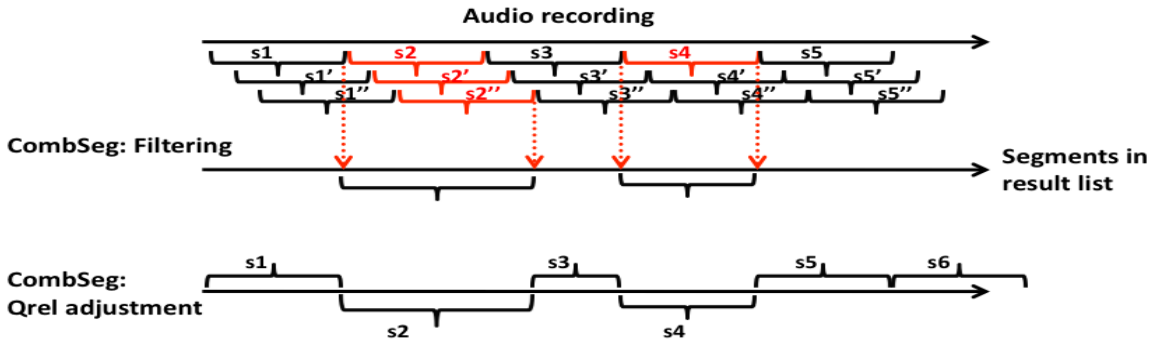


Figure 7.2: Segmentation adjustment per query within qrel after type CombSeg filtering (combination of overlapping segments into longer single ones)

When fixed length segments with a sliding window are used for collection segmentation, we increase the probability of creation of segments with a level of precision and recall of the relevant content within each segment being high enough to help the segment to be retrieved at high rank. At the same time as these segments are of the same length across the collection, they might be used more as anchors that define the region of potential relevance with its boundaries to be adjusted individually for each query. With this assumption we implement two filtering approaches that target either the creation of a result list that contains the regions of relevant information, but does not contain an abundance of long non-relevant content, or one that targets finding regions of relevance that are longer than the initial retrieval units, and hence may contain increased amounts of non-relevant content in addition to the relevant material. These filtering methods and their evaluation are described in the following sections.

7.1.1 Filtering methods

We implement two filtering approaches on retrieved potentially overlapping segments. These are illustrated in Figures 7.1 and 7.2, which show the segments that appear in the result list after filtering. We denote the filtering schemes *RemSeg* and *CombSeg* defined as follows:

- *RemSeg*: Delete overlapping segments that are present at lower ranks in the result list. This approach normalizes the results list, as all segments continue to be of the same length. The user is not forced to listen to audio segments from the same region that will already have been presented to them earlier in the result list.

- *CombSeg*: Combine overlapping segments that are present in the result list into one single segment and place it at the rank of the appearance of highest rank amongst the segments of which it is composed. With this filtering approach the segments in the result list differ in length. Our assumption is that retrieval of overlapping segments in the list shows that the boundaries should be enlarged within this file to create segments that are longer than initial segmentation units, as they are semantically coherent, and the presence of multiple overlapping regions in the retrieved result list indicates an increased chance of relevance of this material.

7.1.2 Evaluation of filtered runs

Changes imposed on the segments by a filtering process may require adjustment in the evaluation framework. This occurs if the relevance is represented in terms of content within a set of segment boundaries, since if the segments are adjusted then so must the relevance data. This problem is avoided when metrics are based on the sentences that constitute the retrieval units, as is the case for the NTCIR-9 SpokenDoc metrics (uMAP, pwMAP, fMAP) that can be used directly on the retrieval results regardless of the segmentation method used.

The qrel files of metrics are based on prior knowledge of the set of segments containing relevant information in the collection, e.g. MAP, mGAP, MASP, introduced in Sections 3.2.3 and 5.2.3, are adjusted. We store the information about the relevance of each non-overlapping segment to indicate the amount of relevant information each adjusted segment contains, in order to use this in assessment of retrieval results.

The adjustment of the qrel files is carried out once the result list has been filtered by using the revised segment boundaries and overall knowledge of the start and end time of the relevant content. Figures 7.1 and 7.2 illustrate this adjusted file segmentation in the qrel file. These adjustments allow us to calculate the evaluation metric values, though they do not enable us to run direct comparison of the runs, since the qrels used vary with each combination of segment length and overlap. However, we can compare the results of further boundary adjustments on the same filtered results (when we suggest a jump-in point for a segment that is different from the segment start time), since the only change is in the jump-in point, and not the details of the relevant segments in the qrel file.

7.2 Filtering experiments on the NTCIR-9 SpokenDoc Collection

We segmented the NTCIR-9 SpokenDoc collection into units of 5 to 15 IPUs, with a sliding window that varied from 1 to 10 IPUs. The values of the segment units were chosen partly based on previous work that has demonstrated that segments of shorter fixed length (Kaneko et al., 2011) produced better results than those of longer average length (Eskevich and Jones, 2011a), see (Akiba et al., 2011) for a direct comparison of results. Another reason for the values chosen is based on analysis of the average length of the relevant segments measured in IPUs. Figure 7.2 shows that for most queries relevant segments are between 1 and 20 IPUs in length, with only 3 queries having relevant passages longer than about 50 IPUs.

Query	AVR length	Query	AVR length
SpokenDoc1-SDR-formal-0060	1.17	SpokenDoc1-SDR-formal-0032	8.75
SpokenDoc1-SDR-formal-0024	1.20	SpokenDoc1-SDR-formal-0027	8.80
SpokenDoc1-SDR-formal-0057	1.70	SpokenDoc1-SDR-formal-0079	9.00
SpokenDoc1-SDR-formal-0058	1.79	SpokenDoc1-SDR-formal-0074	9.09
SpokenDoc1-SDR-formal-0014	2.00	SpokenDoc1-SDR-formal-0013	9.25
SpokenDoc1-SDR-formal-0021	2.00	SpokenDoc1-SDR-formal-0034	9.43
SpokenDoc1-SDR-formal-0010	2.30	SpokenDoc1-SDR-formal-0040	9.53
SpokenDoc1-SDR-formal-0047	2.39	SpokenDoc1-SDR-formal-0044	9.60
SpokenDoc1-SDR-formal-0086	2.44	SpokenDoc1-SDR-formal-0009	10.00
SpokenDoc1-SDR-formal-0078	2.54	SpokenDoc1-SDR-formal-0031	10.25
SpokenDoc1-SDR-formal-0073	3.00	SpokenDoc1-SDR-formal-0072	10.28
SpokenDoc1-SDR-formal-0022	3.11	SpokenDoc1-SDR-formal-0081	10.30
SpokenDoc1-SDR-formal-0059	3.31	SpokenDoc1-SDR-formal-0084	10.50
SpokenDoc1-SDR-formal-0035	3.53	SpokenDoc1-SDR-formal-0012	10.58
SpokenDoc1-SDR-formal-0004	3.60	SpokenDoc1-SDR-formal-0011	11.44
SpokenDoc1-SDR-formal-0064	3.90	SpokenDoc1-SDR-formal-0003	11.60
SpokenDoc1-SDR-formal-0046	3.93	SpokenDoc1-SDR-formal-0048	11.66
SpokenDoc1-SDR-formal-0007	4.13	SpokenDoc1-SDR-formal-0050	11.97
SpokenDoc1-SDR-formal-0020	4.17	SpokenDoc1-SDR-formal-0075	12.09
SpokenDoc1-SDR-formal-0041	4.21	SpokenDoc1-SDR-formal-0052	12.25
SpokenDoc1-SDR-formal-0053	4.26	SpokenDoc1-SDR-formal-0066	12.43
SpokenDoc1-SDR-formal-0018	4.29	SpokenDoc1-SDR-formal-0076	12.71
SpokenDoc1-SDR-formal-0015	4.33	SpokenDoc1-SDR-formal-0042	12.82
SpokenDoc1-SDR-formal-0037	4.52	SpokenDoc1-SDR-formal-0067	13.17
SpokenDoc1-SDR-formal-0028	4.55	SpokenDoc1-SDR-formal-0016	13.61
SpokenDoc1-SDR-formal-0002	4.66	SpokenDoc1-SDR-formal-0061	14.03
SpokenDoc1-SDR-formal-0029	4.75	SpokenDoc1-SDR-formal-0045	14.10
SpokenDoc1-SDR-formal-0019	4.90	SpokenDoc1-SDR-formal-0069	14.90
SpokenDoc1-SDR-formal-0008	4.95	SpokenDoc1-SDR-formal-0085	15.20
SpokenDoc1-SDR-formal-0070	5.20	SpokenDoc1-SDR-formal-0062	15.40
SpokenDoc1-SDR-formal-0026	5.41	SpokenDoc1-SDR-formal-0056	15.53
SpokenDoc1-SDR-formal-0030	5.98	SpokenDoc1-SDR-formal-0051	15.95
SpokenDoc1-SDR-formal-0017	6.40	SpokenDoc1-SDR-formal-0065	16.31
SpokenDoc1-SDR-formal-0033	6.60	SpokenDoc1-SDR-formal-0005	17.33
SpokenDoc1-SDR-formal-0068	6.79	SpokenDoc1-SDR-formal-0083	18.17
SpokenDoc1-SDR-formal-0006	6.85	SpokenDoc1-SDR-formal-0063	18.28
SpokenDoc1-SDR-formal-0023	7.11	SpokenDoc1-SDR-formal-0077	18.48
SpokenDoc1-SDR-formal-0025	7.14	SpokenDoc1-SDR-formal-0080	18.86
SpokenDoc1-SDR-formal-0043	7.18	SpokenDoc1-SDR-formal-0054	19.43
SpokenDoc1-SDR-formal-0049	7.29	SpokenDoc1-SDR-formal-0036	20.20
SpokenDoc1-SDR-formal-0001	7.79	SpokenDoc1-SDR-formal-0055	49.23
SpokenDoc1-SDR-formal-0038	8.40	SpokenDoc1-SDR-formal-0082	92.89
SpokenDoc1-SDR-formal-0039	8.63	SpokenDoc1-SDR-formal-0071	111.75

Table 7.2: Average length of relevant segments (in IPUs) for the NTCIR-9 SpokenDoc task.

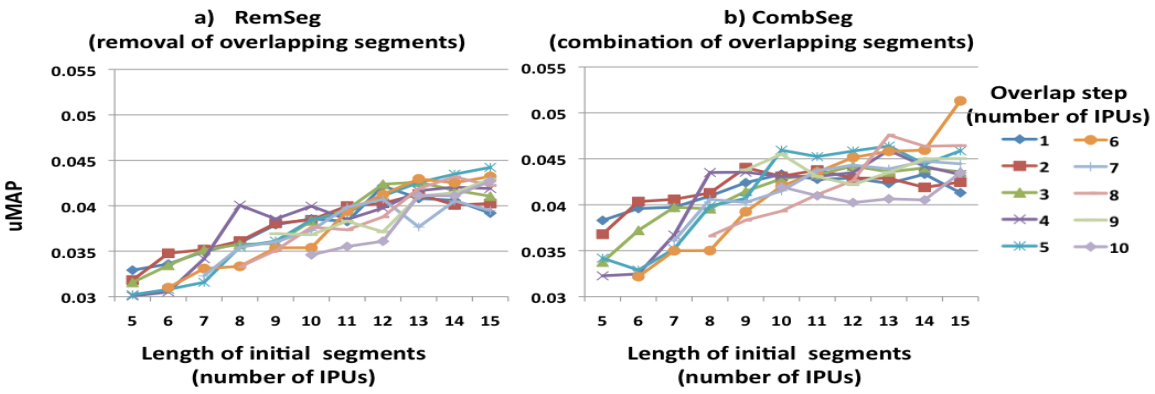


Figure 7.3: uMAP after filtering, ASR transcript

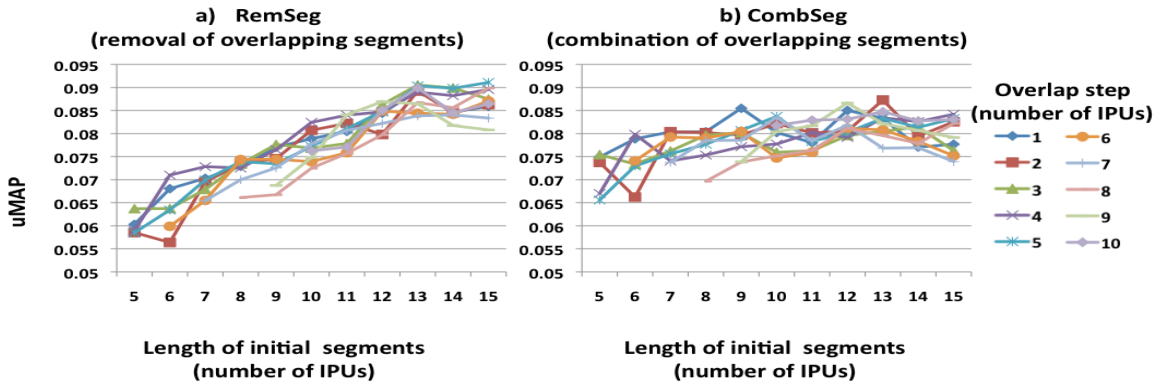


Figure 7.4: uMAP after filtering, MAN transcript

The retrieved results list was evaluated using the IPU-based metrics discussed in detail in the next sections.

7.2.1 uMAP

Figures 7.3 and 7.4 show that longer segments perform better than shorter ones for both ASR and manual transcripts when RemSeg filtering is applied, whereas results for the CombSeg filtering approach only follow this trend for the ASR transcript, and the difference in uMAP values between shorter (5 to 10 IPUs) and longer (10 to 15 IPUs) initial segments for the manual transcript is smaller.

For further analysis of these results we checked the precision within these segments in terms of IPUs and length calculated in seconds for segments that contain relevant content that are taken into account by uMAP and fMAP metrics (Figures 7.5 and 7.6 for precision based on IPUs for ASR and manual transcripts respectively, Figures 7.7 and 7.8 for precision based on time for ASR and manual transcripts

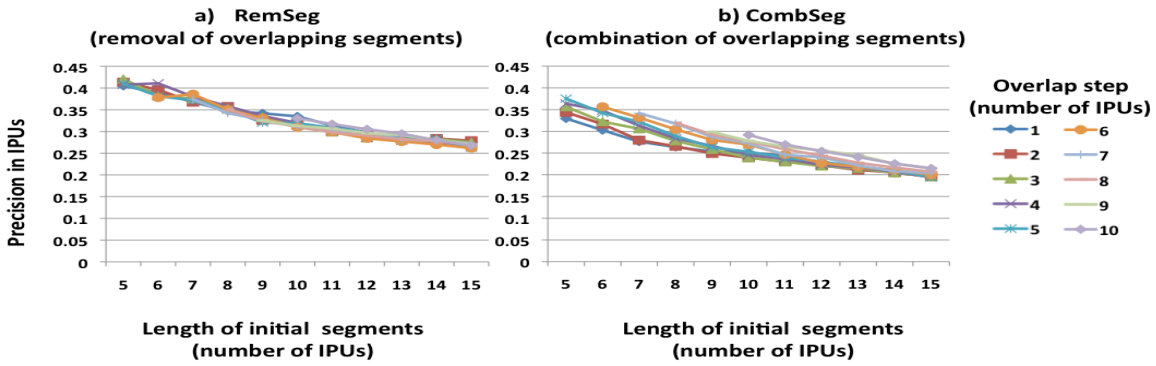


Figure 7.5: Precision of the content within all the segments containing relevant content, calculated in IPUs, ASR transcript

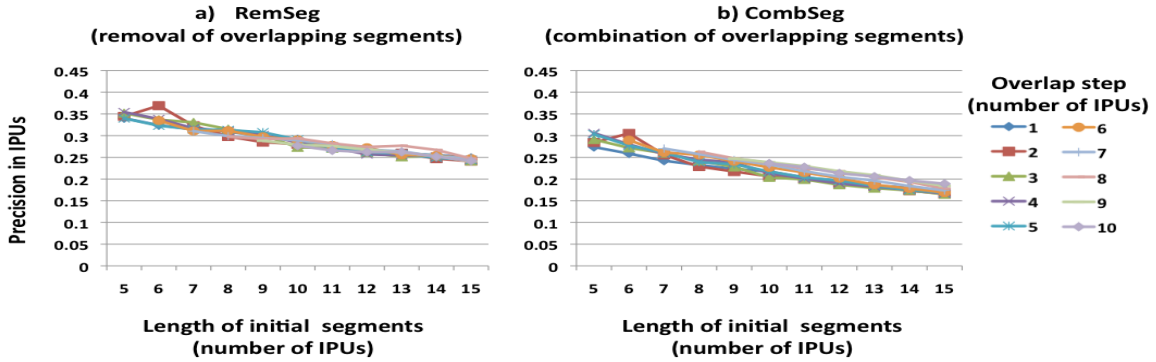


Figure 7.6: Precision of the content within all the segments containing relevant content, calculated in IPUs, MAN transcript

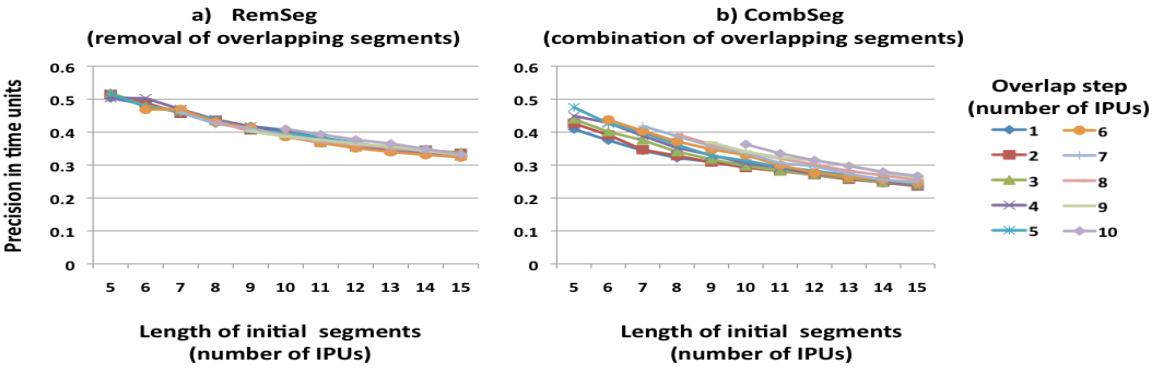


Figure 7.7: Precision of the content within all the segments containing relevant content, calculated in temporal length (seconds), ASR transcript

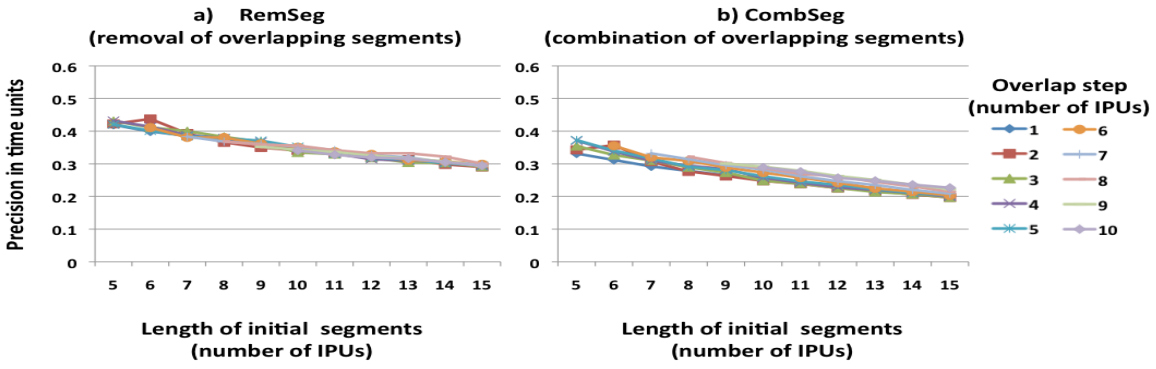


Figure 7.8: Precision of the content within all the segments containing relevant content, calculated in temporal length (seconds), MAN transcript

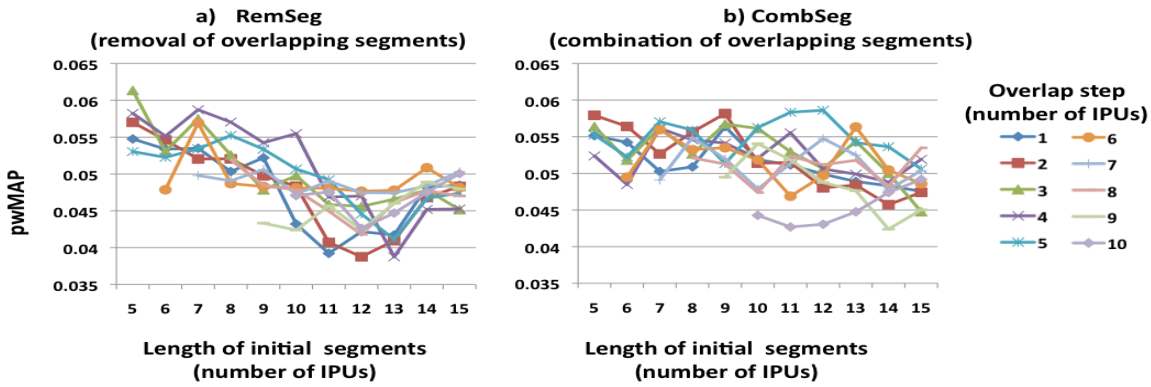


Figure 7.9: pwMAP after filtering, ASR transcript

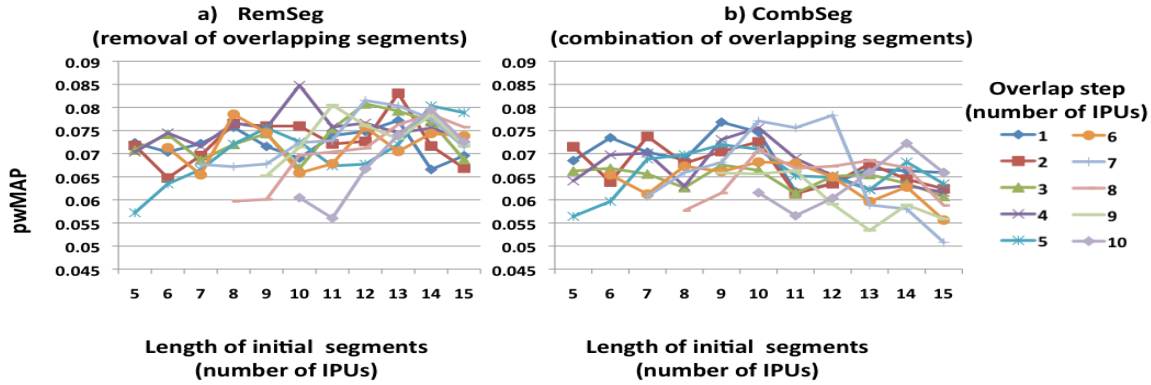


Figure 7.10: pwMAP after filtering, MAN transcript

respectively). As could be expected, precision within segments that are longer before filtering is lower for both transcript types. The CombSeg filtering approach shows lower precision, although of comparable values to RemSeg ones. However, as the uMAP metric shows better performance for longer segments which have lower precision for both RemSeg and CombSeg filtering approaches, it means that these segments are retrieved at higher ranks, and there is less non-relevant content at higher ranks in comparison with runs based on shorter segments.

7.2.2 pwMAP

Figures 7.9 and 7.10 show that filtering affects results for ASR transcript based runs differently for longer segments (between 10 and 15 IPUs). In the case of ASR transcript based runs, the RemSeg filtering approach achieves higher scores for intially shorter segments (between 5 and 10), while for the CombSeg filtering, pwMAP results stay within the same value range for all types of segments.

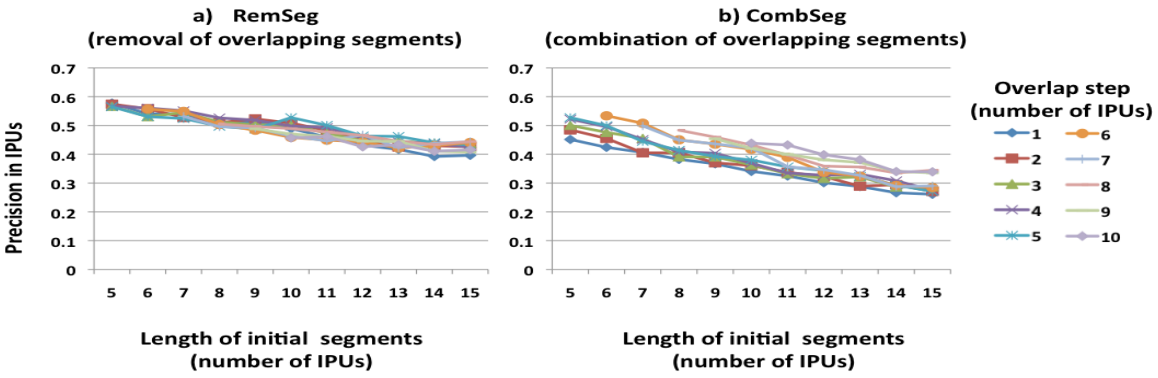


Figure 7.11: Precision of the content within the segments with relevant IPU in the middle, calculated in IPUs, ASR transcript

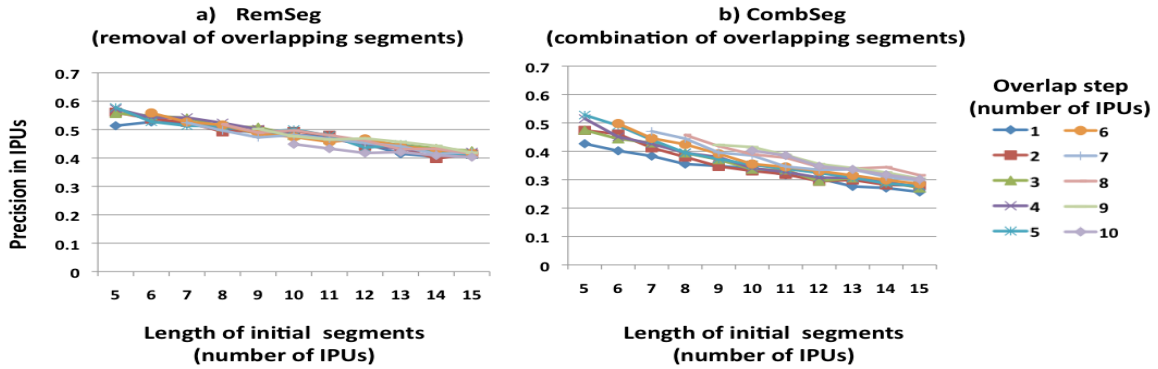


Figure 7.12: Precision of the content within the segments with relevant IPU in the middle, calculated in IPUs, MAN transcript

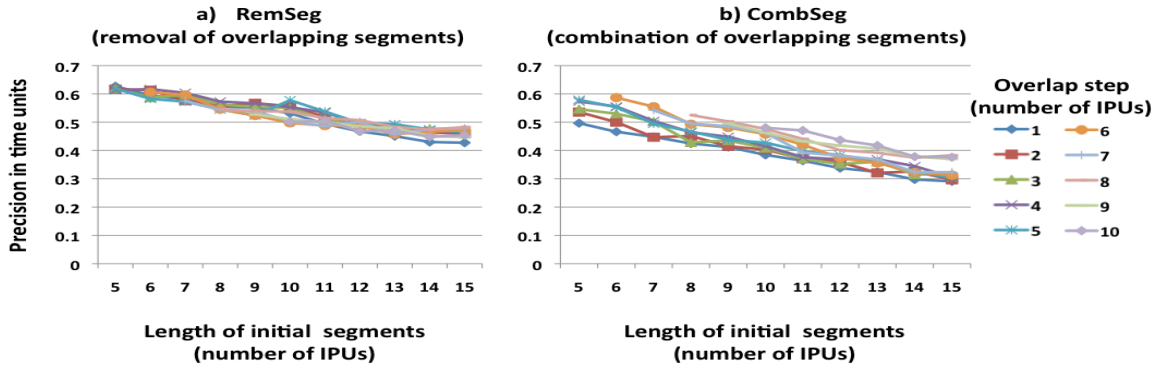


Figure 7.13: Precision of the content within the segments with relevant IPU in the middle, calculated in temporal length (seconds), ASR transcript

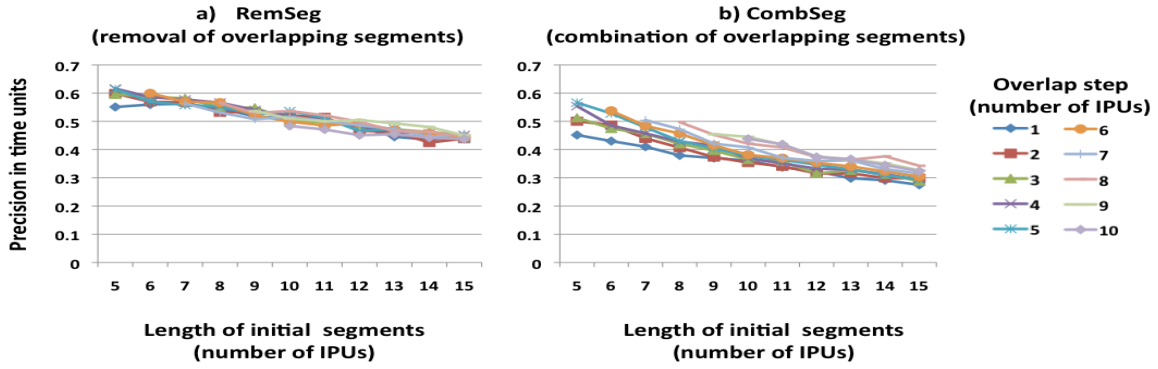


Figure 7.14: Precision of the content within the segments with relevant IPU in the middle, calculated in temporal length (seconds), MAN transcript

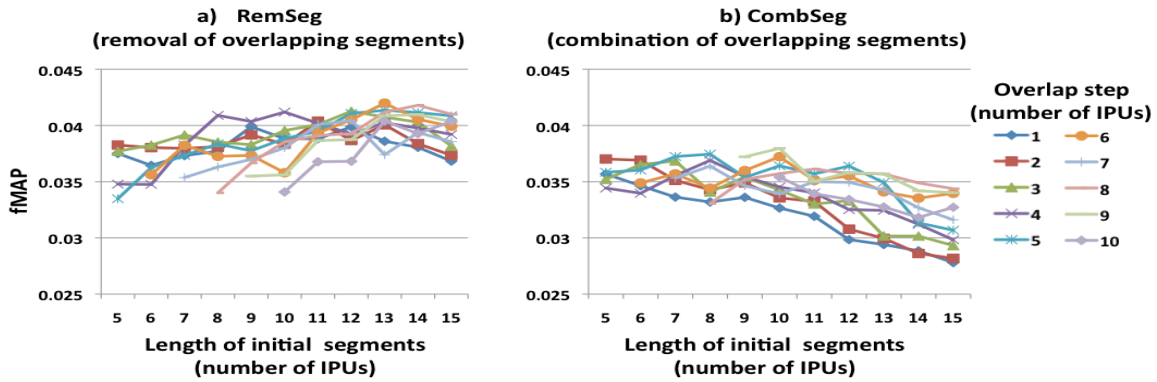


Figure 7.15: fMAP after filtering, ASR transcript

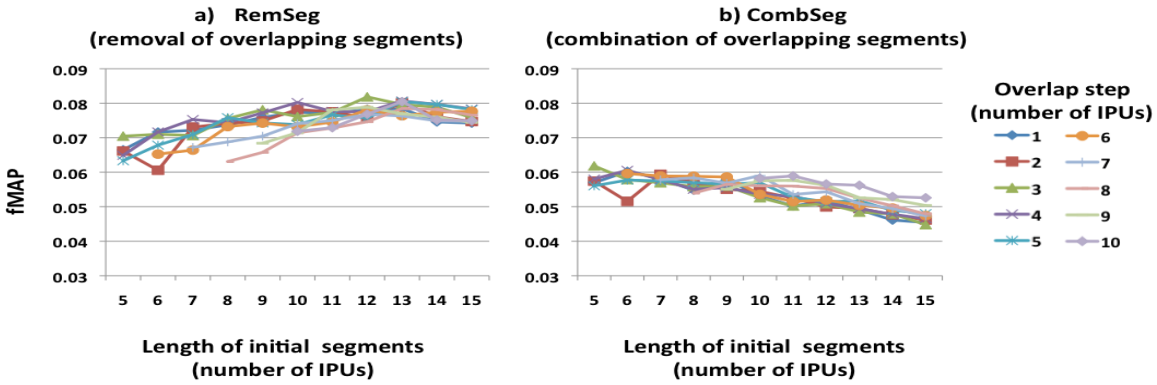


Figure 7.16: fMAP after filtering, MAN transcript

Figures 7.11 – 7.14 show precision of the relevant content for segments included in the calculation of pwMAP, i.e. those segments where the central IPU is relevant. As could be expected, in the case of precision for all segments (Figures 7.5 – 7.8), the overall trend across both types of transcripts and filtering approaches is that the longer the segments are, the lower the average precision within the segments is, with the only difference being that the segments that are used for pwMAP calculation have precision in a higher range of values (0.25–0.7 versus 0.2–0.45 for precision calculated in IPUs, and 0.25–0.7 versus 0.15–0.55 for precision calculated in time units).

7.2.3 fMAP

fMAP addresses both the precision and recall of the relevant IPUs within the retrieved segments. Results for this metric distinguish the two filtering approaches, since RemSeg achieves higher scores for longer segments for both ASR and manual

Overlap Step	Transcript Type					
	ASR			MAN		
	Overlap Step					
	10	30	60	10	30	60
10	–	**0.325**	**0.204**	–	**0.304**	**0.168**
30	0.539	–	**0.425**	0.556	–	**0.413**
60	0.637	0.638	–	0.647	0.649	–

Table 7.3: Average amount of segments with the same qrel boundaries after RemSeg filtering for runs with initial segments of 60 seconds. Values in bold correspond to segments containing relevant content, while non-bold ones represent segments without relevant content.

Overlap Step	Transcript Type					
	ASR			MAN		
	Overlap Step					
	10	30	60	10	30	60
10	–	**0.084**	**0.075**	–	**0.074**	**0.061**
30	0.498	–	**0.152**	0.508	–	**0.173**
60	0.617	0.599	–	0.623	0.614	–

Table 7.4: Average amount of segments with the same qrel boundaries after CombSeg filtering for runs with initial segments of 60 seconds. Values in bold correspond to segments containing relevant content, while non-bold ones represent segments without relevant content.

transcripts, while CombSeg performs better for shorter segments for both types of transcript, see Figures 7.15 and 7.16. In the case of RemSeg filtering on the ASR transcript, pwMAP is lower for longer segments, and uMAP is higher for these segments. Since fMAP is higher for longer segments, we can assume that this method retains more segments with relevant content at the beginning or end of the segments than those with the relevant content in the centre in the top ranked segments.

7.3 Filtering and boundary adjustment experiments for the AMI corpus collection

We follow the same SCR framework, as introduced in Section 7.1 (segmentation with overlapping window followed by filtering of the results), for retrieval experiments using the AMI corpus with the only difference from the NTCIR-9 SpokenDoc based

Overlap Step	Transcript Type							
	ASR				MAN			
	Overlap Step							
	10	30	60	90	10	30	60	90
10	–	**0.308**	**0.125**	**0.163**	–	**0.306**	**0.123**	**0.153**
30	0.542	–	**0.286**	**0.323**	0.559	–	**0.279**	**0.309**
60	0.473	0.482	–	**0.466**	0.491	0.502	–	**0.443**
90	0.667	0.565	0.612	–	0.679	0.586	0.619	–

Table 7.5: Average amount of segments with the same qrel boundaries after RemSeg filtering for runs with initial segments of 90 seconds. Values in bold correspond to segments containing relevant content, while non-bold ones represent segments without relevant content.

Overlap Step	Transcript Type							
	ASR				MAN			
	Overlap Step							
	10	30	60	90	10	30	60	90
10	–	**0.083**	**0.020**	**0.045**	–	**0.064**	**0.014**	**0.035**
30	0.491	–	**0.136**	**0.081**	0.504	–	**0.117**	**0.052**
60	0.430	0.442	–	**0.121**	0.445	0.454	–	**0.090**
90	0.620	0.522	0.554	–	0.629	0.534	0.562	–

Table 7.6: Average amount of segments with the same qrel boundaries after CombSeg filtering for runs with initial segments of 90 seconds. Values in bold correspond to segments containing relevant content, while non-bold ones represent segments without relevant content.

experiments being that the segmentation unit and window overlap step are defined in terms of seconds. We set segmentation boundaries at varying time lengths (from 60 seconds to 180), with a sliding window varying from a small step within the segment to the full length of the segment (from 10 seconds to 180 seconds) that effectually produces a run with no overlapping segments at all.

Since the AMI transcripts contain time stamp information for each word in both types of the transcripts, we carry out segment boundary adjustment of different types as shown in Table 7.1 on retrieval results for both RemSeg and CombSeg filtered runs:

- *Lexical cohesion based segment boundary adjustment:* As retrieval from lexical cohesion based segments might be biased due to the varying length of the segments, we assume that this impact can be avoided when using segments

of fixed length, while use of the lexical cohesion boundaries as jump-in point adjustment might bring the user closer to semantically coherent part of the retrieved segment. If a fixed length retrieved segment contains a lexical cohesion based boundary (C99, TextTiling) then its starting point is used as a jump-in point.

- *Use of the first pause in the segment:* This adjustment demonstrates improved performance in our experiments on other datasets of semi-professional and professional broadcasts, as described in (Eskevich and Jones, 2012) and (Eskevich and Jones, 2013) respectively. Therefore we use the same value to define pause length 0.5 seconds or longer.

- *Use of the longest pause in the segment:.* As discussed previously in Chapter 6, in cases when the content is split between several segments, the non-relevant part of these segments contains additional information that might help in its retrieval, if it is on the same topic, or decrease the results, when it covers another topic of conversation. With segment boundary adjustment we cannot change the rank of such segments, but we might be able to usefully shift the jump-in point of the segment to this potential place of topic change. Here we use the longest pause in the segment as a potential alternative start time, as introduction of new topics or change of topics might change the dynamics of the conversation, thus causing a significant break indicated by a long pause.

- *Use of the loudness (energy peaks) in the segment:* This adjustment aimed to use information about the loudest word in the segment, assuming that this word should be in an area of prominence that consequetly might be correlated to relevant content within the segment. Across all the collection we calculated the loudness within a window of one fixed length across all audio content, these values were then averaged for each word in the transcript according to their time stamps. To obtain the initial loudness results we used the OpenSmile software (Eyben et al., 2010), default configuration, with window length equal

Overlap Step	Transcript Type									
	ASR					MAN				
	Overlap Step									
	10	30	60	90	120	10	30	60	90	120
10	–	**0.333**	**0.169**	**0.143**	**0.147**	–	**0.331**	**0.153**	**0.132**	**0.142**
30	0.559	–	**0.434**	**0.294**	**0.236**	0.580	–	**0.378**	**0.287**	**0.214**
60	0.481	0.561	–	**0.279**	**0.445**	0.496	0.567	–	**0.299**	**0.452**
90	0.302	0.365	0.381	–	**0.221**	0.317	0.377	0.391	–	**0.214**
120	0.691	0.550	0.589	0.346	–	0.695	0.566	0.604	0.361	–

Table 7.7: Average amount of segments with the same qrel boundaries after RemSeg filtering for runs with initial segments of 120 seconds. Values in bold correspond to segments containing relevant content, while non-bold ones represent segments without relevant content.

Overlap Step	Transcript Type									
	ASR					MAN				
	Overlap Step									
	10	30	60	90	120	10	30	60	90	120
10	–	**0.075**	**0.027**	**0.019**	**0.052**	–	**0.059**	**0.017**	**0.013**	**0.037**
30	0.490	–	**0.196**	**0.119**	**0.055**	0.507	–	**0.160**	**0.092**	**0.034**
60	0.409	0.466	–	**0.089**	**0.101**	0.417	0.471	–	**0.084**	**0.081**
90	0.259	0.305	0.319	–	**0.043**	0.269	0.314	0.330	–	**0.033**
120	0.609	0.482	0.501	0.300	–	0.609	0.494	0.506	0.311	–

Table 7.8: Average amount of segments with the same qrel boundaries after CombSeg filtering for runs with initial segments of 120 seconds. Values in bold correspond to segments containing relevant content, while non-bold ones represent segments without relevant content.

to 10 seconds). Then for each segment we find a word with the highest loudness score, and assign the jump-in point to its starting time.

7.3.1 Adjustment of evaluation framework for metrics using segments in case of filtered runs

The evaluation of these runs was done using the metrics of MAP, mGAP, MASP. In all the results, the numbers in the left column represent RemSeg filtering (removal of the overlapping segments), and in the right column, CombSeg filtering (combination of the overlapping segments). In order to be able to analyse the impact of segment boundary adjustment, we calculate all metrics, including MAP, using a window of

Overlap Step	Overlap Step					
	10	30	60	90	120	150
	ASR					
10	–	**0.343**	**0.210**	**0.124**	**0.136**	**0.147**
30	0.571	–	**0.521**	**0.317**	**0.255**	**0.201**
60	0.387	0.522	–	**0.344**	**0.421**	**0.192**
90	0.302	0.369	0.384	–	**0.215**	**0.162**
120	0.246	0.289	0.416	0.285	–	**0.153**
150	0.701	0.541	0.402	0.325	0.280	–
	MAN					
10	–	**0.268**	**0.154**	**0.104**	**0.082**	**0.133**
30	0.589	–	**0.537**	**0.302**	**0.249**	**0.201**
60	0.409	0.534	–	**0.320**	**0.416**	**0.187**
90	0.324	0.382	0.402	–	**0.250**	**0.152**
120	0.263	0.299	0.422	0.299	–	**0.144**
150	0.710	0.558	0.420	0.342	0.298	–

Table 7.9: Average amount of segments with the same qrel boundaries after RemSeg filtering for runs with initial segments of 150 seconds. Values in bold correspond to segments containing relevant content, while non-bold ones represent segments without relevant content.

Overlap Step	Overlap Step					
	10	30	60	90	120	150
	ASR					
10	–	**0.071**	**0.014**	**0.017**	**0.009**	**0.040**
30	0.501	–	**0.126**	**0.119**	**0.082**	**0.019**
60	0.326	0.419	–	**0.084**	**0.165**	**0.022**
90	0.244	0.285	0.311	–	**0.061**	**0.025**
120	0.196	0.226	0.313	0.232	–	**0.024**
150	0.575	0.445	0.330	0.262	0.232	–
	MAN					
10	–	**0.058**	**0.023**	**0.010**	**0.010**	**0.027**
30	0.516	–	**0.106**	**0.079**	**0.085**	**0.021**
60	0.341	0.431	–	**0.080**	**0.196**	**0.015**
90	0.260	0.299	0.326	–	**0.047**	**0.015**
120	0.206	0.228	0.320	0.244	–	**0.021**
150	0.576	0.460	0.344	0.278	0.246	–

Table 7.10: Average amount of segments with the same qrel boundaries after Comb-Seg filtering for runs with initial segments of 150 seconds. Values in bold correspond to segments containing relevant content, while non-bold ones represent segments without relevant content.

Overlap Step	Overlap Step						
	10	30	60	90	120	150	180
	ASR						
10	–	**0.327**	**0.189**	**0.151**	**0.102**	**0.081**	**0.162**
30	0.583	–	**0.414**	**0.299**	**0.199**	**0.167**	**0.204**
60	0.461	0.552	–	**0.317**	**0.469**	**0.186**	**0.354**
90	0.455	0.456	0.419	–	**0.237**	**0.178**	**0.467**
120	0.327	0.332	0.470	0.312	–	**0.149**	**0.316**
150	0.205	0.230	0.231	0.237	0.233	–	**0.157**
180	0.713	0.537	0.495	0.545	0.380	0.229	–
	MAN						
10	–	**0.303**	**0.175**	**0.120**	**0.092**	**0.077**	**0.131**
30	0.608	–	**0.378**	**0.289**	**0.175**	**0.163**	**0.174**
60	0.468	0.559	–	**0.268**	**0.444**	**0.175**	**0.324**
90	0.467	0.466	0.427	–	**0.215**	**0.172**	**0.431**
120	0.335	0.345	0.472	0.326	–	**0.145**	**0.360**
150	0.217	0.242	0.247	0.255	0.247	–	**0.147**
180	0.719	0.559	0.499	0.551	0.387	0.246	–

Table 7.11: Average amount of segments with the same qrel boundaries after RemSeg filtering for runs with initial segments of 180 seconds. Values in bold correspond to segments containing relevant content, while non-bold ones represent segments without relevant content.

Overlap Step	Overlap Step					
	10	30	60	90	120	150
	ASR					
10	–	**0.071**	**0.014**	**0.017**	**0.009**	**0.040**
30	0.501	–	**0.126**	**0.119**	**0.082**	**0.019**
60	0.326	0.419	–	**0.084**	**0.165**	**0.022**
90	0.244	0.285	0.311	–	**0.061**	**0.025**
120	0.196	0.226	0.313	0.232	–	**0.024**
150	0.575	0.445	0.330	0.262	0.232	–
	MAN					
10	–	**0.058**	**0.023**	**0.010**	**0.010**	**0.027**
30	0.516	–	**0.106**	**0.079**	**0.085**	**0.021**
60	0.341	0.431	–	**0.080**	**0.196**	**0.015**
90	0.260	0.299	0.326	–	**0.047**	**0.015**
120	0.206	0.228	0.320	0.244	–	**0.021**
150	0.576	0.460	0.344	0.278	0.246	–

Table 7.12: Average amount of segments with the same qrel boundaries after CombSeg filtering for runs with initial segments of 180 seconds. Values in bold correspond to segments containing relevant content, while non-bold ones represent segments without relevant content.

tolerance (60 seconds), i.e. we consider a segment relevant if the relevant content is present within a window of 60 seconds of its beginning.

Since the qrels are changed according to the filtering results for each run, the values on each figure are not directly comparable. We calculate average amount of segments with same boundaries in qrel relevance files between pairs of runs that differ only by the size of overlapping step while being based on the same type of transcript and same filtering scheme. This data is shown in Tables 7.3 – 7.12. The average is calculated separately for segments that contain relevant content (in bold in the tables) and the remainder of the segments which do not.

Across different values of initial segments RemSeg filtering produces a considerably larger amount of segments with the same boundaries, e.g. for initial segment equal to 60 it is 0.325 between overlapping steps 10 and 30, 0.204 between overlapping steps 10 and 60, 0.425 between overlapping steps 30 and 60 against 0.084, 0.075, 0.152 respectively for the ASR transcript (Figures 7.3 and 7.4). This result may be caused by the nature of CombSeg filtering approach that changes the boundaries of the regions with the relevant content. The average amount of change for the non-relevant segments is within similar range of values across both types of filtering, types of transcripts, and size of initial segment units.

Another trend that is consistent for shorter initial segments (Figures 7.3–7.6), and is present for most of the comparison combinations for longer segments, is that the runs based on ASR transcripts have a higher average amount of segments with the same boundaries than the runs based on the manual transcript. This may be explained by the fact that the better quality of manual transcript allows these runs to find more segments in the region of relevant content, thus increasing the probability of various non-overlapping segments being present in the top of the ranked list.

Overall, this comparison of relevance qrels across different runs confirms that we cannot directly compare the retrieval results. However as the result list does not change when the boundary adjustment is introduced, the results can be compared for the same transcript type, the same initial segment and overlap step, and the

same filtering approach across different boundaries adjustments, i.e. the results in each column for the same metric are comparable. For example, RemSeg results in the left column in Figures 7.17 and 7.18 are calculated for the same result list of each filtered run, and only the start times of the segments are adjusted using the C99 segments boundaries for the calculations in Figure 7.18.

Figures 7.17-7.22 and 7.23-7.28 show MAP values for ASR and manual transcripts; Figures 7.29-7.34 and 7.35-7.40 - mGAP respectively; Figures 7.41-7.46 and 7.47-7.52 - MASP respectively; and precision of relevant content in the segments is shown in Figures 7.53-7.58 and 7.59-7.64 respectively. In the following sections we analyse how these scores reflect the trends in SCR, and which techniques improve its effectiveness.

7.3.2 Impact of filtering

CombSeg filtering for the shortest segments (60 segments) and longer non-overlapping segments (initial length and overlap step equal to 90/90, 120/120, 150/150) achieves the highest MAP scores for both ASR (0.216-0.234) and manual (0.221-0.268) transcripts when compared to other runs following the same filtering approach and RemSeg filtering, see Figures 7.17 and 7.23. This means that these runs manage to retrieve the relevant content at higher ranks.

MAP values for RemSeg filtering show less variability in scores compared to CombSeg filtering, see Figures 7.17 and 7.23. Runs with segments of length 90 seconds and longer containing non-overlap segments get higher MAP scores than overlapping ones for CombSeg filtering approach for both ASR and manual transcripts. mGAP and MASP scores show similar behaviour (Figures 7.29 and 7.35, and 7.41 and 7.47).

7.3.3 Impact of boundary adjustment

Use of C99 and TextTiling start times for jump-in point adjustment: This type of boundary adjustment is shown to be more effective for the CombSeg filtered runs, while results for RemSeg filtering stay the same or differ insignificantly, see Figures 7.18-7.19 and 7.24-7.25 compared to baseline filtered runs results before boundary adjustement presented in Figures 7.17 and 7.23 respectively.

For CombSeg filtering cases with overlapped sliding window show improvement in the results and decreased scores for the runs without it for both ASR and manual transcript measured using MAP and mGAP (with initial segment equal to 90 seconds and longer), see Figures 7.18-7.19 and 7.24-7.25 for MAP, Figures 7.30-7.31 and 7.36-7.37 for mGAP. CombSeg filtering with longer initial segments can produce longer segments as result of overlap. In the cases when the relevant content is shorter than these combined segments, or some segments of non-relevant content are grouped together, the use of lexical cohesion based boundaries enables us to get closer to the beginning of different sections of large segments, i.e. potentially closer to the beginning of the relevant content, if it is present in the segment closer to its end. This method enables the amount of non-relevant content for the user to audition to be reduced. The latter statement is confirmed by the scores achieved by the MASP scores for these runs, see Figures 7.42-7.43 and 7.48-7.49. Figures 7.53-7.55 and 7.59-7.61 confirm this argument demonstrating that both C99 and TextTiling boundary adjustment increase precision of relevant content when compared to initial filtered results.

MASP results for CombSeg filtering demonstrate more difference between C99 and TextTiling segmentation, as can be seen by the latter version of adjustment achieving higher scores than the former one, as well as the initial filtering before any adjustment, see right columns of the Figures 7.48-7.49 against 7.42-7.43, and 7.41 and 7.47.

Use of pauses for boundary adjustment: Use of first pause[1] in the segments does not show significant impact on the results across MAP, mGAP and MASP metrics (Figures 7.20 and 7.26 for MAP, Figures 7.20 and 7.26 for mGAP, Figures 7.20 and 7.26 for MASP).

Use of the longest pause in the segment shows improvement only for some of the segmentation runs with overlap (90 seconds) using the manual transcript with CombSeg filtering across all the metrics, and for runs with short segments evaluatec using MASP (Figures 7.21 and 7.27 for MAP, Figures 7.21 and 7.27 for mGAP, Figures 7.21 and 7.27 for MASP).

In previous experiments, boundary adjustment using pauses was implemented for a known-item task (Eskevich and Jones, 2013), and in the case of ad hoc retrieval experiment this approach appear to have little impact on the results. This is caused by the fact that the right adjustment of a boundary for the only relevant segment which is calculated by the evaluation metrics produces more significant impact, and it is neutralized in case of the ad-hoc task with more than one relevant segment that are taken into account by the scores.

Use of loudness for boundary adjustment: Another method of boundary adjustment relied on the calculation of the average loudness for each word in the segment. This type of adjustment achieved the lowest scores across the metrics as compared to the non-adjusted result runs (Figures 7.22 and 7.28 for MAP, Figures 7.22 and 7.28 for mGAP, Figures 7.22 and 7.28 for MASP). This suggests that the most prominent word in the segment might not be the best candidate for a jump-in point. Potentially the beginning of the sentence containing this prominent word might be a better candidate, as the relevance assessment was created manually, and people tend to set the boundaries of the relevant content in terms of full sentences.

[1]We consider the distance between 2 words that is longer than 0.5 seconds a pause.

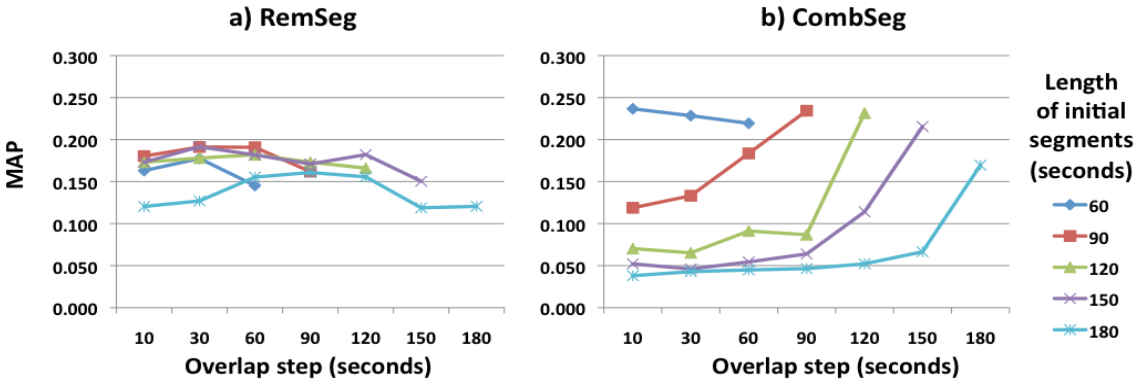


Figure 7.17: MAP after filtering, ASR transcript

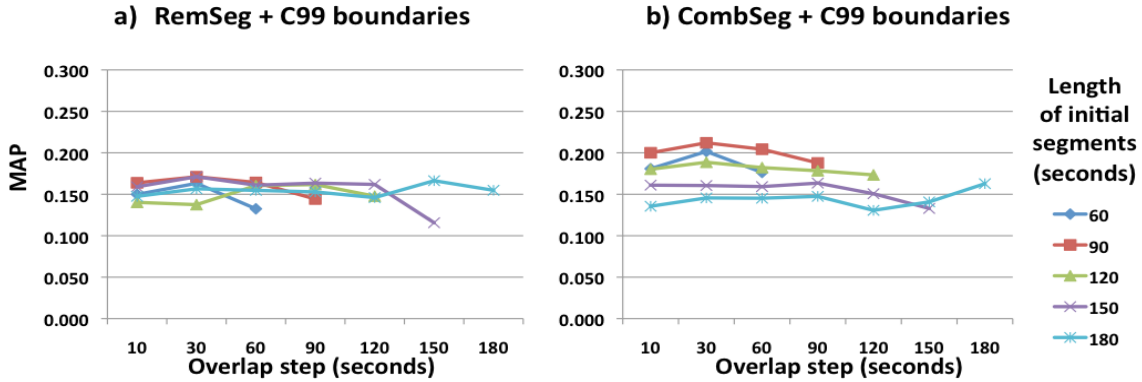


Figure 7.18: MAP after filtering and use of C99 segment boundaries, ASR transcript

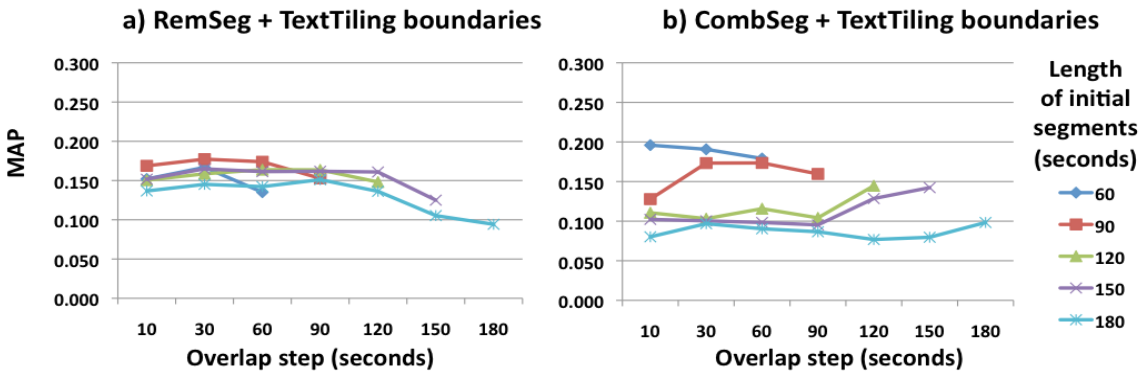


Figure 7.19: MAP after filtering and use of TextTiling segment boundaries, ASR transcript

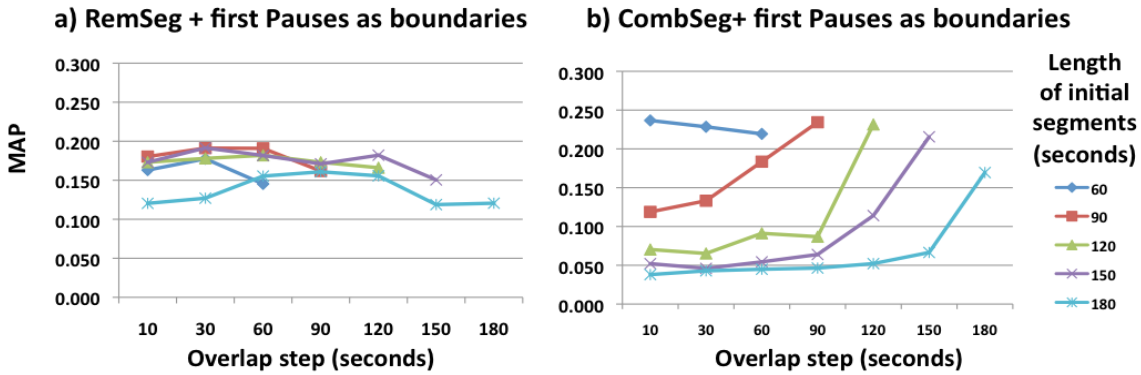


Figure 7.20: MAP after filtering and use of first pause as segment boundaries, ASR transcript

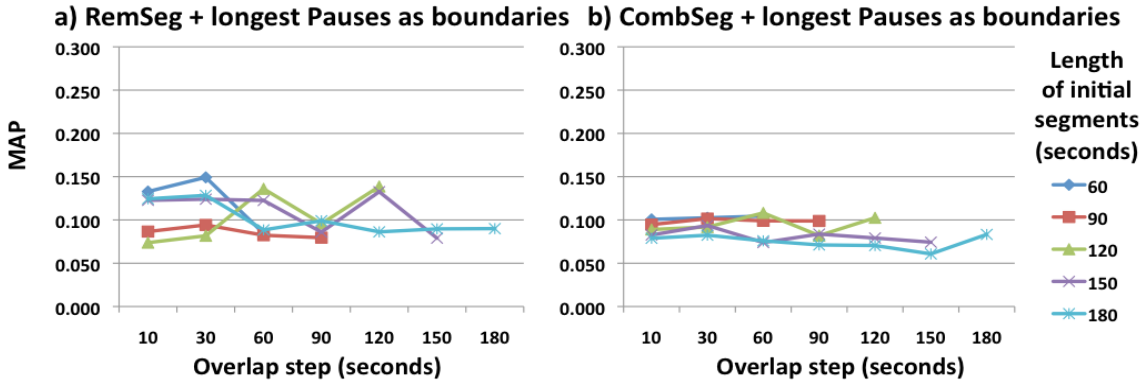


Figure 7.21: MAP after filtering and use of longest pause as segment boundaries, ASR transcript

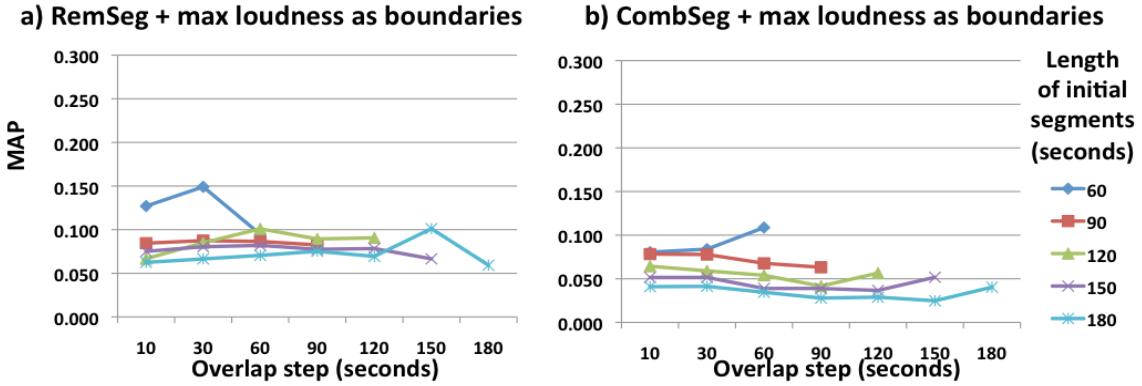


Figure 7.22: MAP after filtering and use of a word with maximum loudness as segment boundaries, ASR transcript

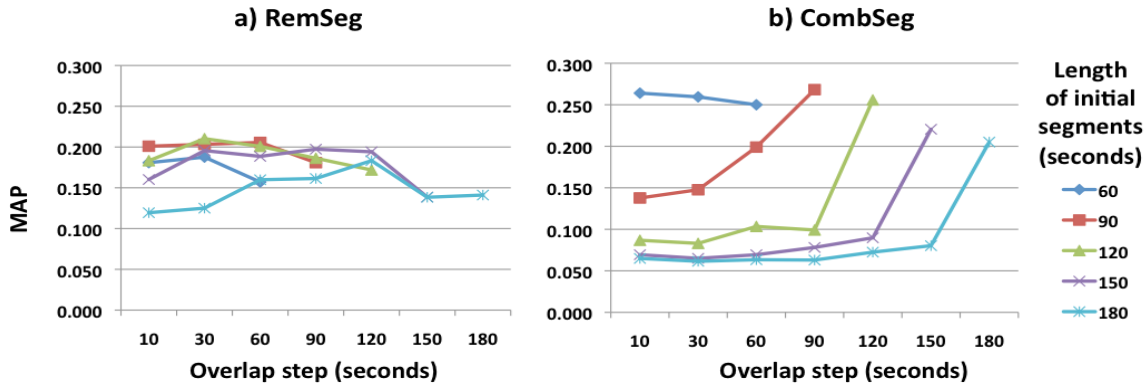


Figure 7.23: MAP after filtering, manual transcript

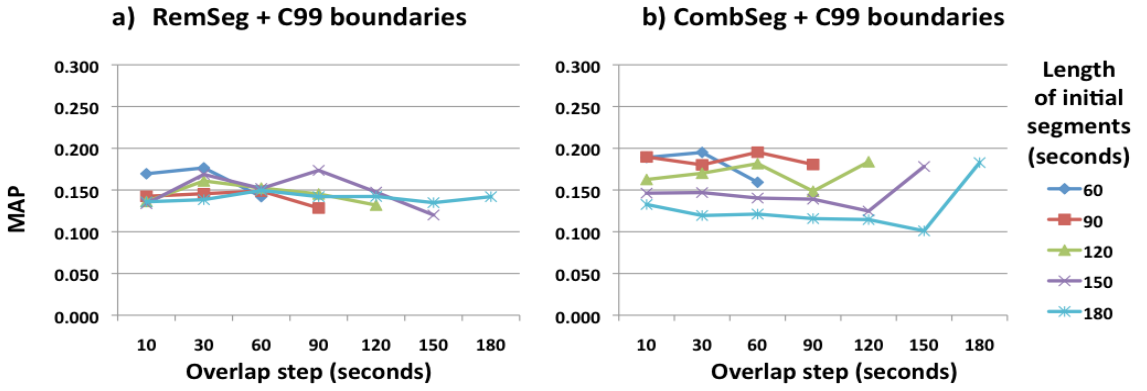


Figure 7.24: MAP after filtering and use of C99 segment boundaries , manual transcript

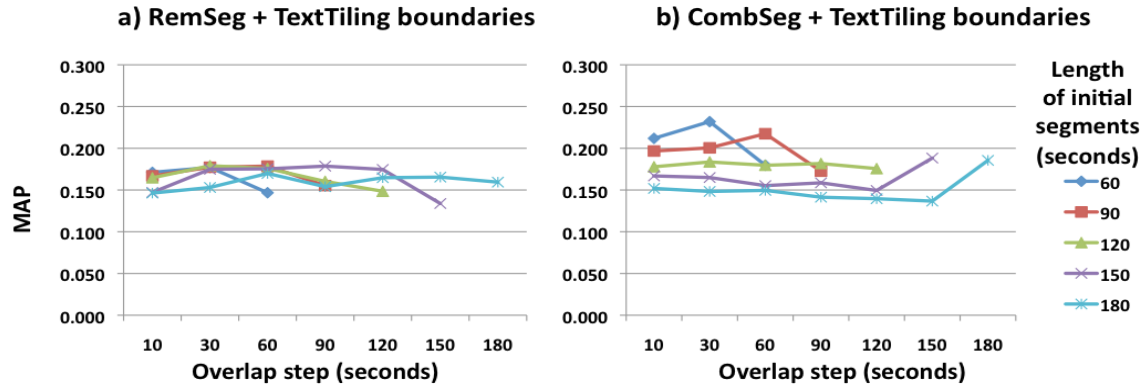


Figure 7.25: MAP after filtering and use of TextTiling segment boundaries, manual transcript

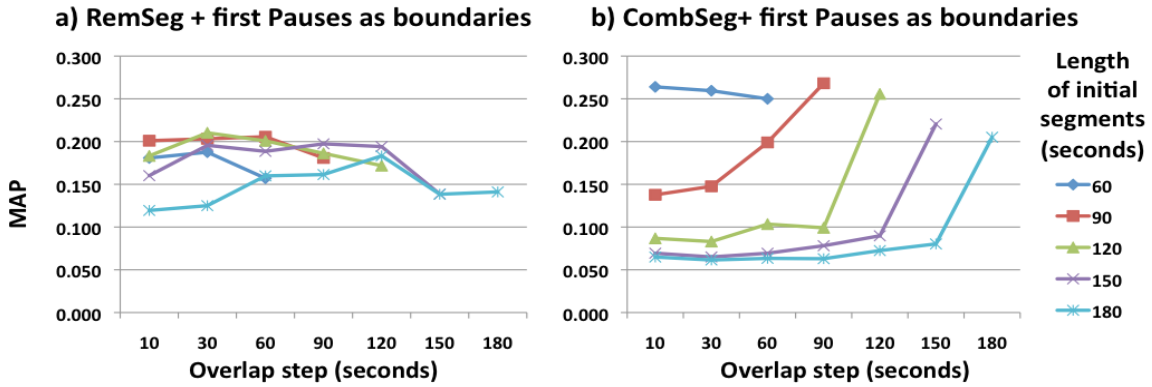


Figure 7.26: MAP after filtering and use of first pause as segment boundaries, manual transcript

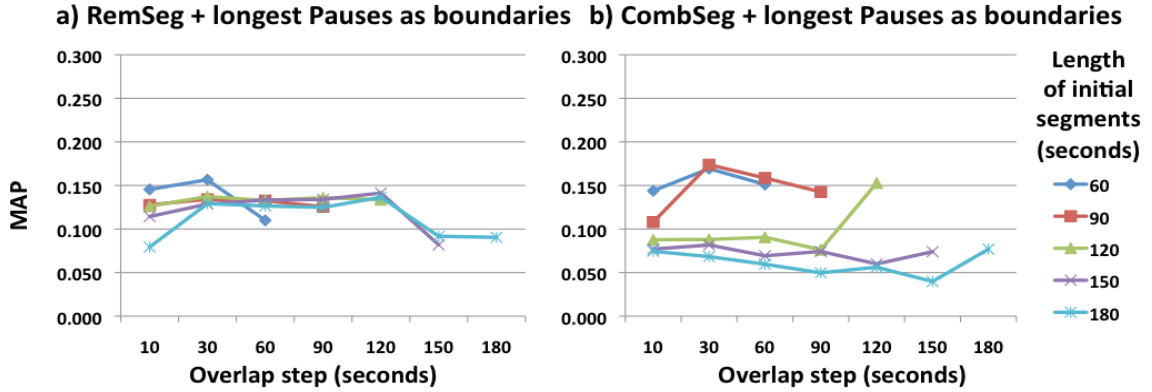


Figure 7.27: MAP after filtering and use of longest pause as segment boundaries, manual transcript

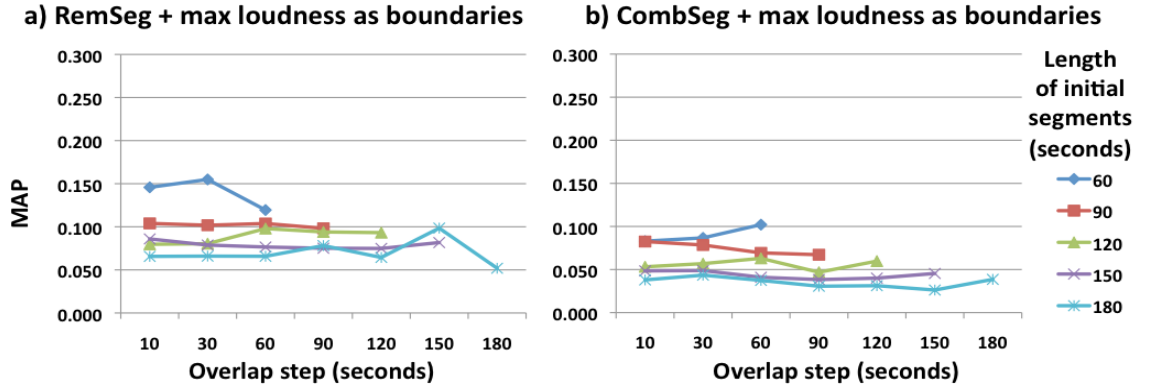


Figure 7.28: MAP after filtering and use of a word with maximum loudness as segment boundaries, manual transcript

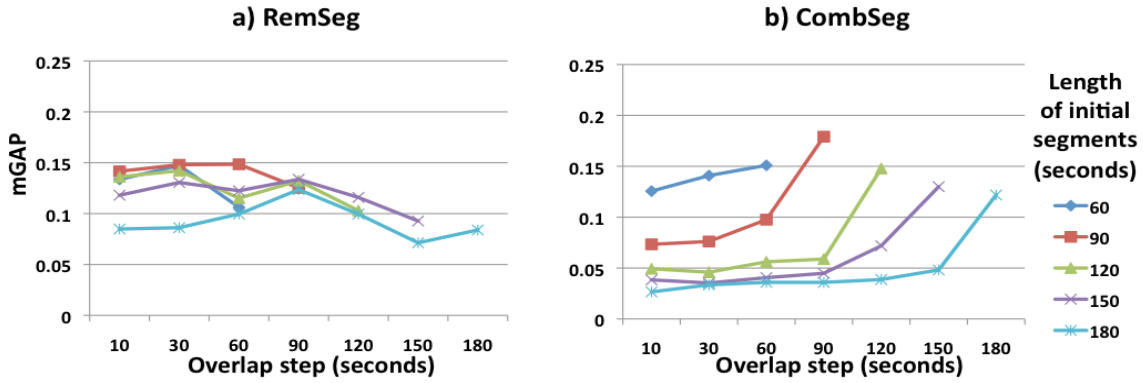


Figure 7.29: mGAP after filtering, ASR transcript

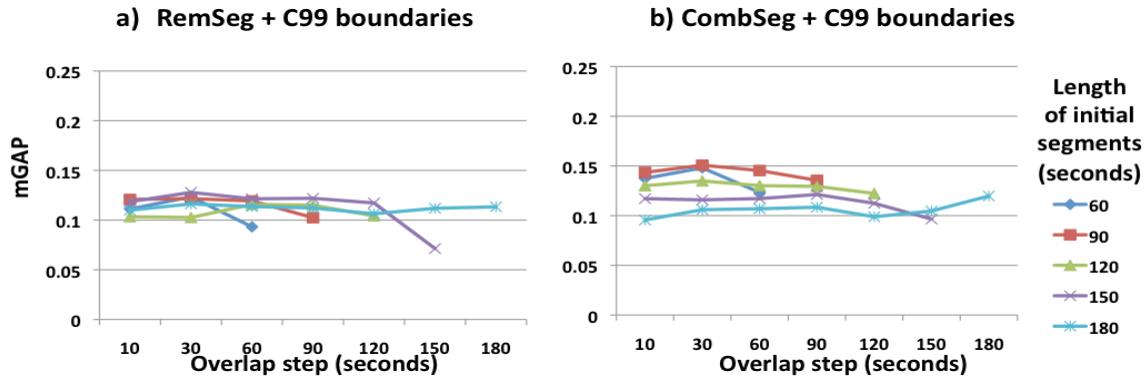


Figure 7.30: mGAP after filtering and use of C99 segment boundaries, ASR transcript

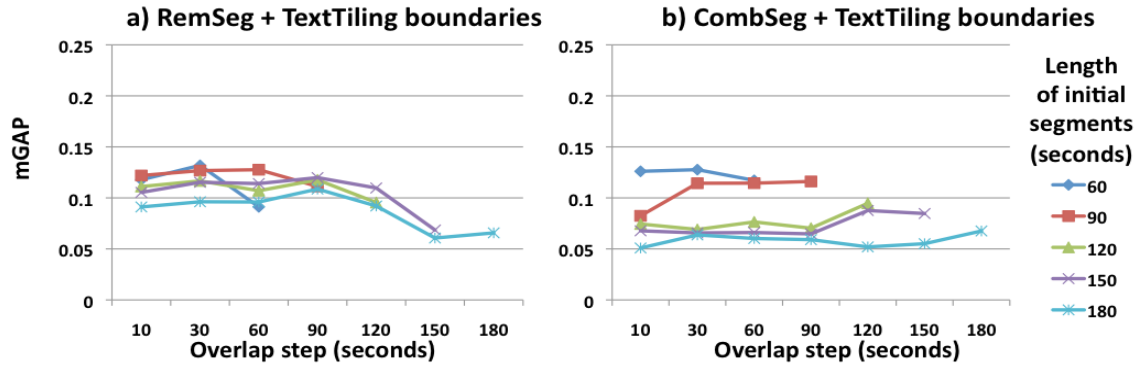


Figure 7.31: mGAP after filtering and use of TextTiling segment boundaries, ASR transcript

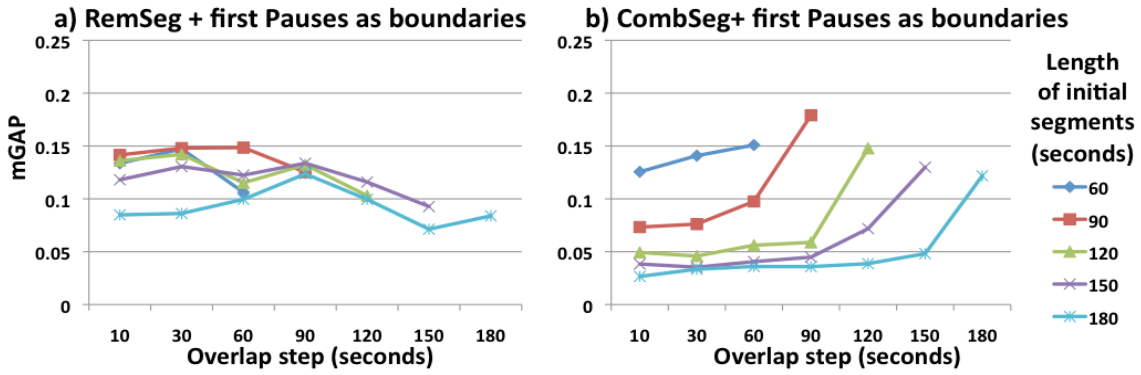


Figure 7.32: mGAP after filtering and use of first pause as segment boundaries, ASR transcript

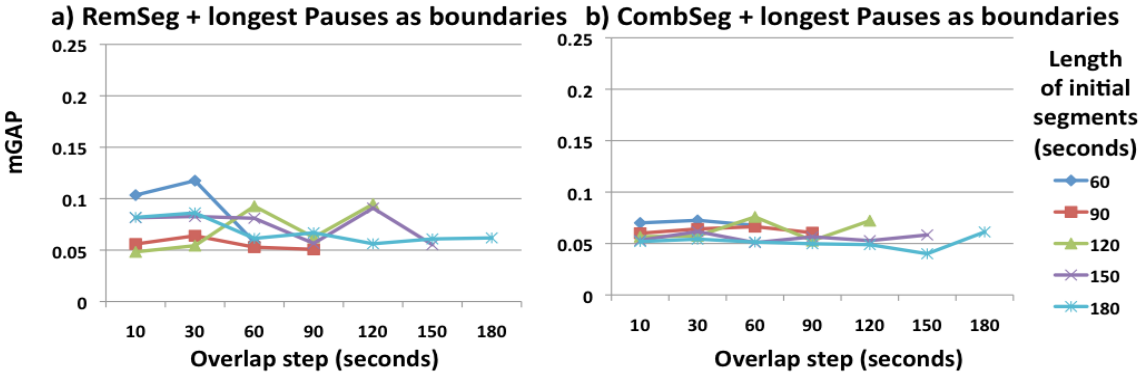


Figure 7.33: mGAP after filtering and use of longest pause as segment boundaries, ASR transcript

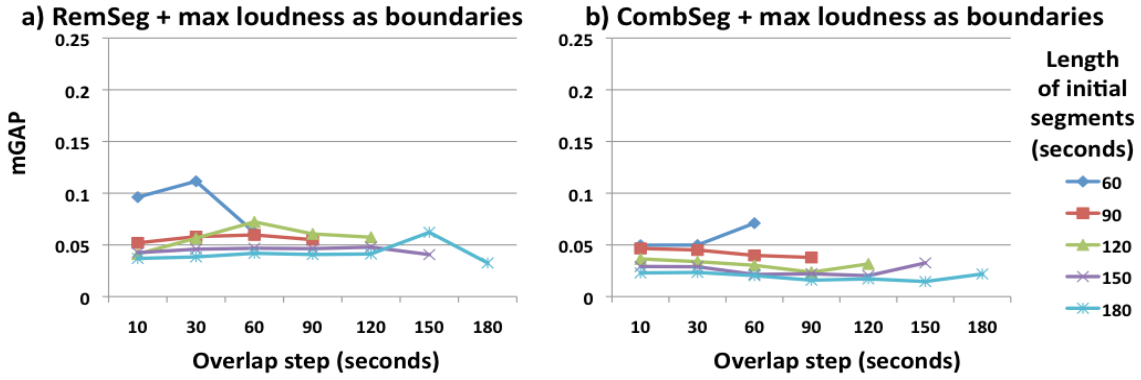


Figure 7.34: mGAP after filtering and use of a word with maximum loudness as segment boundaries, ASR transcript

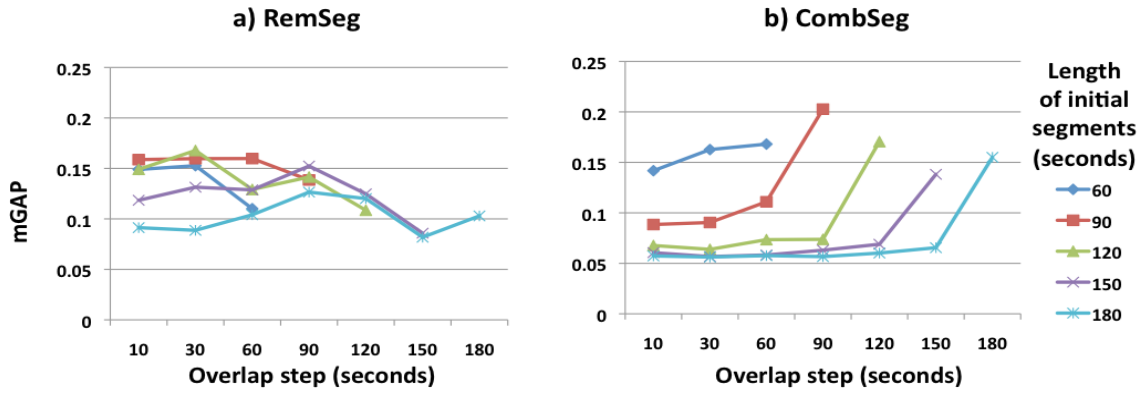


Figure 7.35: mGAP after filtering, manual transcript

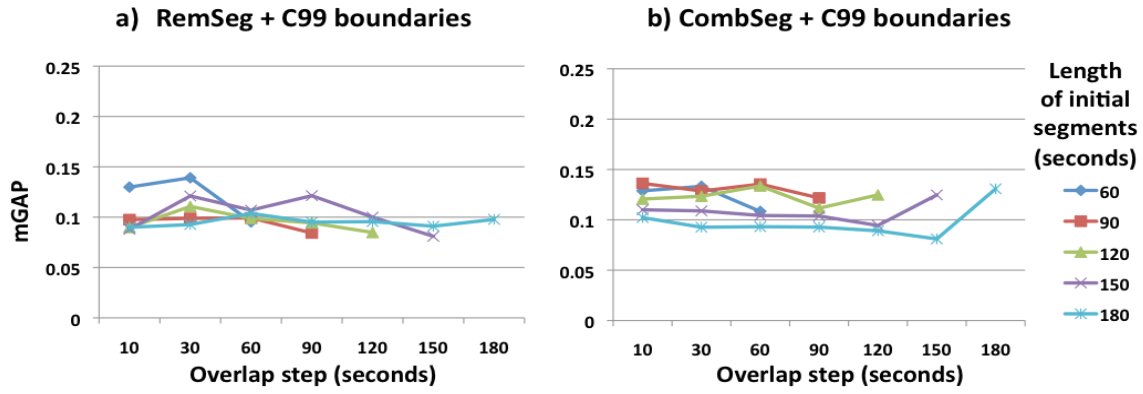


Figure 7.36: mGAP after filtering and use of C99 segment boundaries , manual transcript

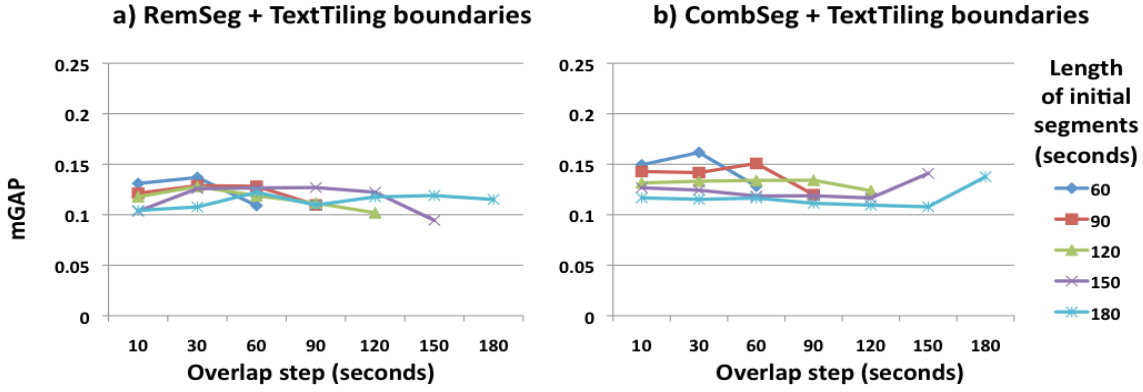


Figure 7.37: mGAP after filtering and use of TextTiling segment boundaries, manual transcript

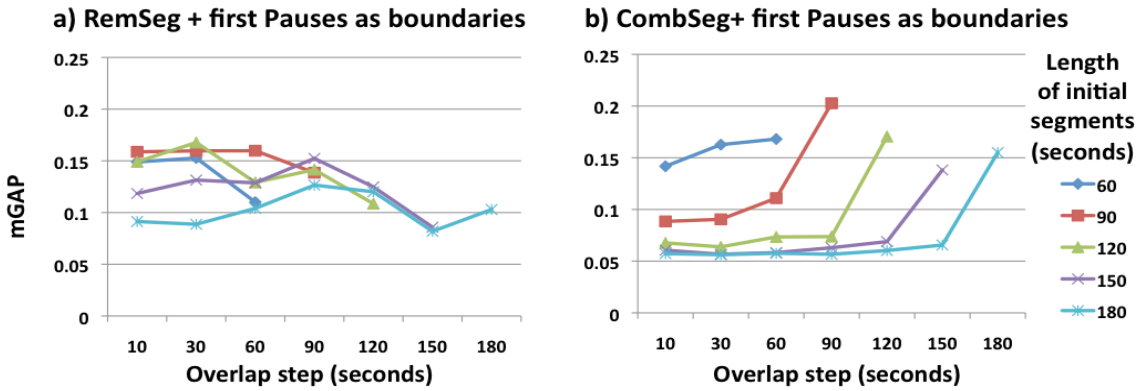


Figure 7.38: mGAP after filtering and use of first pause as segment boundaries, manual transcript

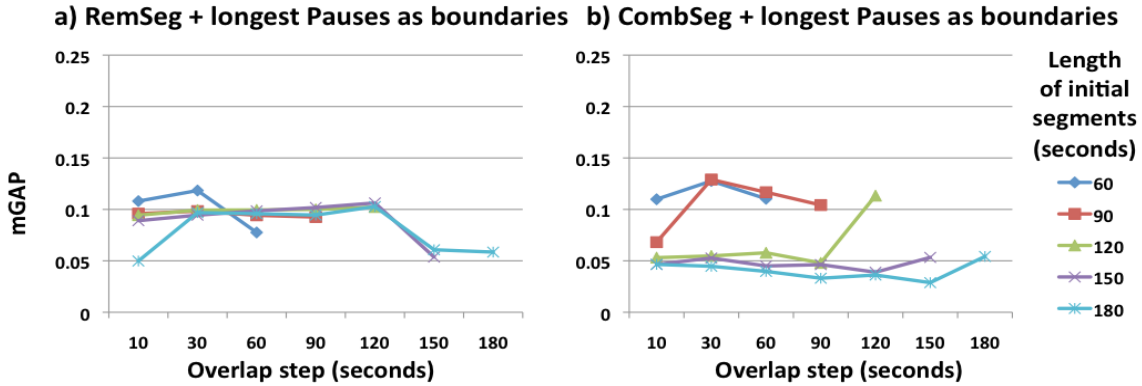


Figure 7.39: mGAP after filtering and use of longest pause as segment boundaries, manual transcript

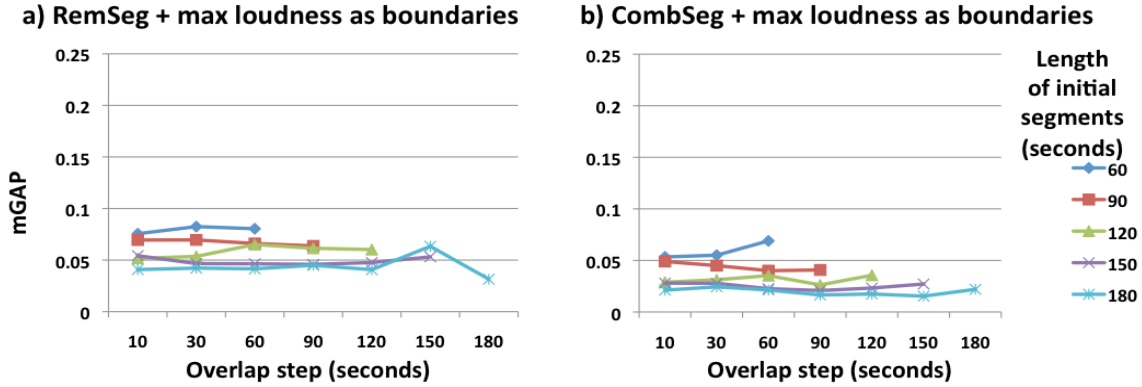


Figure 7.40: mGAP after filtering and use of a word with maximum loudness as segment boundaries, manual transcript

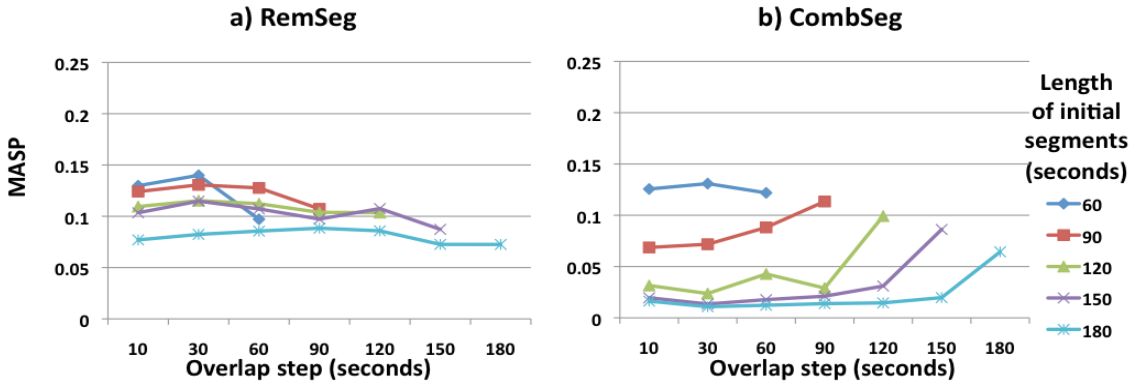


Figure 7.41: MASP after filtering, ASR transcript

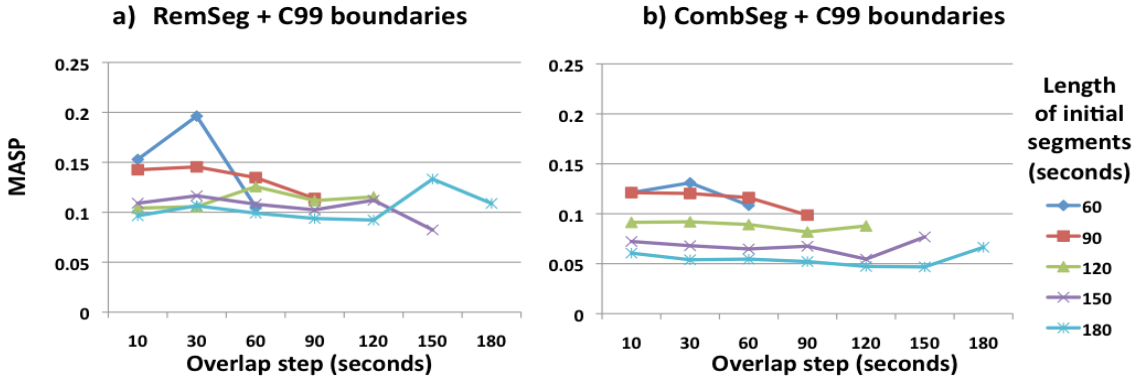


Figure 7.42: MASP after filtering and use of C99 segment boundaries, ASR transcript

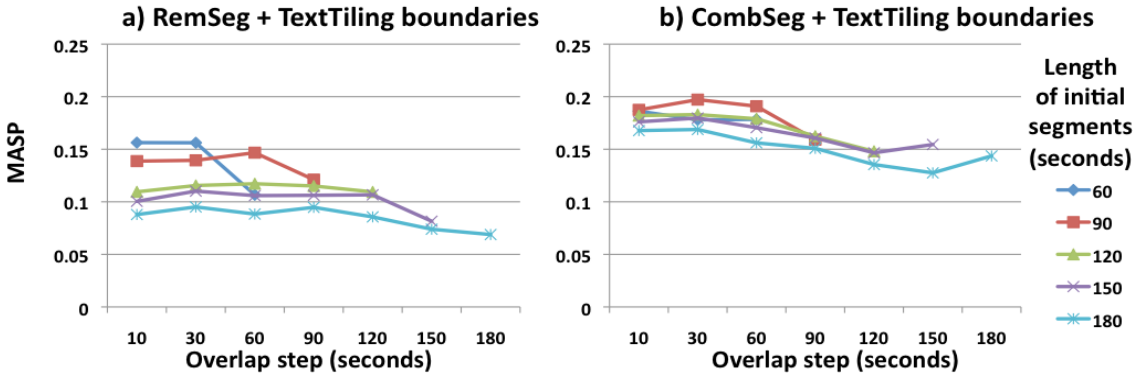


Figure 7.43: MASP after filtering and use of TextTiling segment boundaries, ASR transcript

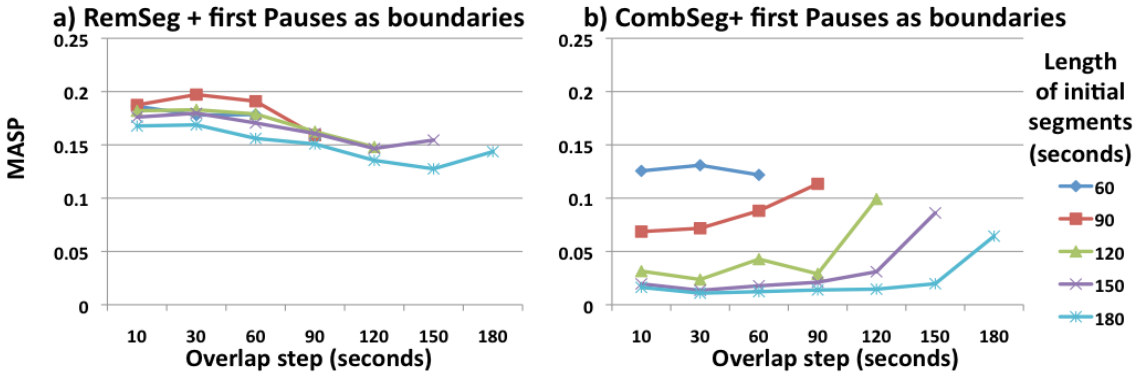


Figure 7.44: MASP after filtering and use of first pause as segment boundaries, ASR transcript

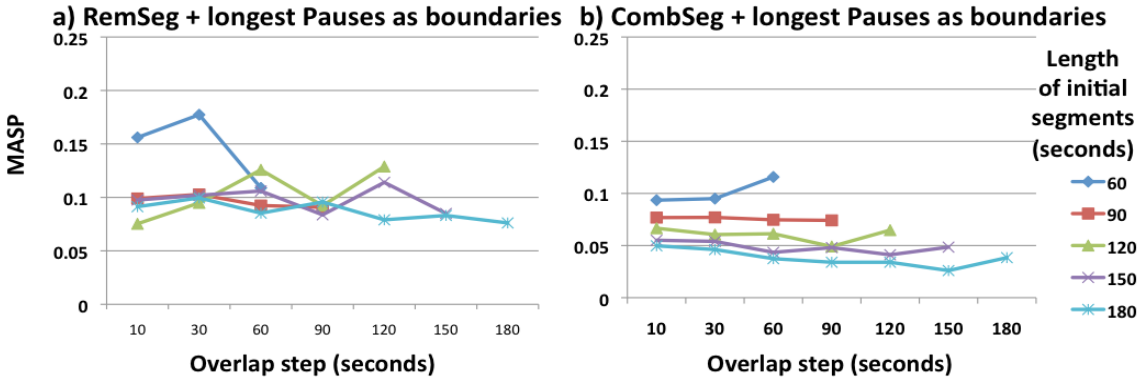


Figure 7.45: MASP after filtering and use of longest pause as segment boundaries, ASR transcript

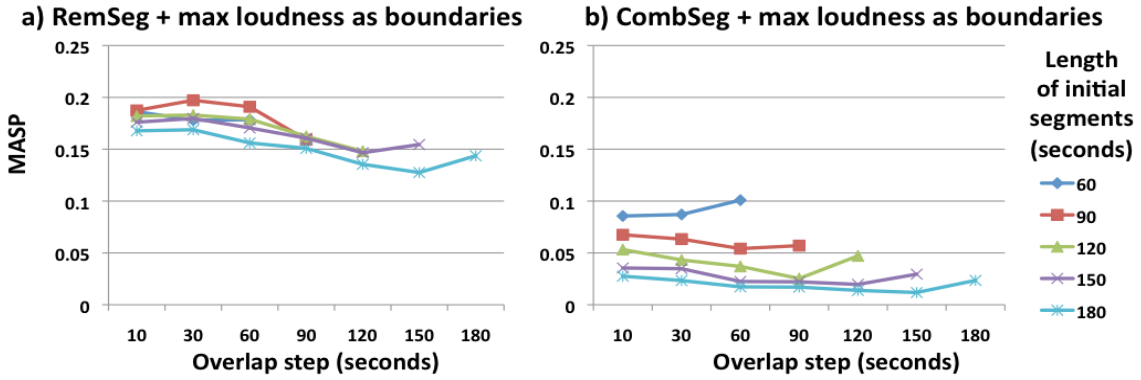


Figure 7.46: MASP after filtering and use of a word with maximum loudness as segment boundaries, ASR transcript

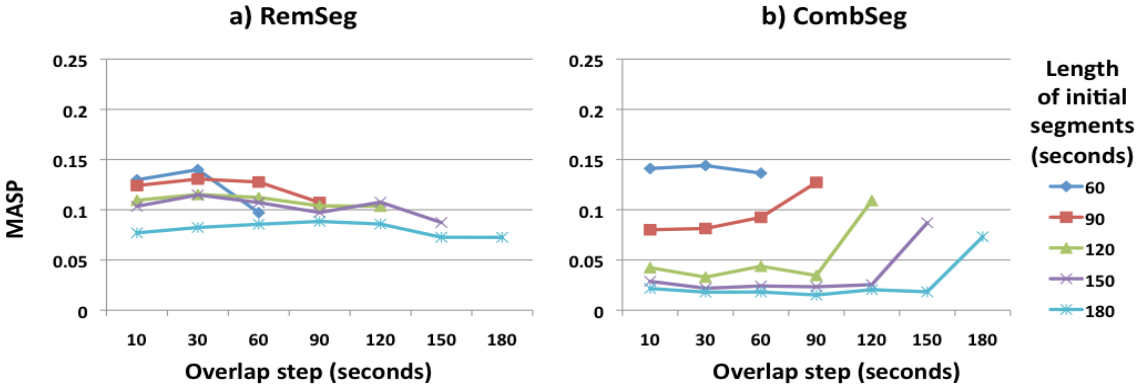


Figure 7.47: MASP after filtering, manual transcript

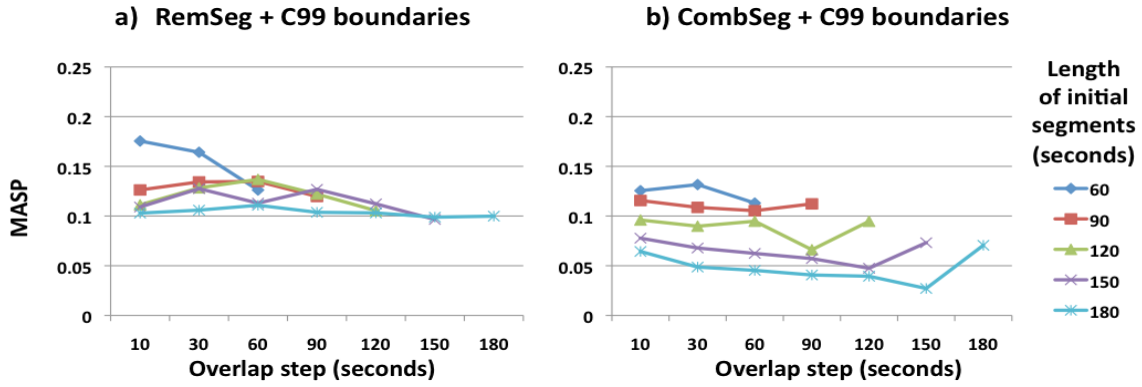


Figure 7.48: MASP after filtering and use of C99 segment boundaries , manual transcript

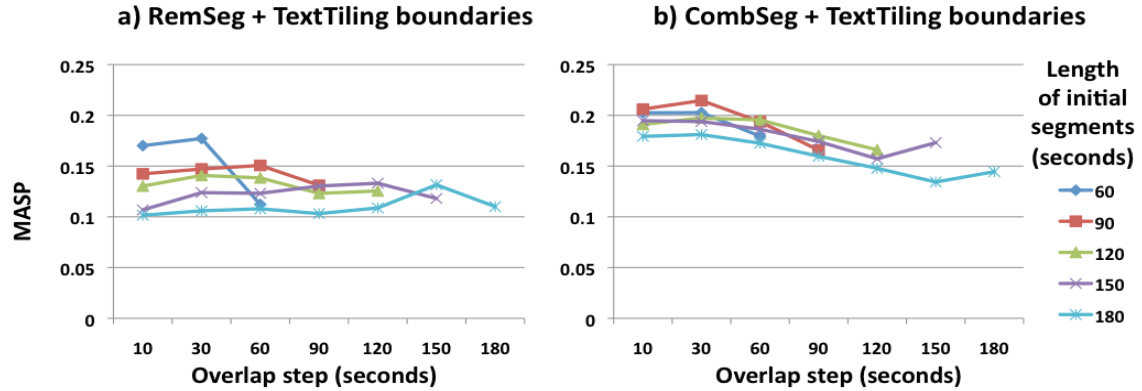


Figure 7.49: MASP after filtering and use of TextTiling segment boundaries, manual transcript

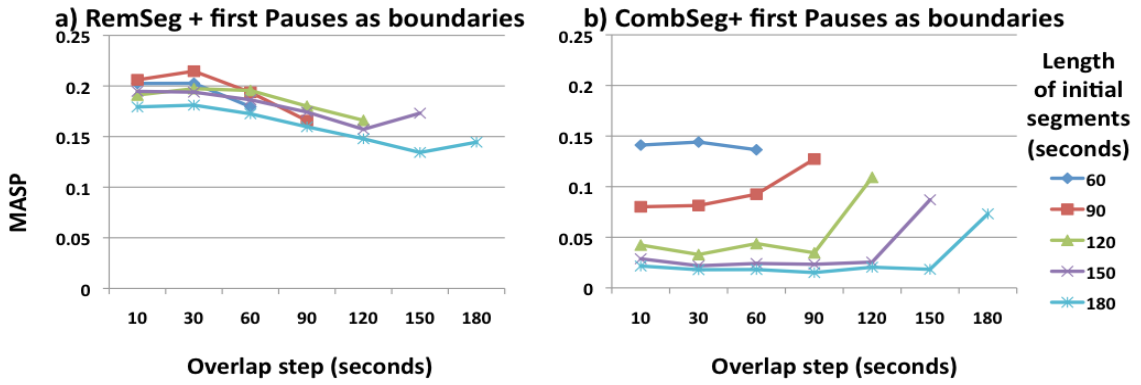


Figure 7.50: MASP after filtering and use of first pause as segment boundaries, manual transcript

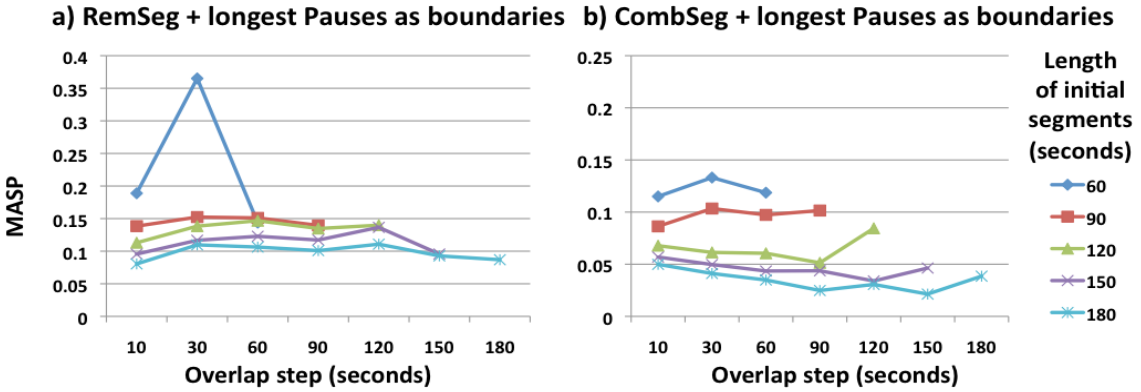


Figure 7.51: MASP after filtering and use of longest pause as segment boundaries, manual transcript

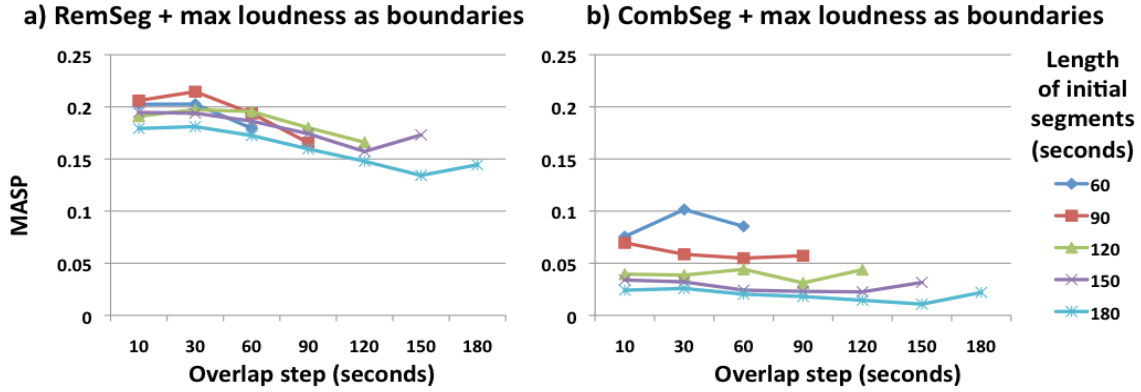


Figure 7.52: MASP after filtering and use of a word with maximum loudness as segment boundaries, manual transcript

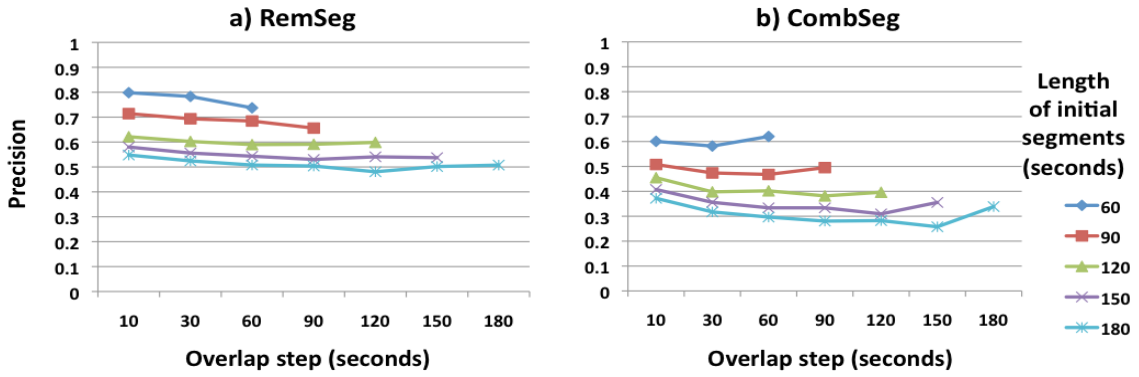


Figure 7.53: Average precision of relevant content in the segment after filtering, ASR transcript

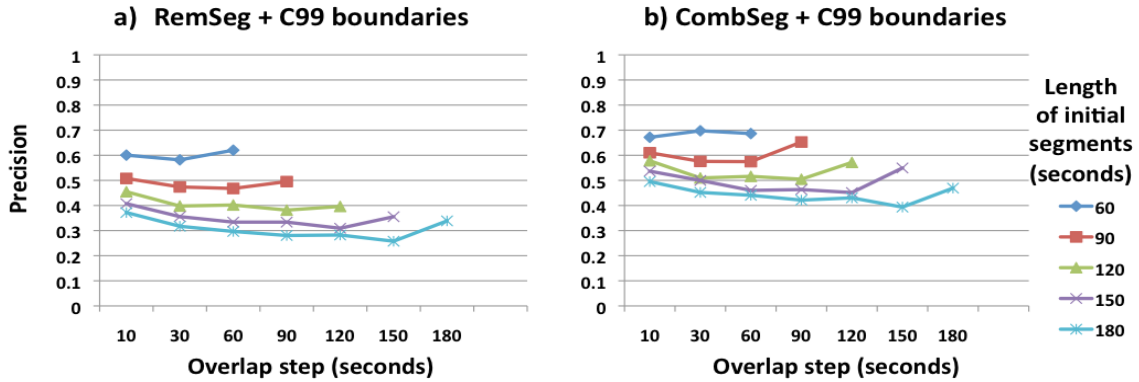


Figure 7.54: Average precision of relevant content in the segment after filtering and use of C99 segment boundaries, ASR transcript

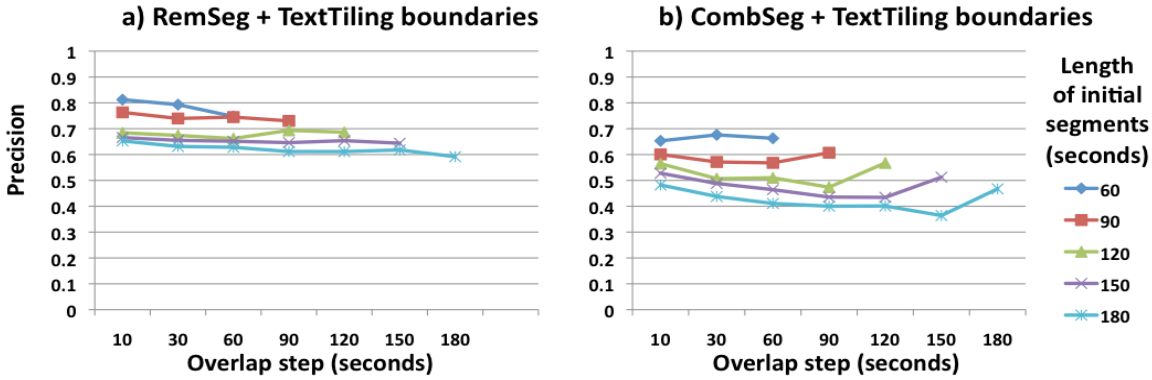


Figure 7.55: Average precision of relevant content in the segment after filtering and use of TextTiling segment boundaries, ASR transcript

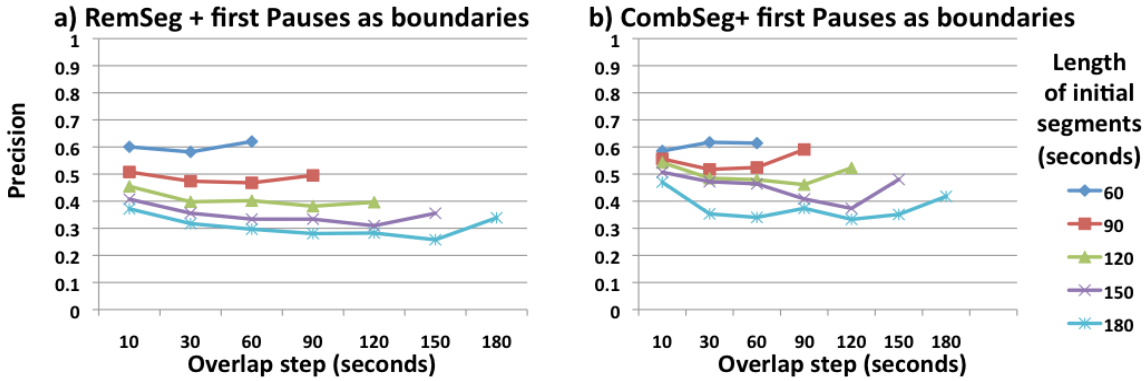


Figure 7.56: Average precision of relevant content in the segment after filtering and use of first pause as segment boundaries, ASR transcript

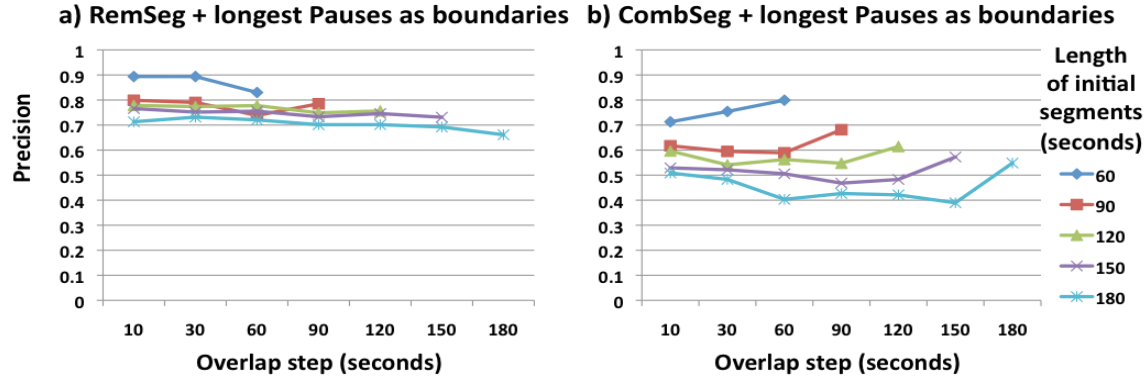


Figure 7.57: Average precision of relevant content in the segment after filtering and use of longest pause as segment boundaries, ASR transcript

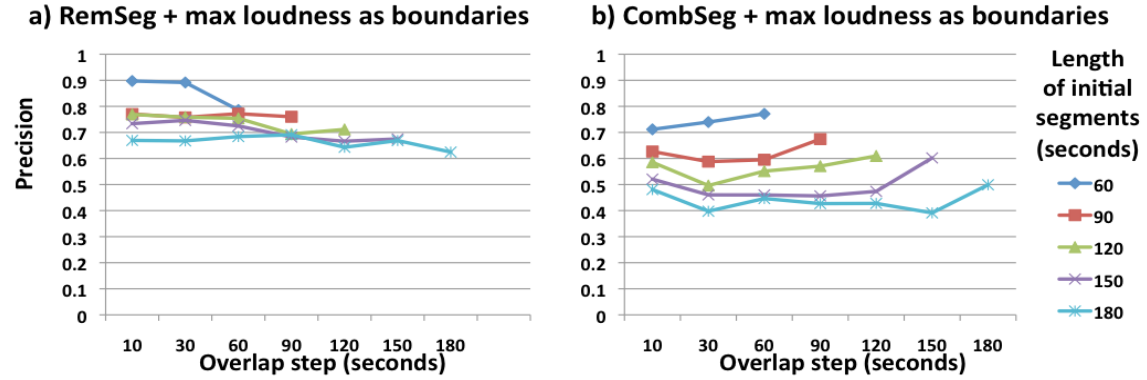


Figure 7.58: Average precision of relevant content in the segment after filtering and use of a word with maximum loudness as segment boundaries, ASR transcript

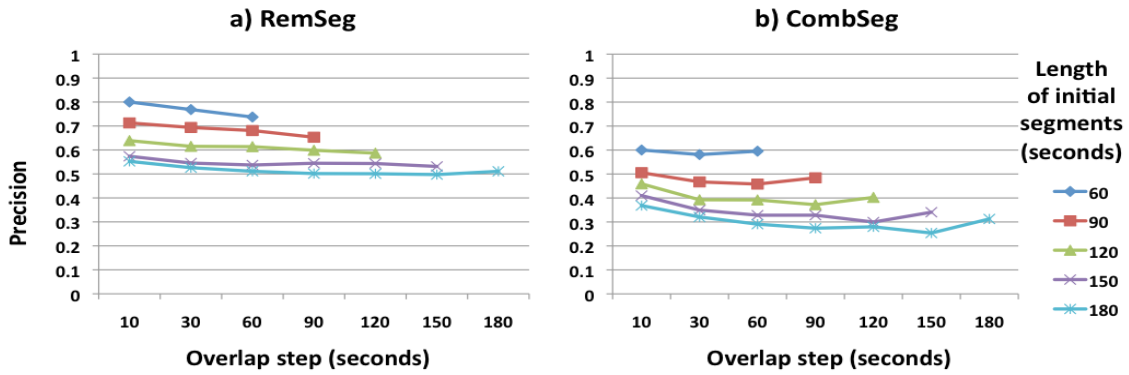


Figure 7.59: Average precision of relevant content in the segment after filtering, manual transcript

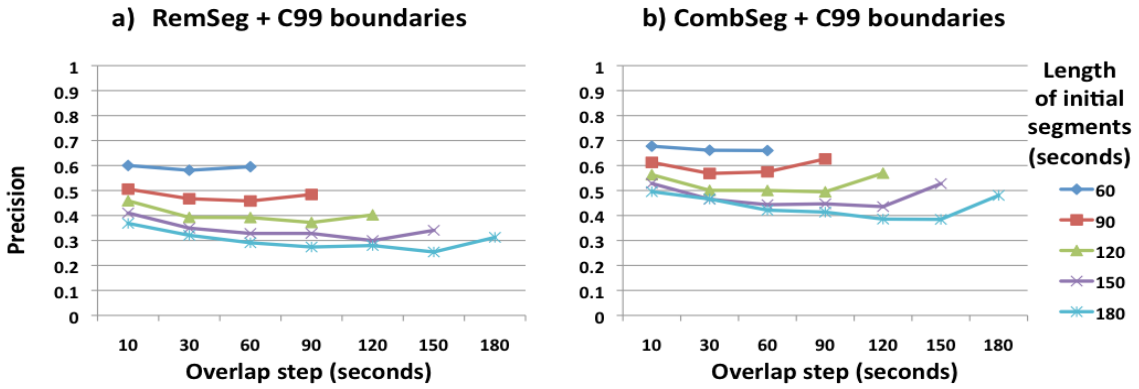


Figure 7.60: Average precision of relevant content in the segment after filtering and use of C99 segment boundaries , manual transcript

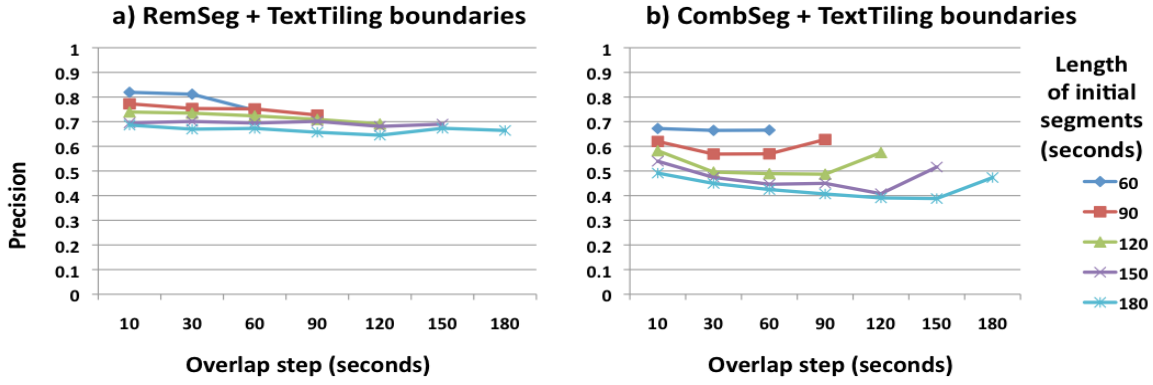


Figure 7.61: Average precision of relevant content in the segment after filtering and use of TextTiling segment boundaries, manual transcript

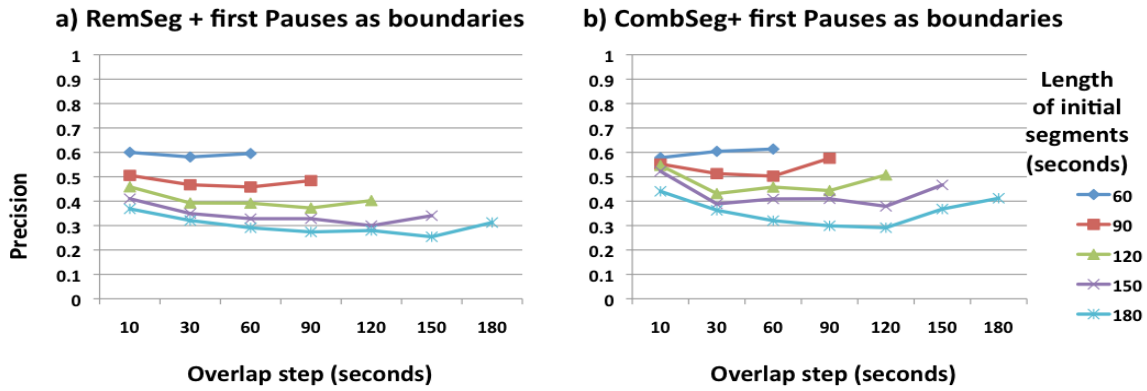


Figure 7.62: Average precision of relevant content in the segment after filtering and use of first pause as segment boundaries, manual transcript

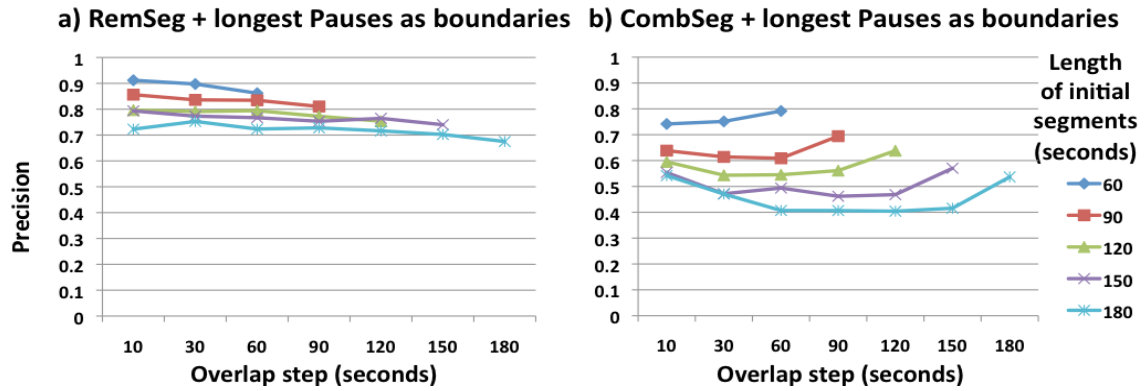


Figure 7.63: Average precision of relevant content in the segment after filtering and use of longest pause as segment boundaries, manual transcript

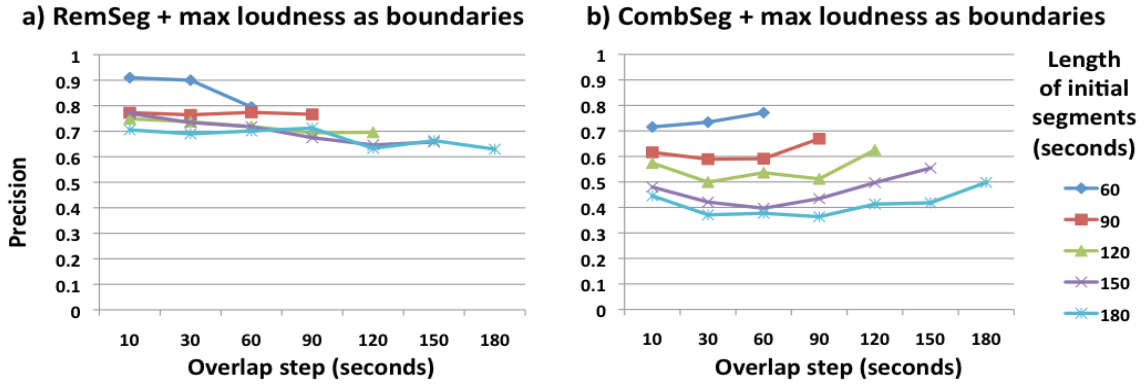


Figure 7.64: Average precision of relevant content in the segment after filtering and use of a word with maximum loudness as segment boundaries, manual transcript

7.4 Summary

In this chapter we implemented two approaches to filtering of ranked results containing overlap (removal and combination), and investigated potential improvements that can be gained when these segments are adjusted using information based on semantic (segmentation boundaries that are based on lexical cohesion) or acoustic (pauses, level of loudness) content.

Results on the AMI corpus collection demonstrate that CombSeg filtering achieves higher scores across a range of evaluation metrics for shorter segments which is consistent with the fMAP metric results for the NTCIR-9 SpokenDoc data. When boundary adjustment based on lexical cohesion methods is implemented, CombSeg results for longer segments are improved, since this affects precision of the relevant content in these segments and decreases the amount of non-relevant content that is present at the top retrieval ranks

Use of acoustic information, though helping to improve effectiveness in some cases, does not significantly change the performance.

In the next chapter we deal with imperfect segmentation results not by explicitly changing the boundaries, but by using the context of the segments to expand the segment list of terms that are used for retrieval. In cases when this document expansion experiments are carried out on segments with fixed length, boundary adjustment using lexical cohesion information is implemented to continue discussion of its usefulness in different frameworks of SCR.

Chapter 8

Document Expansion for SCR

Effective information retrieval relies on the match between the search terms in the query entered by the user and the relevant documents in the collection being searched. In the case of SCR, the indexed contents of the relevant documents may not match the query effectively due to ASR errors, common knowledge between conversation participants that allows them to avoid using relevant terms, imperfect segmentation of the content that brings too much of non-relevant data into a segment or divides the relevant content across segments, or the relevant document is too short to match the query reliably.

Previous work has shown that for manually segmented spoken documents based on broadcast news, document expansion using an external data collection containing documents closely related to the target collection improves the retrieval results, even in case of high word error rates, especially for short documents (Singhal and Pereira, 1999). This improvement can arise both due to the expansion terms introducing terms actually spoken which are missing due to ASR errors and to the introduction of additional terms that are closely related to the spoken topic, but were not used in the conversation.

In this chapter we explore the effect of using a standard document expansion method and new variations of combinations of document expansion methods to seek to improve SCR performance.

8.1 Document expansion in Information Retrieval

Although an external collection proved to be useful for document expansion (Singhal and Pereira, 1999), topically related external collections may not always be available. Therefore alternative methods of document expansion using the target retrieval collection itself have been explored in other studies on document expansion. The main idea is to use the document to be expanded as a query to the target collection, then to assume that the top N retrieved documents are relevant, and finally to extract additional terms from these documents, and to add them to the original document and reweight the terms in the document (Lavrenko and Croft, 2001; Tao et al., 2006).

In the case of audio/video recordings across time, variation in speech quality means that ASR errors will not be spread uniformly. Thus, once the content has been segmented, depending on the segmentation approach adopted, there may be search units of different length and varying level of errors. This means that the search units within the collection with lower error rate might serve as a potentially useful source for the effective document expansion.

8.2 Document Expansion for SCR

For our SCR experiments we implement previously developed text document expansion methods (Lavrenko and Croft, 2001), we examine the use of immediate context and introduce novel ways of combining them. These are summarised in Figure 8.1 and described in more details below:

1. **AD:** This approach assumes that the adjacent context of a document can provide useful contextual expansion since speakers may cover the same topic in the preceding or following parts of the conversation. The speakers might use different words that better describe the topic or pronounce the words relevant to the topic more clearly in other context that allow better ASR performance.

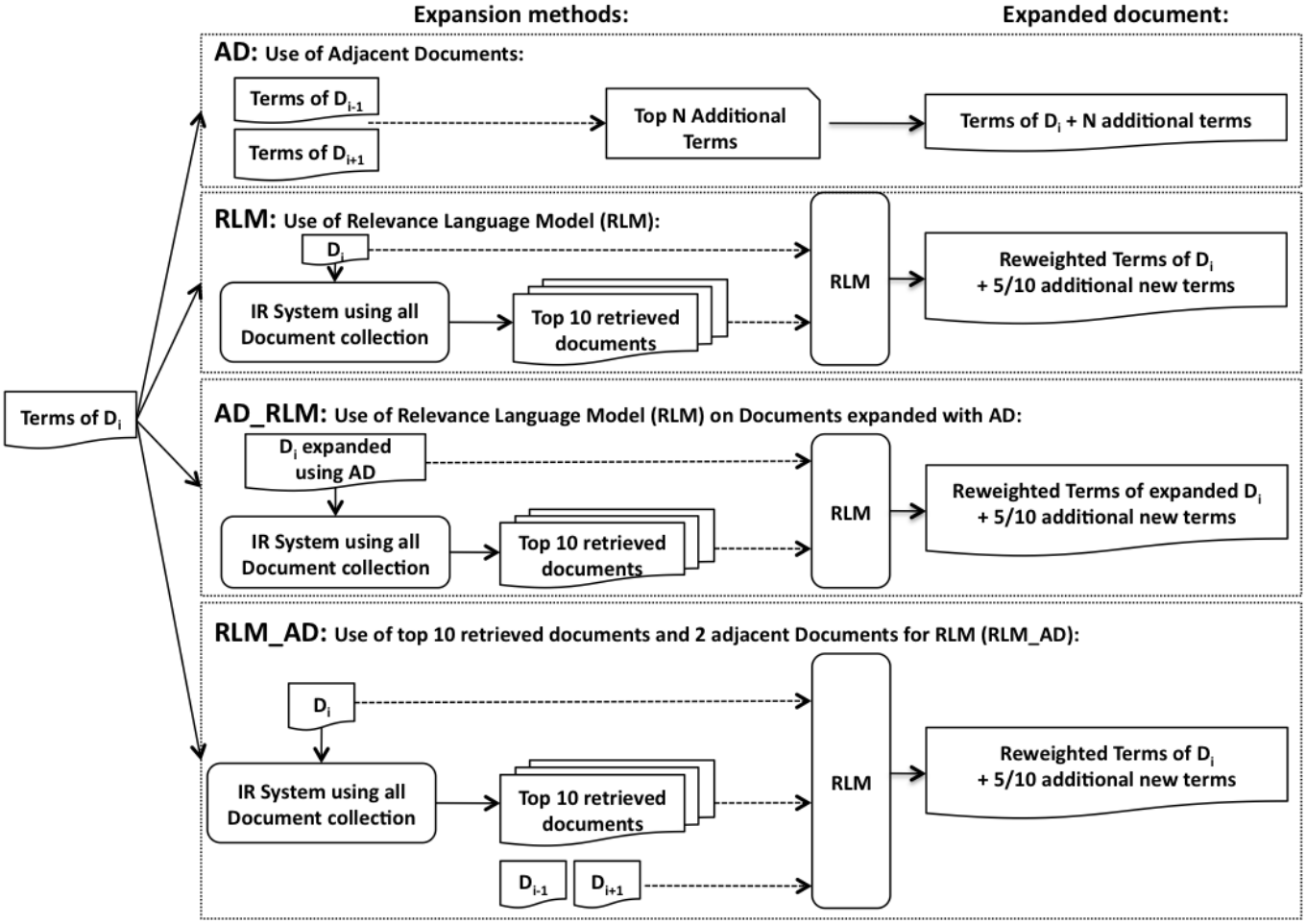


Figure 8.1: Different document expansion schemes.

We take terms $t_1, ..., t_n$ of 2 neigbouring documents, and calculate their weight using the standard *tf-idf* equation, as introduced in Equation 2.1 in Section 2.1.2. Afterwards we add N of these terms to the original document.

2. **RLM:** This approach assumes that there will be other segments covering the same topic within the collection which may provide useful expansion terms, therefore the initial document is used as a query to target the collection. The top 10 retrieved documents are considered to be of topical context, and are used to reweight the terms of the segment and as a source of expansion terms using standard relevance language model (RLM).

   Within this document expansion approach, we first run retrieval within the collection using each of its documents as a query. Afterwards, we calculate

the *tf-idf* weights of the terms within these documents, and add them to the each original document in the collection.

3. **AD_RLM:** This approach follows the same procedure as RLM with only one significant difference that the initial document already contains the expansion terms taken from the immediate context, i.e. AD procedure have already been carried out on it.

4. **RLM_AD:** This approach first retrieves a number of potential relevant documents within the collection (we use top 10 documents as in the RLM procedure), then use these 10 documents and 2 adjacent documents to reweight the terms of the initial document and add N expansion terms.

The AD approach might correct the errors in segmentation, as some parts of the relevant content may have been cut off, and with this method more relevant terms might be added to the document. It has previously been shown that documents with better ASR transcripts tend to be retrieved at higher ranks (Sanderson and Shou, 2007), which assumes that the methods using the retrieved top ranked documents (RLM, AD_RLM, RLM_AD) will tend to retrieve more reliable documents to use as source for additional terms.

We explored varying values of N terms to be added for expansion, and focus on using the N value equal to 10, since this gave optimal and consistent results across different segmentation and document expansion methods in trial experiments.

8.3 Retrieval setup

For the experiments described in this chapter, we chose four segmentation methods that produce segments of similar average length and represent different segmentation approaches: lexical cohesion based (C99, TextTiling), and 2 examples of fixed length segments using information about length in terms of transcript words (len_400, and len_nsw_150), see Tables 6.1 and 8.1. All retrieval and expansion experiments are

Segmentation Algorithm	Average Segment Length for manual transcript	
	words	seconds
C99	346.6	129.1
TextTiling	363.6	136.0
Len_400	400	154.8
Length_nsw_150	416	160.0

Table 8.1: Average segment length (words, seconds)

implemented and carried out within the framework of the open-source Terrier toolkit (Ounis et al., 2006) [1]. We use the Terrier implementation of Hiemstra's language modelling method (Hiemstra, 2001), which is the same as used in all previous experiments throughout this thesis[2], and here we vary with *lambda* value from the Equation 2.3 from its minimum to maximum.

8.4 Results and Analysis

In this section we describe our results for SCR experiments using document expansion. First, we overview overall results for all segmentation methods and varying settings of the IR system. Second, we show a few examples of system performance on different types of queries.

8.4.1 Document expansion performance with varying λ value

In order to evaluate the proposed methods we vary the settings of the retrieval language model, assigning more or less weight to the terms present in the document, as introduced in Equation 2.3, and calculate evaluation results using MAP, mGAP, MASP, and MASDWP metrics for the segmentation types defined in previous section.

The main parameter of the IR system based on language modelling that we have

[1] www.terrier.org

[2] We switch from SMART to Terrier in these experiments simply for the practical reason of methods implementation. There is no significant difference in results due to different retrieval systems.

to set up before running SCR experiments is the value of λ that defines the importance of the words from the query to be present in the result document (Hiemstra, 2001). We vary this value from 0.05 to 0.95 (the smaller is the value, the less importance the presence of the query term in a document has) to assess document expansion performance in these settings. Experimental results are presented in Figures 8.2-8.5 for MAP, Figures 8.6-8.9 for mGAP, Figures 8.10-8.13 for MASP, Figures 8.14-8.17 for MASDWP, and Figures 8.18-8.21 for average precision of the relevant content in the segments.

Across all metrics the variation of λ does not result in significant variation in in the runs based on segments with fixed length (len_400, len_nsw_150).

Runs based on segments with different length tend to have higher MAP and mGAP scores when λ value is higher, while their MASP and MASDWP scores decrease for the same increasing λ value.

When the λ value increases, the language model assigns more value to the presence of exact query terms in the document. The expansion terms added to the documents help to improve the ranking of relevant segments, as demonstrated by higher MAP and mGAP (Figures 8.2-8.3 and 8.6-8.7), although at the same time bring longer non-relevant segments into the top ranking positions. This is demonstrated by the fact that MASP and MASDWP values decrease (Figures 8.10-8.11 and 8.14-8.15), while precision of the relevant content stays on average almost the same (Figures 8.18-8.19).

8.4.2 Statistical significance across different document expansion methods

In order to assess the effectiveness of new document expansion methods in these SCR experiments, we check these results for statistical significance. Within each segmentation and transcript type, we select the best MAP score for each document expansion method, as well as a baseline that does not use any document expan-

sion; and calculate a Wilcoxon signed-rank test with significance level equal to 0.05 (Wilcoxon, 1945) for these MAP values, and further for mGAP, MASP, MASDWP values that this run with same λ value scores. We present results of these significance tests in Tables 8.2– 8.5 for lexical-cohesion based runs, and in Tables 8.6–8.9 for fixed length based runs, where statistically significant combinations are marked in bold font. Since we combine two segmentation types into one table, we put the λ values in the brackets for each run type, e.g. in Table 8.2 the title of the first line called "Baseline (0.4/0.25)" means that for the C99 run we use baseline run with the λ value equal to 0.4, and equal to 0.25 for the TextTiling run.

Overall, these tables show our *RLM_AD* document exansion method achieves statistically significant higher MAP scores, not only over the *Baseline*, but also over *AD* and *RLM* methods for the segments containing ASR transcript for all types of segmentation. In the case of the manual transcript, while *RLM_AD* approach outperforms the *Baseline*, *AD* and *RLM* methods for fixed length segmentation based runs, it has statistically significant improvement only over the *Baseline* and *RLM* methods for lexical cohesion based runs. This suggests that the terms from immediate context of a document have more importance in improving SCR performance for manual transcript. In the case of the ASR transcript these terms might not be correctly recognised, therefore the combination of the terms from the immediate and broader context within the collection is needed to enhance ranking of the relevant content with significant SCR effectiveness improvement. The same trend is followed in statistical significance of the mGAP scores. Statistical significance of improvement of the MASP and MASDWP scores is not as straighforward, as we take for comparison the runs with λ values that correspond to the highest MAP scores.

8.4.3 Effect of Document expansion on example queries

As the previous section showed, introduced document expansion methods can achieve statistical significant improvement over baseline results, in this section we focus on two example queries that demonstrate the processes that underlie these improve-

ments.

We take the runs with the λ value of 0.85 that produce highest MAP scores as shown on Figure 8.2, and these scores are also shown in details in Tables 8.10-8.14. We take our example queries from the C99 segmentation runs. One query is the same as the one used in Section 6.2, shown in Figure 4.3, with its scores being shown in Table 8.15. The other query example, see Figure 8.22, we use has larger improvement across the scores for all document expansion methods and overall significantly higher scores as compared with the other example query, as shown in Table 8.15.

Figures 8.24 and 8.26 show that the top 50 ranks for query 21 contains considerably more non-relevant content (the range of values for query 21 is between 2500 and 11237 seconds, while for query 13 it varies between 300 and 1820 seconds), while the range of the amount of the relevant content present in the top 50 fits within the same range of values for both queries.

Figures 8.23 and 8.25 show that the relevant content that is present in the top 50 retrieved segments, and may be moved down the retrieval list in one of the runs based on different transcript is segmented differently for the queries. In case of query 21, most of the content is contained in the segments that contain jump-in points, while most of the relevant content of the query 13 is spread across the segments that are following the ones with the jump-in point.

For further analysis we introduce Figures 8.27-8.28 that show average distance to the jump-in for the segments in the top 50 that contain relevant content, and Figures 8.29-8.30 that reflects average precision of the relevant content in those segments. As discussed before in Chapter 6, it can be seen that for the baseline runs and all document expansion methods the segments that are present in the top 50 for both asr_man and man runs have their strating points close to the actual jump-in point whether it is in the segment or in the previous one, Figures 8.27-8.28 these segments also contain high precision of the relevant content in them, Figures 8.29-8.30. RLM and RLM_AD methods move asr_man transcript based runs within the top 50 ranks

($a\_m <= 50 < m$) the segments that have lower precision of the relevant content and are further from the jump-in points for the query 13, see Figures 8.29 and 8.27.

8.5 Discussion

We have shown that our document expansion approaches using immediately adjacent and topical collection context improves the retrieval effectiveness. The use of immediate context of different type allows to address the issues of incorrect segmentation, as potentially important terms from the context are being added to the segment. Combination of these adjacent documents with those from the topical collection context enables addition of the terms that potentially come from the segments with better ASR quality, and thus are more reliable.

Improvement over the baseline and state-of-the art document expansion methods is statistically significant for fixed length segmentation and most of the runs that use lexically coherent segments. Variation of the IR system settings that assigns importance of the term presence in the documents shows consistently better performance for the document expansion runs across a set of metrics.

In the next chapter we overview the work described in all previous chapters of the thesis and summarize the insights on the SCR framework that this work covered. Basing our discussion on the gained knowledge we introduce future work potetial directions.

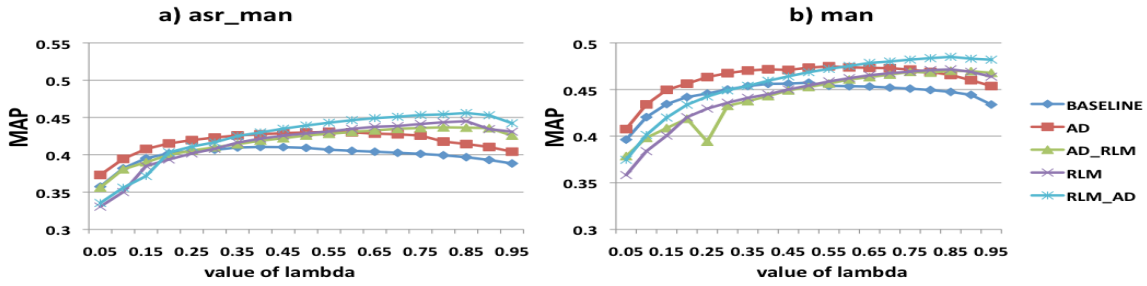


Figure 8.2: MAP for various document expansion methods based on the c99 segmentation with varying λ value

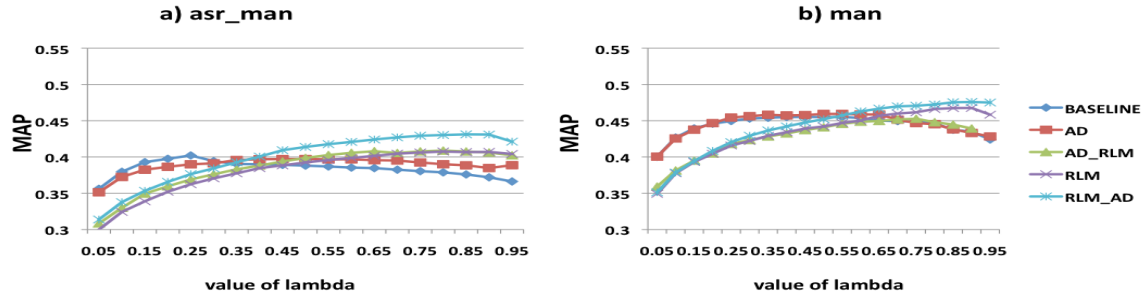


Figure 8.3: MAP for various document expansion methods based on the TextTiling segmentation with varying λ value

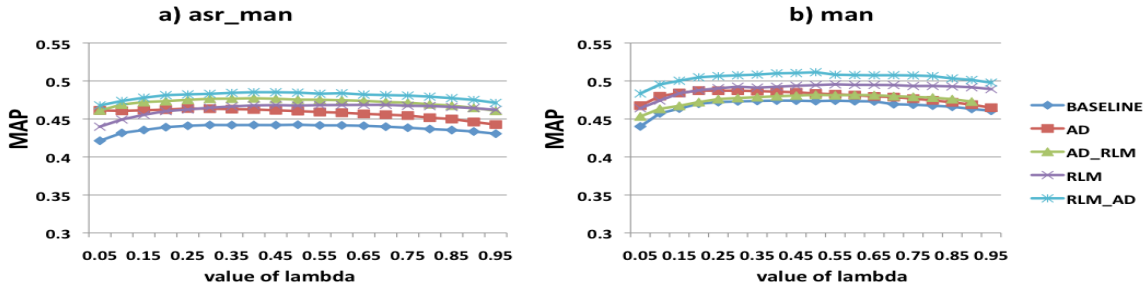


Figure 8.4: MAP for various document expansion methods based on the len_400 segmentation with varying λ value

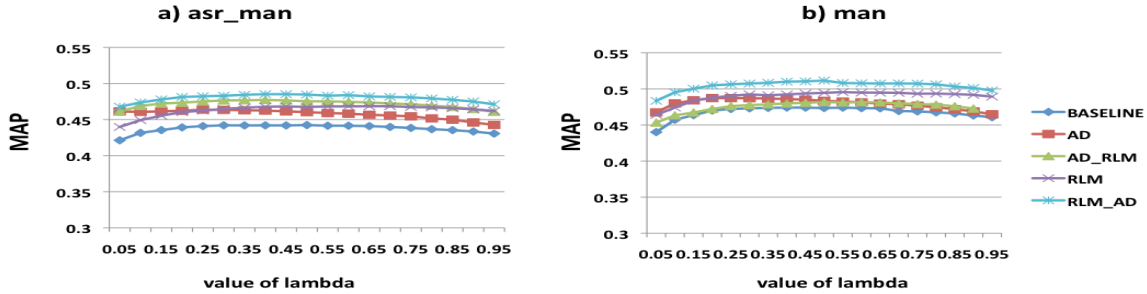


Figure 8.5: MAP for various document expansion methods based on the len_nsw_150 segmentation with varying λ value

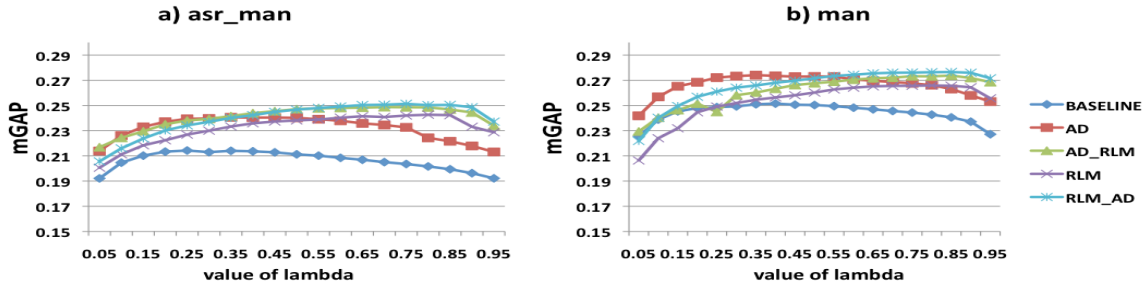


Figure 8.6: mGAP for various document expansion methods based on the c99 segmentation with varying λ value

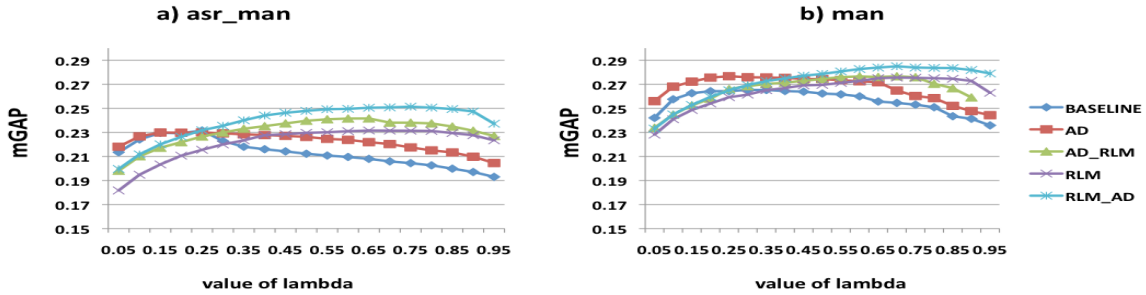


Figure 8.7: mGAP for various document expansion methods based on the TextTiling segmentation with varying λ value

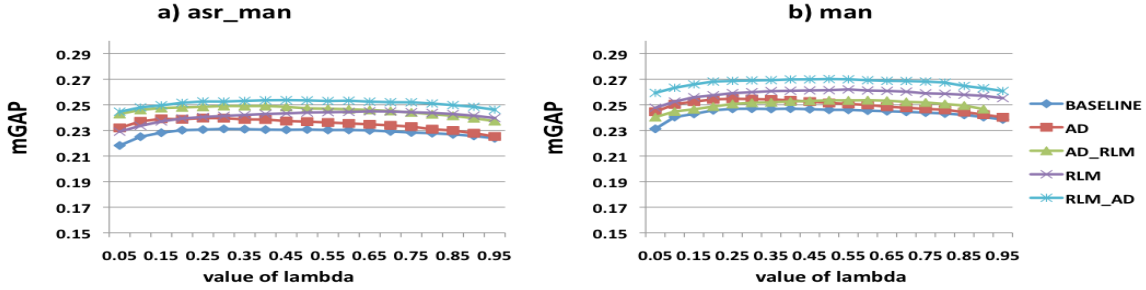


Figure 8.8: mGAP for various document expansion methods based on the len_400 segmentation with varying λ value

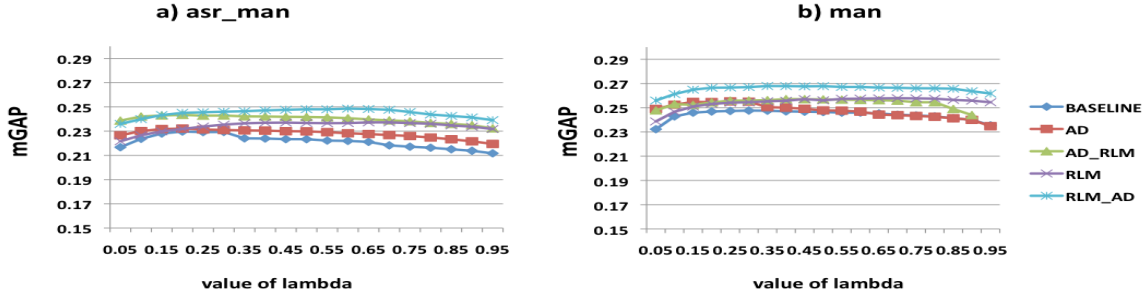


Figure 8.9: mGAP for various document expansion methods based on the len_nsw_150 segmentation with varying λ value

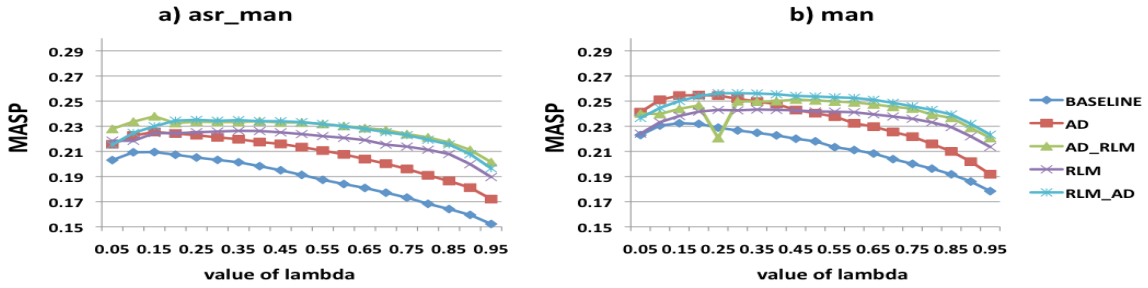


Figure 8.10: MASP for various document expansion methods based on the c99 segmentation with varying λ value

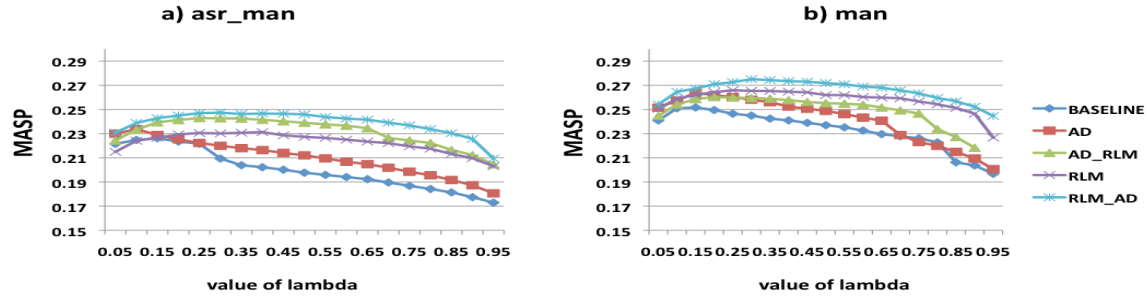


Figure 8.11: MASP for various document expansion methods based on the TextTiling segmentation with varying λ value

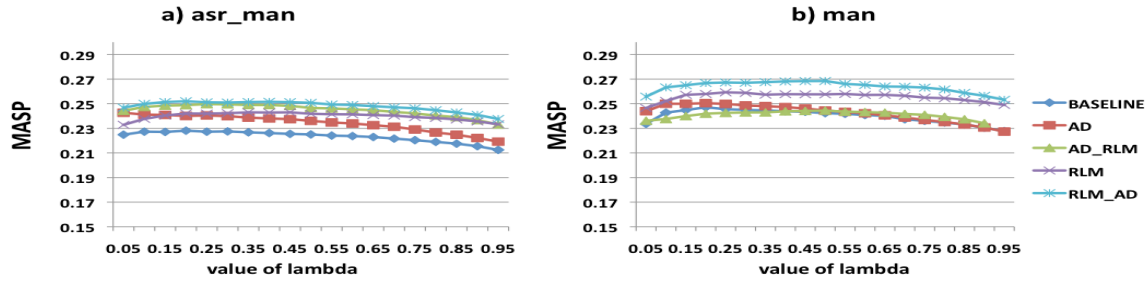


Figure 8.12: MASP for various document expansion methods based on the len_400 segmentation with varying λ value

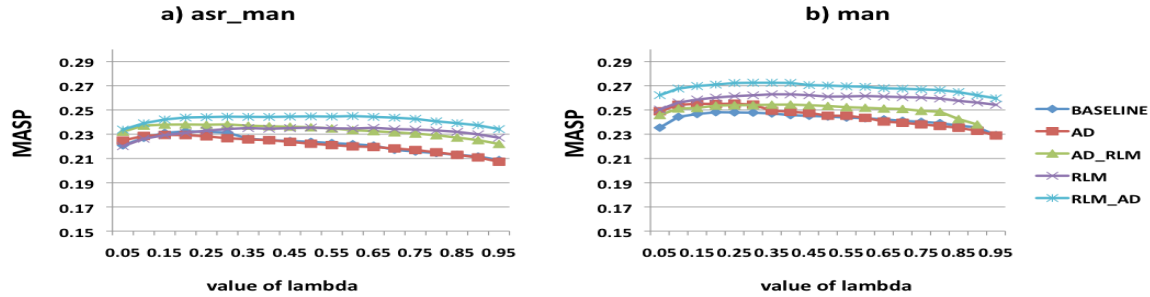


Figure 8.13: MASP for various document expansion methods based on the len_nsw_150 segmentation with varying λ value

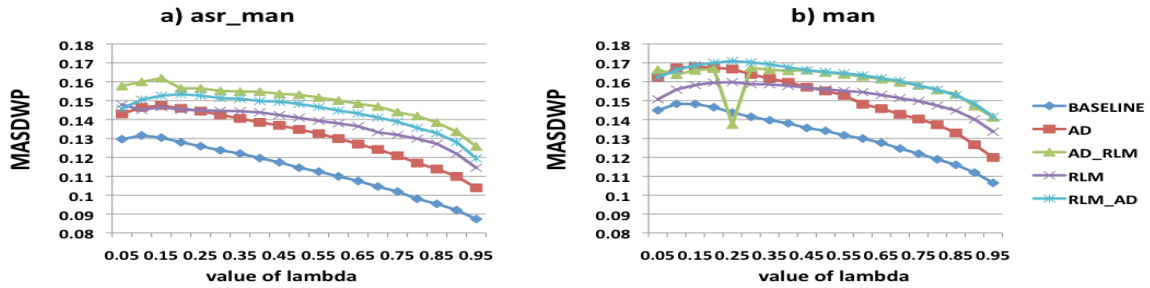


Figure 8.14: MASDWP for various document expansion methods based on the c99 segmentation with varying λ value

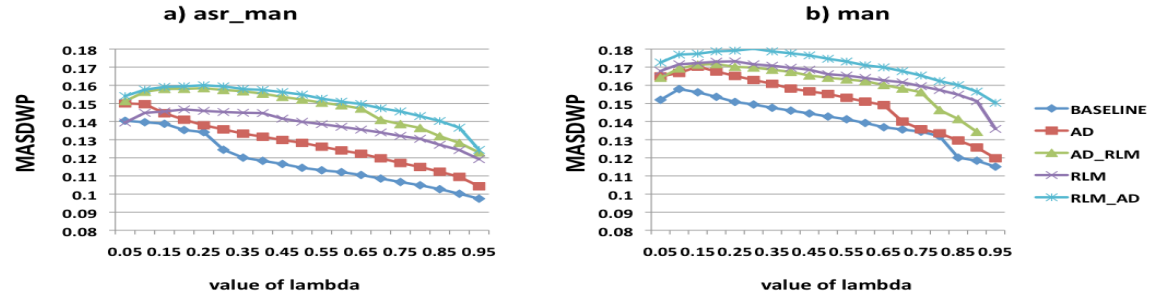


Figure 8.15: MASDWP for various document expansion methods based on the TextTiling segmentation with varying λ value

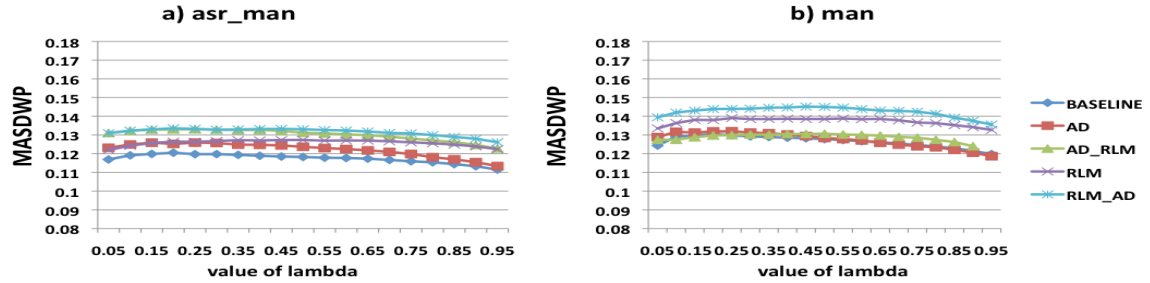


Figure 8.16: MASDWP for various document expansion methods based on the len_400 segmentation with varying λ value

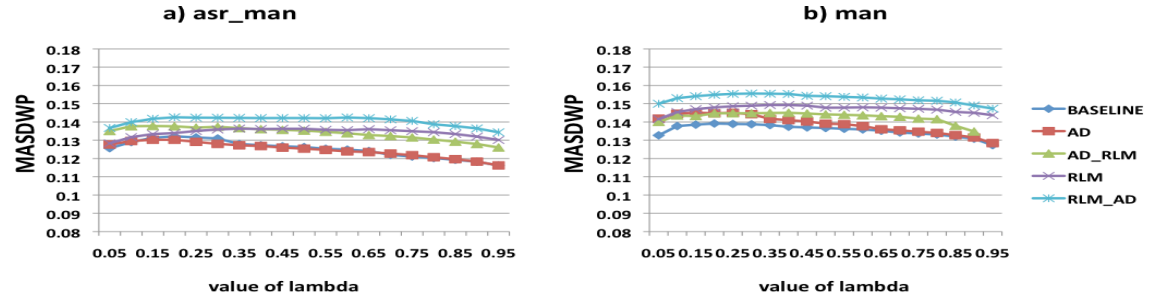


Figure 8.17: MASDWP for various document expansion methods based on the len_nsw_150 segmentation with varying λ value

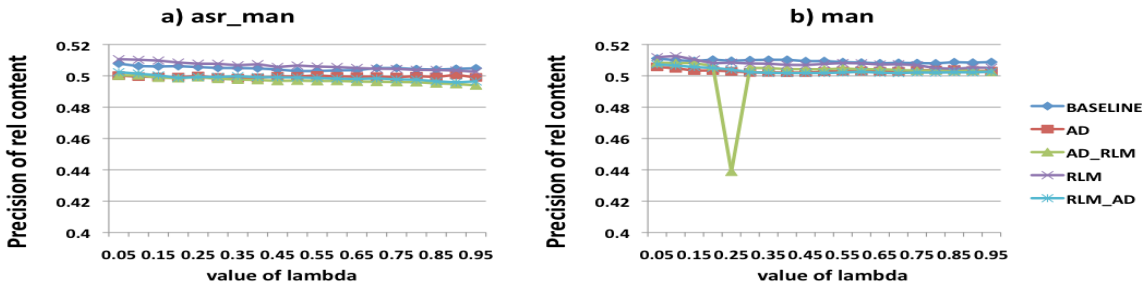


Figure 8.18: Precision of relevant content for various document expansion methods based on the c99 segmentation with varying λ value

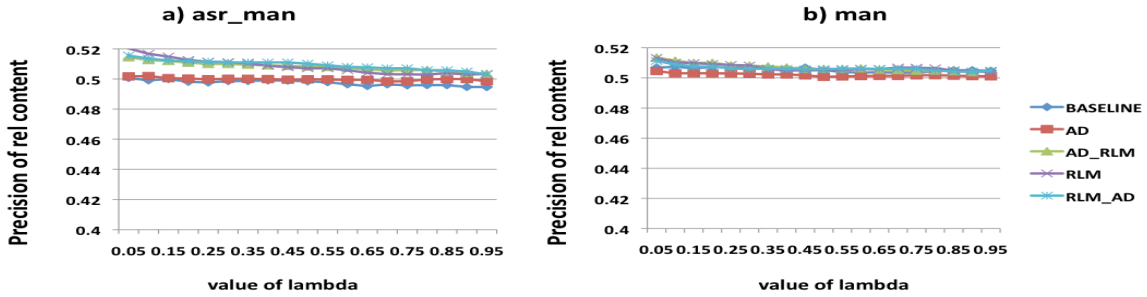


Figure 8.19: Precision of relevant content for various document expansion methods based on the TextTiling segmentation with varying λ value

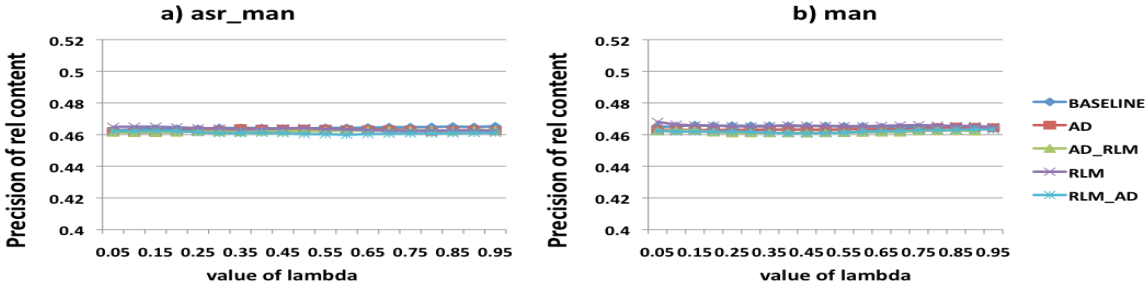


Figure 8.20: Precision of relevant contentfor various document expansion methods based on the len_400 segmentation with varying λ value

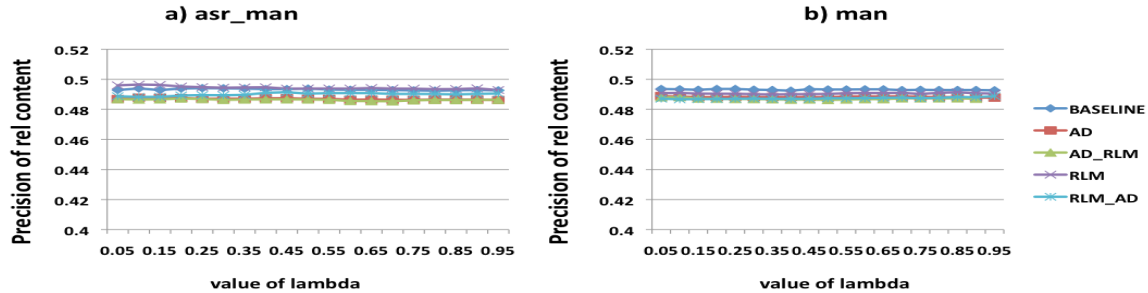


Figure 8.21: Precision of relevant contentfor various document expansion methods based on the len_nsw_150 segmentation with varying λ value

MAP: ASR_MAN				
Types of document expansion methods				
	AD (0.55/0.5)	AD_RLM (0.85/0.65)	RLM (0.85/0.8)	RLM_AD (0.85/0.85)
Baseline (0.4/0.25)	C99, TT	C99, TT	**C99, TT**	**C99, TT**
AD (0.55/0.5)	–	C99, TT	C99, TT	**C99, TT**
RLM (0.85/0.8)	C99, TT	C99, TT	–	**C99, TT**
MAP: MAN				
Types of document expansion methods				
	AD (0.55/0.5)	AD_RLM (0.85/0.75)	RLM (0.85/0.9)	RLM_AD (0.85/0.9)
Baseline (0.5/0.55)	C99, TT	C99, TT	**C99, TT**	**C99, TT**
AD (0.55/0.5)	–	C99, TT	C99, TT	C99, TT
RLM (0.85/0.9)	C99, TT	**C99, TT**	–	**C99, TT**

Table 8.2: Statistical significance of the highest MAP scores for lexical-cohesion based runs (C99, TextTiling), statistical significance in bold.

mGAP: ASR_MAN				
Types of document expansion methods				
	AD (0.55/0.5)	AD_RLM (0.85/0.65)	RLM (0.85/0.8)	RLM_AD (0.85/0.85)
Baseline (0.4/0.25)	**C99**, TT	**C99**, TT	**C99, TT**	**C99, TT**
AD (0.55/0.5)	–	**C99, TT**	C99, TT	**C99, TT**
RLM (0.85/0.8)	C99, TT	C99, TT	–	**C99, TT**
mGAP: MAN				
Types of document expansion methods				
	AD (0.55/0.5)	AD_RLM (0.85/0.75)	RLM (0.85/0.9)	RLM_AD (0.85/0.9)
Baseline (0.5/0.55)	C99, TT	C99, TT	**C99, TT**	**C99, TT**
AD (0.55/0.5)	–	C99, TT	C99, TT	C99, TT
RLM (0.85/0.9)	C99, TT	C99, TT	–	**C99, TT**

Table 8.3: Statistical significance of the highest mGAP scores for lexical-cohesion based runs (C99, TextTiling), statistical significance in bold.

<top>
<num> Number: 13
<title>
Tool Training
Try out whiteboard
Every participant should draw their favorite animal and sum up their favorite characteristics of that animal
<top>

Figure 8.22: Example of information present in a slide, AMI Corpus (query 13).

MASP: ASR_MAN				
Types of document expansion methods				
	AD (0.55/0.5)	AD_RLM (0.85/0.65)	RLM (0.85/0.8)	RLM_AD (0.85/0.85)
Baseline (0.4/0.25)	C99, TT	**C99, TT**	**C99,** TT	**C99, TT**
AD (0.55/0.5)	–	**C99, TT**	C99, TT	C99, **TT**
RLM (0.85/0.8)	C99, TT	C99, **TT**	–	**C99, TT**
MASP: MAN				
Types of document expansion methods				
	AD (0.55/0.5)	AD_RLM (0.85/0.75)	RLM (0.85/0.9)	RLM_AD (0.85/0.9)
Baseline (0.5/0.55)	**C99, TT**	**C99, TT**	**C99, TT**	**C99, TT**
AD (0.55/0.5)	–	C99, TT	C99, TT	C99, TT
RLM (0.85/0.9)	C99, TT	C99, TT	–	**C99, TT**

Table 8.4: Statistical significance of the highest MASP scores for lexical-cohesion based runs (C99, TextTiling), statistical significance in bold.

MASDWP: ASR_MAN				
Types of document expansion methods				
	AD (0.55/0.5)	AD_RLM (0.85/0.65)	RLM (0.85/0.8)	RLM_AD (0.85/0.85)
Baseline (0.4/0.25)	**C99,** TT	**C99, TT**	**C99,** TT	**C99,** TT
AD (0.55/0.5)	–	**C99, TT**	C99, TT	C99, **TT**
RLM (0.85/0.8)	C99, TT	**C99, TT**	–	**C99, TT**
MASDWP: MAN				
Types of document expansion methods				
	AD (0.55/0.5)	AD_RLM (0.85/0.75)	RLM (0.85/0.9)	RLM_AD (0.85/0.9)
Baseline (0.5/0.55)	**C99, TT**	**C99, TT**	**C99, TT**	**C99, TT**
AD (0.55/0.5)	–	C99, **TT**	**C99,** TT	C99, TT
RLM (0.85/0.9)	**C99**, TT	C99, **TT**	–	**C99, TT**

Table 8.5: Statistical significance of the best retrieval scores (MASDWP) for lexical-cohesion based runs (C99, TextTiling), statistical significance in bold.

MAP: ASR_MAN				
Types of document expansion methods				
	AD (0.25/0.2)	AD_RLM (0.4/0.3)	RLM (0.7/0.7)	RLM_AD (0.4/0.6)
Baseline (0.5/0.3)	len_400, len_nsw_150	**len_400, len_nsw_150**	**len_400, len_nsw_150**	**len_400, len_nsw_150**
AD (0.25/0.2)	– –	**len_400, len_nsw_150**	len_400, len_nsw_150	**len_400, len_nsw_150**
RLM (0.4/0.6)	len_400, len_nsw_150	len_400, len_nsw_150	– –	**len_400, len_nsw_150**
MAP: MAN				
Types of document expansion methods				
	AD (0.25/0.2)	AD_RLM (0.4/0.3)	RLM (0.7/0.7)	RLM_AD (0.4/0.6)
Baseline (0.5/0.3)	len_400, len_nsw_150	len_400, len_nsw_150	**len_400, len_nsw_150**	**len_400, len_nsw_150**
AD (0.25/0.2)	– –	len_400, len_nsw_150	len_400, len_nsw_150	**len_400, len_nsw_150**
RLM (0.4/0.6)	len_400, len_nsw_150	**len_400**, len_nsw_150	– –	**len_400, len_nsw_150**

Table 8.6: Statistical significance of the highest MAP scores for fixed length based runs (len_400, len_nsw_150), statistical significance in bold.

mGAP: ASR_MAN				
Types of document expansion methods				
	AD (0.25/0.2)	AD_RLM (0.4/0.3)	RLM (0.7/0.7)	RLM_AD (0.4/0.6)
Baseline (0.5/0.3)	len_400, len_nsw_150	**len_400, len_nsw_150**	**len_400, len_nsw_150**	**len_400, len_nsw_150**
AD (0.25/0.2)	– –	**len_400, len_nsw_150**	len_400, len_nsw_150	**len_400, len_nsw_150**
RLM (0.4/0.6)	len_400, len_nsw_150	len_400, len_nsw_150	– –	**len_400, len_nsw_150**
mGAP: MAN				
Types of document expansion methods				
	AD (0.25/0.2)	AD_RLM (0.4/0.3)	RLM (0.7/0.7)	RLM_AD (0.4/0.6)
Baseline (0.5/0.3)	len_400, **len_nsw_150**	len_400, len_nsw_150	**len_400, len_nsw_150**	**len_400, len_nsw_150**
AD (0.25/0.2)	– –	len_400, len_nsw_150	**len_400,** len_nsw_150	**len_400, len_nsw_150**
RLM (0.4/0.6)	**len_400,** len_nsw_150	**len_400**, len_nsw_150	– –	len_400, **len_nsw_150**

Table 8.7: Statistical significance of the highest mGAP retrieval scores for fixed length based runs (len_400, len_nsw_150), statistical significance in bold.

MASP: ASR_MAN				
Types of document expansion methods				
	AD (0.25/0.2)	AD_RLM (0.4/0.3)	RLM (0.7/0.7)	RLM_AD (0.4/0.6)
Baseline (0.5/0.3)	**len_400, len_nsw_150**	**len_400, len_nsw_150**	**len_400, len_nsw_150**	**len_400, len_nsw_150**
AD (0.25/0.2)	– –	len_400, len_nsw_150	len_400, len_nsw_150	**len_400, len_nsw_150**
RLM (0.4/0.6)	len_400, len_nsw_150	len_400, len_nsw_150	– –	**len_400, len_nsw_150**
MASP: MAN				
Types of document expansion methods				
	AD (0.25/0.2)	AD_RLM (0.4/0.3)	RLM (0.7/0.7)	RLM_AD (0.4/0.6)
Baseline (0.5/0.3)	len_400, len_nsw_150	len_400, len_nsw_150	**len_400, len_nsw_150**	**len_400, len_nsw_150**
AD (0.25/0.2)	– –	len_400, len_nsw_150	**len_400,** len_nsw_150	**len_400, len_nsw_150**
RLM (0.4/0.6)	**len_400,** len_nsw_150	**len_400**, **len_nsw_150**	– –	**len_400, len_nsw_150**

Table 8.8: Statistical significance of the highest MASP retrieval scores for fixed length based runs (len_400, len_nsw_150), statistical significance in bold.

MASDWP: ASR_MAN				
Types of document expansion methods				
	AD (0.25/0.2)	AD_RLM (0.4/0.3)	RLM (0.7/0.7)	RLM_AD (0.4/0.6)
Baseline (0.5/0.3)	len_400, len_nsw_150	**len_400, len_nsw_150**	**len_400, len_nsw_150**	**len_400, len_nsw_150**
AD (0.25/0.2)	– –	len_400, len_nsw_150	len_400, len_nsw_150	len_400, **len_nsw_150**
RLM (0.4/0.6)	len_400, len_nsw_150	len_400, len_nsw_150	– –	**len_400, len_nsw_150**
MASDWP: MAN				
Types of document expansion methods				
	AD (0.25/0.2)	AD_RLM (0.4/0.3)	RLM (0.7/0.7)	RLM_AD (0.4/0.6)
Baseline (0.5/0.3)	len_400, len_nsw_150	len_400, len_nsw_150	**len_400, len_nsw_150**	**len_400, len_nsw_150**
AD (0.25/0.2)	– –	len_400, len_nsw_150	**len_400,** len_nsw_150	**len_400, len_nsw_150**
RLM (0.4/0.6)	**len_400,** len_nsw_150	**len_400**, len_nsw_150	– –	len_400, **len_nsw_150**

Table 8.9: Statistical significance of the best retrieval scores (MASDWP) for fixed length based runs (len_400, len_nsw_150), statistical significance in bold.

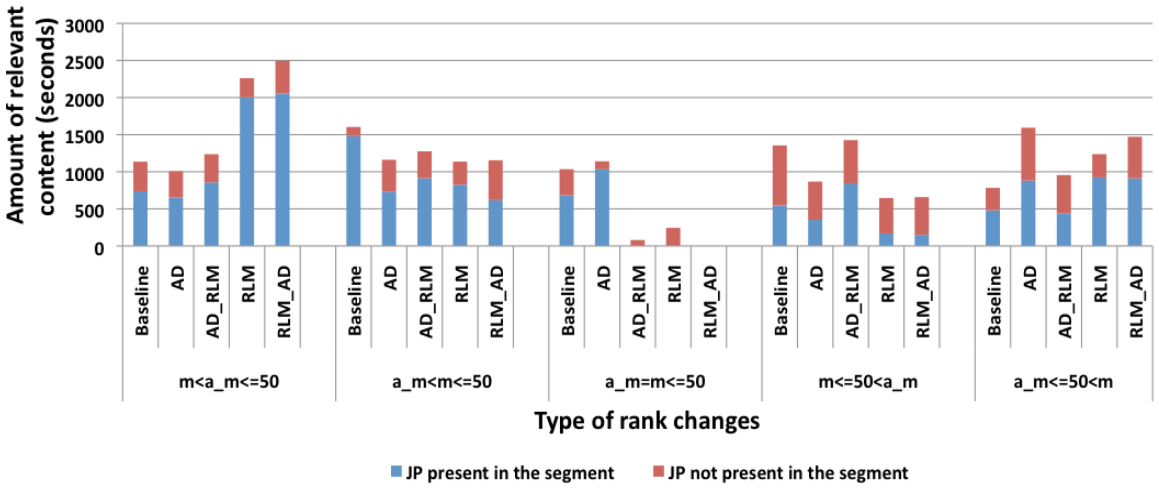


Figure 8.23: Amount of relevant content in the top 50 ranks in asr_man and manual runs based on C99 segmentation, query 21.

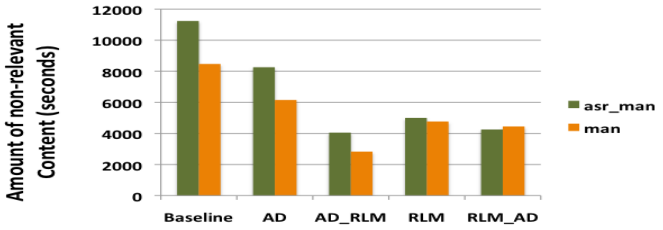


Figure 8.24: Amount of non-relevant content in the top 50 ranks in asr_man and manual runs based on C99 segmentation, query 21.

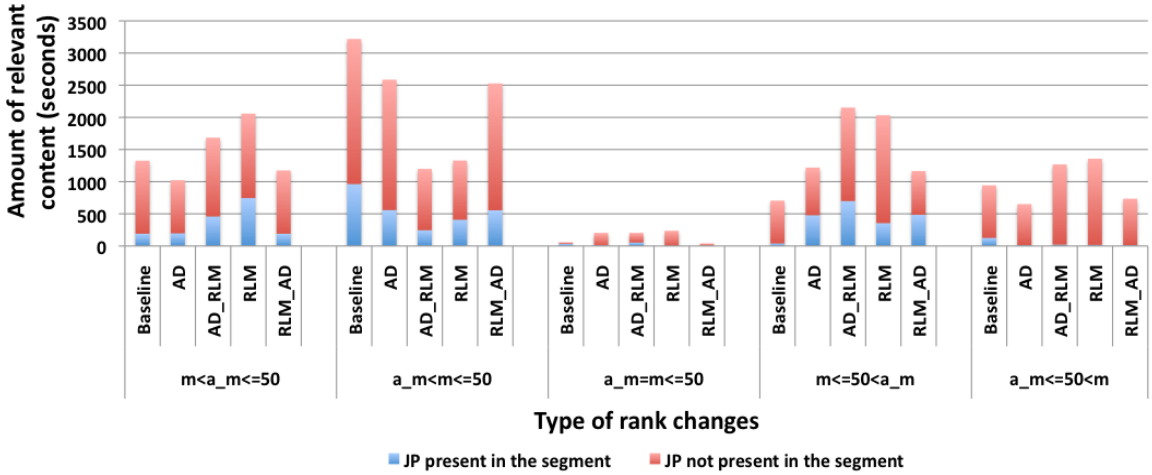


Figure 8.25: Amount of relevant content in the top 50 ranks in asr_man and manual runs based on C99 segmentation, query 13.

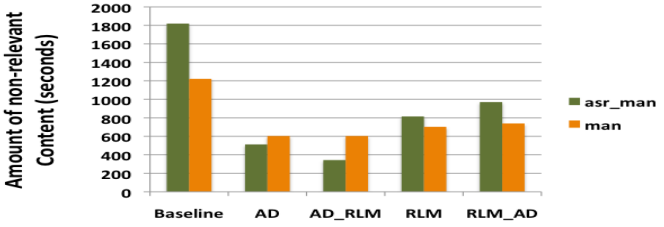


Figure 8.26: Amount of non-relevant content in the top 50 ranks in asr_man and manual runs based on C99 segmentation, query 13.

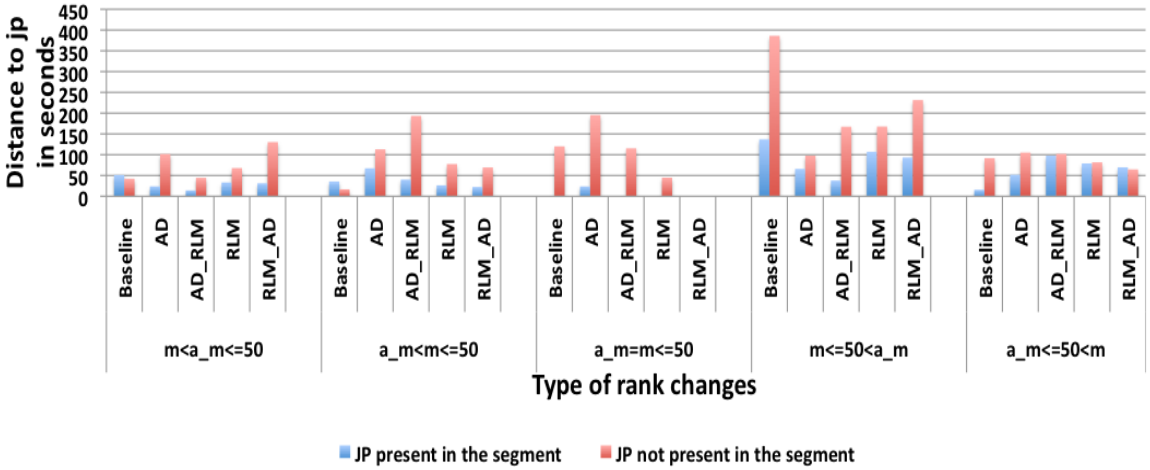


Figure 8.27: Average distance to the jump-in points in the top 50 ranks in asr_man and manual runs based on C99 segmentation, query 21.

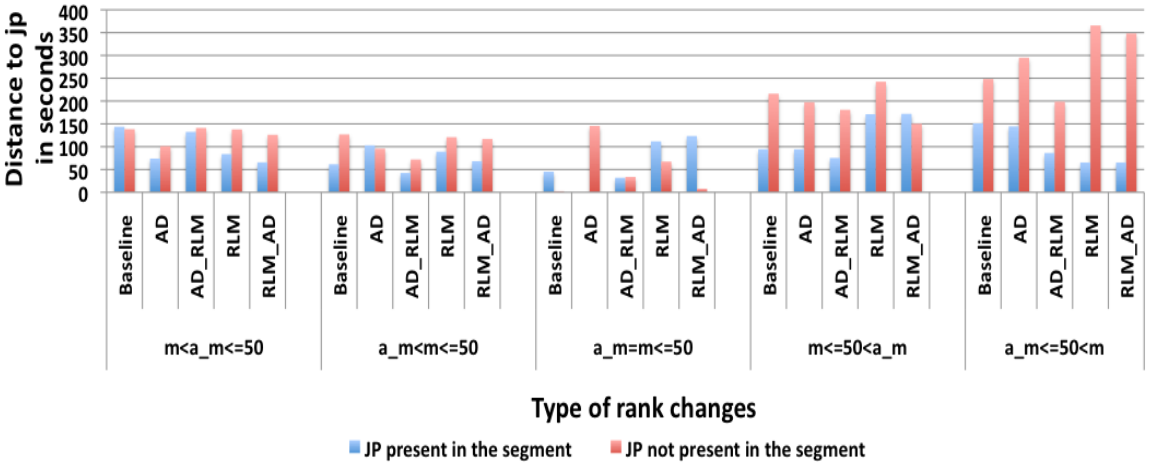


Figure 8.28: Average distance to the jump-in points in the top 50 ranks in asr_man and manual runs based on C99 segmentation, query 13.

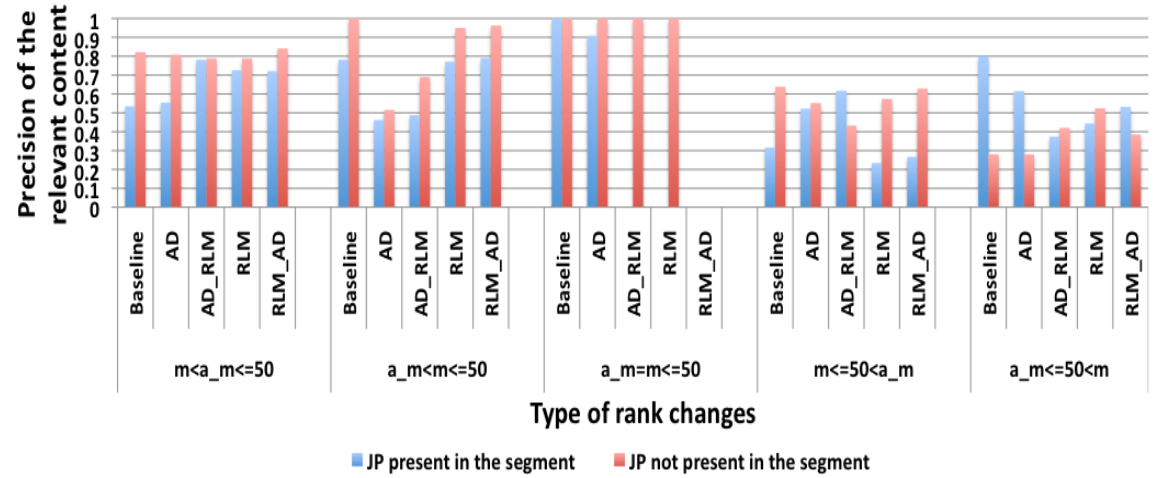


Figure 8.29: Average precision of the relevant content in the segment in the top 50 ranks in asr_man and manual runs based on C99 segmentation, query 21.

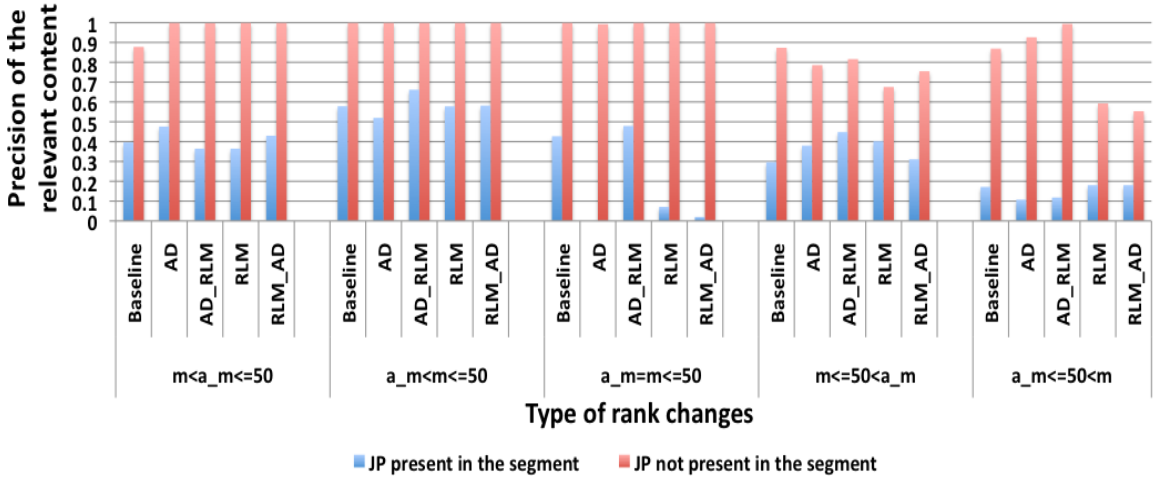


Figure 8.30: Average precision of the relevant content in the segment in the top 50 ranks in asr_man and manual runs based on C99 segmentation, query 13.

Transcript type: asr_man				
Expansion method	Segmentation type			
	c99	tt	len_400	len_nsw_150
BASELINE	0.397	0.376	0.435	0.407
AD	0.414	0.388	0.450	0.418
AD_RLM	0.436	0.408	0.468	0.437
RLM	0.445	0.407	0.466	0.434
RLM_AD	0.456	0.431	0.477	0.447

Transcript type: man				
Expansion method	Segmentation type			
	c99	tt	len_400	len_nsw_150
BASELINE	0.447	0.437	0.466	0.453
AD	0.466	0.438	0.472	0.450
AD_RLM	0.470	0.448	0.478	0.468
RLM	0.472	0.467	0.493	0.477
RLM_AD	0.485	0.475	0.503	0.490

Table 8.10: MAP scores for varying segmentation types and document expansion methods applied, $\lambda = 0.85$

Transcript type: asr_man				
Expansion method	Segmentation type			
	c99	tt	len_400	len_nsw_150
BASELINE	0.199	0.200	0.227	0.215
AD	0.222	0.213	0.230	0.223
AD_RLM	0.247	0.235	0.241	0.236
RLM	0.242	0.229	0.243	0.235
RLM_AD	0.250	0.249	0.250	0.243

Transcript type: man				
Expansion method	Segmentation type			
	c99	tt	len_400	len_nsw_150
BASELINE	0.241	0.243	0.242	0.241
AD	0.263	0.252	0.244	0.241
AD_RLM	0.274	0.270	0.251	0.255
RLM	0.266	0.275	0.258	0.256
RLM_AD	0.277	0.283	0.265	0.266

Table 8.11: mGAPscores for varying segmentation types and document expansion methods applied, $\lambda = 0.85$

Transcript type: asr_man				
Expansion method	Segmentation type			
	c99	tt	len_400	len_nsw_150
BASELINE	0.164	0.181	0.218	0.213
AD	0.186	0.192	0.225	0.213
AD_RLM	0.217	0.217	0.239	0.227
RLM	0.208	0.213	0.237	0.232
RLM_AD	0.216	0.230	0.243	0.239
Transcript type: man				
Expansion method	Segmentation type			
	c99	tt	len_400	len_nsw_150
BASELINE	0.192	0.206	0.233	0.237
AD	0.210	0.215	0.233	0.235
AD_RLM	0.237	0.234	0.239	0.249
RLM	0.229	0.251	0.253	0.258
RLM_AD	0.239	0.257	0.259	0.265

Table 8.12: MASP scores for varying segmentation types and document expansion methods applied, $\lambda = 0.85$

Transcript type: asr_man				
Expansion method	Segmentation type			
	c99	tt	len_400	len_nsw_150
BASELINE	0.095	0.103	0.114	0.119
AD	0.114	0.112	0.117	0.120
AD_RLM	0.138	0.132	0.126	0.129
RLM	0.127	0.127	0.125	0.133
RLM_AD	0.133	0.140	0.129	0.138
Transcript type: man				
Expansion method	Segmentation type			
	c99	tt	len_400	len_nsw_150
BASELINE	0.116	0.120	0.123	0.132
AD	0.133	0.130	0.122	0.1330
AD_RLM	0.154	0.146	0.127	0.142
RLM	0.145	0.155	0.135	0.146
RLM_AD	0.153	0.160	0.139	0.151

Table 8.13: MASDWP scores for varying segmentation types and document expansion methods applied, $\lambda = 0.85$

Transcript type: asr_man				
Expansion method	Segmentation type			
	c99	tt	len_400	len_nsw_150
BASELINE	0.504	0.496	0.465	0.493
AD	0.499	0.500	0.463	0.487
AD_RLM	0.495	0.505	0.462	0.486
RLM	0.504	0.504	0.463	0.494
RLM_AD	0.496	0.506	0.461	0.490

Transcript type: man				
Expansion method	Segmentation type			
	c99	tt	len_400	len_nsw_150
BASELINE	0.509	0.505	0.465	0.493
AD	0.504	0.501	0.465	0.488
AD_RLM	0.503	0.505	0.462	0.487
RLM	0.504	0.505	0.465	0.491
RLM_AD	0.502	0.504	0.463	0.488

Table 8.14: Average Precision for varying segmentation types and document expansion methods applied, $\lambda = 0.85$

QUERY 21: Transcript type: asr_man					
	MAP	mGAP	MASP	MASDW	AVRPrecision
BASELINE	0.2252	0.1536	0.1691	0.112	0.5446
AD	0.2419	0.1696	0.1573	0.1083	0.5487
AD_RLM	0.2855	0.209	0.2304	0.1668	0.5342
RLM	0.2818	0.1986	0.2319	0.1586	0.5417
RLM_AD	0.2857	0.2041	0.2425	0.1679	0.5358
QUERY 21: Transcript type: man					
	MAP	mGAP	MASP	MASDW	AVRPrecision
BASELINE	0.2802	0.1873	0.2093	0.1359	0.5557
AD	0.3135	0.2226	0.2228	0.1596	0.5280
AD_RLM	0.3314	0.2393	0.2691	0.1951	0.5445
RLM	0.3335	0.2322	0.2795	0.1909	0.5494
RLM_AD	0.3451	0.2446	0.2940	0.2048	0.5308
QUERY 13: Transcript type: asr_man					
	MAP	mGAP	MASP	MASDW	AVRPrecision
BASELINE	0.5418	0.4362	0.4082	0.3240	0.7788
AD	0.7372	0.5965	0.5609	0.4502	0.7643
AD_RLM	0.7843	0.6398	0.6415	0.5190	0.7475
RLM	0.6570	0.5298	0.5332	0.4266	0.7793
RLM_AD	0.6981	0.5664	0.5609	0.4508	0.7731
QUERY 13: Transcript type: man					
	MAP	mGAP	MASP	MASDW	AVRPrecision
BASELINE	0.6111	0.4876	0.4731	0.3739	0.7703
AD	0.7984	0.6493	0.6020	0.4853	0.7713
AD_RLM	0.8078	0.6613	0.6295	0.5118	0.7544
RLM	0.7196	0.5862	0.5624	0.4548	0.7484
RLM_AD	0.7491	0.6122	0.5832	0.4724	0.7452

Table 8.15: MAP, mGAP, MASP, MASDWP metric scores and average precision of the relevant content for one example query (queries 21 and 13) using Baseline and all document expansion methods on the collection with C99 segmentation.

Chapter 9

Conclusions and Future Work

This concluding chapter summarizes the contributions of this thesis to the state-of-the-art SCR research field. We provide answers to the research questions introduced in Chapter 1, and outline potential direction for future work that can be based on these findings.

9.1 Summary of the thesis contributions

In this thesis we overviewed SCR experiments across varying types of conversational content: lectures, meeting, and Internet TV. We addressed such problems of dataset creation as choice of potential real case scenario, and use of crowdsourcing technologies for query set creation and relevance assessment; we analyzed and extended evaluation strategies that enable us to reflect various aspects of user experience, in Chapters 4 and 5 respectively. We showed that the power of the crowd proves to be useful for collecting the queries for challenging video collection in natural language, as well as for getting the feedback from the users about their general interest in a given SCR task at the same time. Even though this approach requires an extensive manual analysis on the part of any researcher setting up a task, we are rewarded with a good coverage of collection on a big scale within our final query/relevance set.

Our detailed analysis of the impact of the ASR errors and segmentation variation on the SCR results, presented in Chapter 6, highlighted that these do not only decrease the ranking, but also promote relevant content, as non-relevant content is affected. Through all the data types precision and recall of the relevant content within the segmentation unit plays an essential role in its efficient retrieval, no matter what weighting scheme is used as the IR model, as it was shown in Chapters 6-8.

In Chapter 7, we investigated segmentation approaches that can potentially address precision and recall issues, and we demonstrated how boundary adjustement using the lexical-cohesion based segmentation when imposed over fixed length segmentation methods improves the results. These findings extend the discussion about evaluation framework because traditional IR metrics do not allow cross comparison between all the runs created using this approach due to segment length differences in the runs and their relevance representation.

New document expansion methods that use a combination of immediate document context and overall collection robustly improve SCR effectiveness, as they target both imperfect content segmentation and make use of collection knowledge, see Chapter 8.

9.2 Answers to research questions and further discussion

In this section we return to the research questions introduced in Chapter 1 to discuss the answers that this thesis has reported, and outline potential follow-up research goals that these helped to define.

- *Research Question 1 (RQ1): What is the relationship between ASR errors in the transcript and retrieval behaviour?*

  Our experiments described in Chapter 6 have confirmed previous studies that

argued that better ASR quality of the relevant segments influence the retrieval effectiveness of those. However, our analysis extended this knowledge to the understanding that lower ASR quality has impact on the non-relevant content that moves down the retrieved rank lists and enables improvement of segments that had lower retrieval ranks otherwise.

On the other hand, spoken content that is created in situations when conversation participants share common knowledge about the topic of the discussion, may be hard to retrieve even in cases of high ASR quality. Since this common knowledge may be present in the other modalities (slides or any other additional information shown on a screen or printed and thus accessible to conversation participants in the room) or it may become a tacit information that was discussed at previous meetings, lectures or any other recordings, in these cases effective SCR can benefit from the use of external sources of data, e.g. metadata as discussed in Chapter 6, or immediate or general collection context as described in Chapter 8.

- *Research Question 2 (RQ2): Do current evaluation techniques sufficiently reflect all the most important aspects of the user experience in SCR?*

  Our experiments have shown that even though MAP evaluation metric is useful in reflecting the general ranking of the content, it does not correspond to the amount of effort on the part of the user that browsing through the audio/video content implies. Our introduced metrics MASP and MASDWP take into account the amount of relevant and non-relevant content that the user is exposed to, and use information about jump-in points. These metrics require time information about the segments and are not associated with content segmentation into sentences or utterances, which enables their use for the experiments with the data that has time stamps information and lacks punctuation information, and the relevance assessment is stored in terms of time.

- *Research Question 3a (RQ3a): How does segmentation of spoken data affect retrieval behaviour? Research Question 3b (RQ3b): What are the characteristics of a segmentation method that maximizes SCR effectiveness?*

  As discussed in Chapter 6-8, high precision and recall of the relevant content within the segment and closeness of the start time to the actual jump-in point enable their retrieval at high ranks. The segmentation into units on the same topic is preferable, as it might recover the fact that the actual relevant segments do not have many terms overlapping with the query. Splitting of the relevant content across several segments decreases their ranking.

  Unified segmentation of the collection into units that are used for all queries does not allow flexiblity of providing the user with segments of different size depending on the specificity of their queries. Therefore we invesigated different approaches of filtering for the runs with overlapping segments (removal or combination of these segments), see Chapter 7 for details.

- *Research Question 4 (RQ4): How can regions of different speech recognition quality be identified and processed in order to improve overall speech retrieval performance (detection, special treatment in the speech retrieval process)? Otherwise, how can the regions of good recognition be used most efficiently in retrieval process?*

  In this thesis we focused on the analysis of the behaviour of the speech segments with varying quality, and use of the regions that potentially contain good recognition results.

- *Research Question 5 (RQ5): Can we implement a meaningful approach to SCR of conversational content incorporating task specific segmentation?*

  Use of the context documents has proven to improve results for different types of segmentation: immediate context can resolve potential segmentation disfluencies, and use of top ranked segments enables document expansion based on

terms from the documents that might have better ASR scores. Our experiments have shown that fixed length segmentation results can be improved by implementing further bounday adjustment using lexical cohesion information, see Chapter 7 for details. Opening the discussion about use of audio signal features for boundary adjustment, we started with low level aspects as words loudness and pauses between them. On different types of data, they proved to improve the SCR effectiveness when used before indexing (addition of boundaries in the collection) or after the retrieval stage (boundary adjustment for general segments from the collection). These initial experiments suggest further work using these features across datasets, and gradual incorporation of more complicated prosodic features that reflect the structure of the content and may help to better understand importance of terms for retrieval.

9.3 Proposed future directions for SCR research

A number of possible future research directions can be introduced based on the work and analysis of SCR experiments framework presented in this thesis.

These include futher development of filtering, boundary adjustment, and document expansion techniques using deeper knowledge of the spoken content transcript, i.e. going beyond 1-best transcript or making use of the confidence score information in the weighting scheme of the documents.

Considering that spoken data may contain relevant content for some queries that varies not only in recording quality and subsequently ASR transcript quality, but also in terms of the amount of information present and tacit knowledge between conversation participants, while not being explicitly spoken, a complex approach to query processing that assumes these different scenarios need to be investigated.

In general, all the assumptions about use case scenarios introduced and implied throughout the experiments described in this thesis represent a separate field for potential analysis and exploration. The user behaviour and interaction with a SCR

system might differ depending on the type of device that is being used, on the nature of the task, and the level of user engagement. These aspects can be further incorporated into the evaluation strategy of the SCR results.

9.3.1 Use of acoustic features and transcript quality

As our SCR performance analysis across different tasks has shown, ASR transcript quality influences retrieval in a complex way as its impact is tied to the segmentation quality and closeness to the jump-in points of the segments to the relevant content. Therefore potential ways to improve SCR effectiveness lie in further exploitation of the acoustic features and statistical details of the transcript itself.

Boundary adjustment

In a real-life scenario users that issue an information request may already have different knowledge about the content, thus they may require broader or narrower scope of topic coverage, i.e. different jump-in points within the same files for each information request. In Chapter 7 we examined filtering of the overlapping context that finds segments of different length and start time within the same collection for each individual query. We also opened the discussion about the use of knowledge about acoustic features of the segments to adjust the boundaries of already retrieved segments. We started with low level potential jump-in points such as pauses and the loudest words in each segment. While at this stage pauses produce better results, we believe that use of prominent words, or maybe utterances that contain them, have more potential to improve SCR effectiveness. The next step is to pay attention to the prosodic structure of the whole utterances present in the segments and what speaker intention these utterances might represent, thus adjusting the boundaries accordingly if this type of intent is assumed in a query.

Selection of candidate terms for document expansion or reduction

In the SCR experiments reported in this thesis, we did not make any distinction between the words in the transcript, apart from filtering out the common words. However, as the ASR system output contains different confidence levels that these are the correct representation of what was pronounced, in further work we could take this information into account. Low confidence scores may suggest that these words may need to be filtered out before indexing stage as unreliable and potentially errorful to reduce the documents to their reliable content only.

On the other hand, as the active vocabulary of the ASR system may differ from the topic vocabulary of the audio, the words with low confidence score are not to be discarded without further analysis. The ASR system weight can be taken into account by the SCR system while indexing the documents, and especially when choosing the terms for document expansion.

Knowledge about confidence of the ASR system in the words can be combined with their prominence level and roles in the prosodic structure of the sentences to define better candidates for document expansion or deletion of terms.

9.3.2 Retrieval of relevant content with tacit information

Within the framework of spoken content recording, relevant information can be expressed in terms that can be incorrectly recognized by the ASR system. However, 100% correctness of the ASR transcript does not guarantee presence of the pronounced relevant terms as they may be unknown to the vocabulary of ASR. These types of issues are addressed by use of metadata that might be assigned to the audio files or videos. In real case scenarios the absence of the query terms from the documents may be overcome by use of external information provided by the metadata or other knowledge source, e.g. use of text from the slides shown at the time of a lecture or meeting.

In other cases, the relevant terms might be used earlier in the recording, but

further on referred to using pronouns. Document expansion that targets solving pronoun resolution might be an additional approach to improve retrieval effectiveness of these segments, or might be combined with current methods as we already use the immediate document context.

Overall, one query might require the retrieval of segments with the relevant information being explicitly expressed in the content, or implied by the speakers. As these two types of segment require different approaches to achieve the best retrieval performance, we assume that a combination of the retrieved lists might represent a solution that covers all types of relevant content.

9.4 Concluding remarks

The work completed in this thesis targeted deep analysis of SCR. We analyzed all stages of the retrieval process, from collection creation using state-of-the-art crowdsourcing platforms and traditional manual choice of data for specific tasks such as meeting search, segmentation of the content and its retrieval, to evaluation methods and techniques that achieve significantly better results. We implemented different segmentation approaches and tracked their influence on system behaviour, we opened potential new research directions of use of acoustic information to address specific SCR challenges. We hope that this work will inspire further investigations in SCR, and will provide valuable inspiration for real-world applications.

Bibliography

Adcock, J., Cooper, M., Denoue, L., Pirsiavash, H., and Rowe, L. A. (2010). TalkMiner: a lecture webcast search engine. In *Proceedings of the International Conference on Multimedia*, ACM MM 2010, pages 241–250, Florence, Italy.

Akiba, T., Aikawa, K., Itoh, Y., Kawahara, T., Nanjo, H., Nishizaki, H., Yasuda, N., Yamashita, Y., and Itou, K. (2008). Test collections for spoken document retrieval from lecture audio data. In *Proceedings of the 6th Language Resources and Evaluation Conference (LREC)*, Marrakech, Morocco.

Akiba, T., Aikawa, K., Itoh, Y., Kawahara, T., Nanjo, H., Nishizaki, H., Yasuda, N., Yamashita, Y., and Itou, K. (2009). Construction of a test collection for spoken document retrieval from lecture audio data. *Journal of Information Processing*, 17:82–94.

Akiba, T., Nishizaki, H., Aikawa, K., Hu, X., Itoh, Y., Kawahara, T., Nakagawa, S., Nanjo, H., and Yamashita, Y. (2013). Overview of the NTCIR-10 SpokenDoc-2 Task. In *NTCIR-10 Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access*, pages 573–587, Tokyo, Japan.

Akiba, T., Nishizaki, H., Aikawa, K., Kawahara, T., and Matsui, T. (2011). Overview of the IR for spoken documents task in NTCIR-9 workshop. In *NTCIR-9 Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access*, pages 223–235, Tokyo, Japan.

Allan, J. (2003). Robust techniques for organizing and retrieving spoken documents. *EURASIP J. Appl. Signal Process.*, 2003:103–114.

Aly, R., Verschoor, T., and Ordelman, R. (2011). UTwente does Rich Speech Retrieval at MediaEval 2011. In *Working Notes Proceedings of the MediaEval 2011 Workshop*, volume 807 of *CEUR Workshop Proceedings*, Pisa, Italy.

Baeza-Yates, R. A. and Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

Billerbeck, B. and Zobel, J. (2005). Document expansion versus query expansion for ad-hoc retrieval. In *Proceedings of the Tenth Australasian Document Computing Symposium*, pages 34–41, Sydney, Australia.

Brown, E., Srinivasan, S., Coden, A., Ponceleon, D., Cooper, J., and Amir, A. (2001). Toward speech as a knowledge resource. *IBM Systems Journal*, 40(4):985–1001.

Brown, M. G., Foote, J. T., Jones, G. J. F., Spärck Jones, K., and Young, S. J. (1994). Video mail retrieval by voice: An overview of the Cambridge/Olivetti retrieval system.

Brown, M. G., Foote, J. T., Jones, G. J. F., Spärck Jones, K., and Young, S. J. (1996). Open-vocabulary speech indexing for voice and video mail retrieval. In *Proceedings of the fourth ACM international conference on Multimedia (MULTIMEDIA '96)*, pages 307–316, Boston, Massachusetts, USA.

Büttcher, S., Clarke, C. L. A., and Cormack, G. V. (2010). *Information Retrieval: Implementing and Evaluating Search Engines*. MIT Press.

Byrne, W., Doermann, D., Franz, M., Gustman, S., Hajic, J., Oard, D., Picheny, M., Psutka, J., Ramabhadran, B., Soergel, D., Ward, T., and Zhu, W.-J. (2004). Automatic recognition of spontaneous speech for access to multilingual oral history archives. *IEEE Transactions on Speech and Audio Processing*, 12(4):420–435.

Carletta, J. (2007). Unleashing the killer corpus: experiences in creating the multi-everything AMI meeting corpus. *Language Resources and Evaluation Journal*, 41(2):181–190.

Chelba, C., Hazen, T. J., and Saralar, M. (2008). Retrieval and browsing of spoken content. *IEEE Signal Processing Mag*, pages 39–49.

Chibelushi, C. and Thelwall, M. (2009). Text mining for meeting transcript analysis to extract key decision elements. In *Proceedings of the International MultiConference of Engineers and Computer Scientists (IMECS 2009)*, pages 710–715, Hong Kong, China.

Cho, E., Fügen, C., Hermann, T., Kilgour, K., Mediani, M., Mohr, C., Niehues, J., Rottmann, K., Saam, C., Stüker, S., and Waibel, A. (2013). A real-world system for simultaneous translation of german lectures. In *14th Annual Conference of the International Speech Communication Association (INTERSPEECH 2013)*, pages 3473–3477, Lyon, France.

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, Seattle, Washington, USA.

Cieri, C., Graff, D., Liberman, M., Martey, N., and Strassel, S. (1999). The tdt-2 text and speech corpus. In *IN PROCEEDINGS OF DARPA BROADCAST NEWS WORKSHOP*, pages 57–60. Morgan Kaufmann.

Cormack, G. V., Palmer, C. R., and Clarke, C. L. A. (1998). Efficient construction of large test collections. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR'98)*, pages 282–289, Melbourne, Australia.

Denoue, L., Hilbert, D., Adcock, J., Billsus, D., and Cooper, M. (2005). Projectorbox: Seamless presentation capture for classrooms. In Richards, G., edi-

tor, *Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 2005*, pages 1986–1991.

Dharanipragada, S., Franz, M., and Roukos, S. (1998). Audio-indexing for broadcast news. In *Proceedings of the 8th Text Retrieval Conference (TREC-7)*, pages 63–67.

Dielmann, A. and Renals, S. (2007). Automatic Meeting Segmentation Using Dynamic Bayesian Networks. *Multimedia, IEEE Transactions on*, 9(1):25–36.

Efron, M., Organisciak, P., and Fenlon, K. (2012). Improving retrieval of short texts through document expansion. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval (SIGIR '12)*, pages 911–920, Portland, Oregon, USA.

Eskevich, M. and Jones, G. J. (2011a). DCU at the NTCIR-9 SpokenDoc Passage Retrieval Task. In *NTCIR-9 Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access*, pages 257–260, Tokyo, Japan.

Eskevich, M. and Jones, G. J. F. (2011b). Dcu at mediaeval 2011: Rich speech retrieval. In *Working Notes Proceedings of the MediaEval 2011 Workshop*, volume 807 of *CEUR Workshop Proceedings*, Pisa, Italy.

Eskevich, M. and Jones, G. J. F. (2012). DCU Search Runs at MediaEval 2012: Search and Hyperlinking Task. In *MediaEval*, volume 927 of *CEUR Workshop Proceedings*, Pisa, Italy. CEUR-WS.org.

Eskevich, M. and Jones, G. J. F. (2013). Time-based Segmentation and Use of Jump-in Points in DCU Search Runs at the Search and Hyperlinking Task at MediaEval 2013. In *Proceedings of MediaEval 2013*, Barcelona, Spain.

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*, volume 927 of *CEUR Workshop Proceedings*, Pisa, Italy. CEUR-WS.org.

Eskevich, M., Jones, G. J. F., Larson, M., and Ordelman, R. (2012b). Creating a data collection for evaluating rich speech retrieval. In *Proceedings of the 8th Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey.

Eskevich, M., Jones, G. J. F., Wartena, C., Larson, M., Aly, R., Verschoor, T., and Roeland (2012c). Comparing retrieval effectiveness of alternative content segmentation methods for internet video search. In *10th International Workshop on Content-Based Multimedia Indexing (CBMI 2012)*, pages 1–6, Annecy, France.

Eyben, F., Wöllmer, M., and Schuller, B. (2010). Opensmile: The munich versatile and fast open-source audio feature extractor. In *Proceedings of the International Conference on Multimedia (MM '10)*, pages 1459–1462, Firenze, Italy.

Fiscus, J. G., Ajot, J., and Garofolo, J. S. (2007). The rich transcription 2007 meeting recognition evaluation. In *Proceedings of Multimodal Technologies for Perception of Humans, International Evaluation Workshops CLEAR 2007 and RT 2007*, pages 373–389, Baltimore, MD, USA.

Foote, J. T., Jones, G. J. F., Spärck Jones, K., and Young, S. J. (1995). Talker-independent keyword spotting for information retrieval. In *Fourth European Conference on Speech Communication and Technology (EUROSPEECH 1995)*, Madrid, Spain.

Galley, M., McKeown, K. R., Fosler-Lussier, E., and Jing, H. (2003). Discourse segmentation of multi-party conversation. In *Proceedings of the 41st annual meeting of the Association for Computational Linguistics (ACL–03)*, pages 562–569, Sapporo, Japan.

Garofolo, J. S., Auzanne, C. G. P., and Voorhees, E. M. (2000a). The TREC spoken document retrieval track: A success story. In *Proceedings of RIAO 2000*, pages 1–20, Paris, France.

Garofolo, J. S., Lard, J., Auzanne, C. G. P., and Voorhees, E. M. (2000b). 2000 TREC-9 spoken document retrieval (SDR) track evaluation specification. NIST.

Garofolo, J. S., Voorhees, E. M., Auzanne, C. G. P., and Stanford, V. M. (1999). Spoken document retrieval: 1998 evaluation and investigation of new metrics. In *Proceedings of the ESCA workshop: Accessing information in spoken audio*, pages 1–7, Cambridge, United Kingdom.

Garofolo, J. S., Voorhees, E. M., Auzanne, C. G. P., Stanford, V. M., and Lund, B. A. (1998). TREC-7 spoken document retrieval track overview and results. In *Proceedings of the 7th Text Retrieval Conference (TREC-7)*. NIST.

Garofolo, J. S., Voorhees, E. M., Stanford, V. M., and Spärck Jones, K. (1997). TREC-6 1997 spoken document retrieval track overview and results. In *Proceedings of the 7th Text Retrieval Conference (TREC-6)*, pages 83–91. NIST.

Gauvain, J.-L., Lamel, L., and Adda, G. (2002). The LIMSI broadcast news transcription system. *Speech Communication*, 37(1-2):89–108.

Gauvain, J.-L., Lamel, L., Barras, C., Adda, G., and de Kercadio, Y. (2000). The LIMSI SDR System for TREC-9. In *Proceedings of the 9th Text Retrieval Conference TREC-9*. NIST.

Glass, J., Hazen, T. J., Hetherington, L., and Wang, C. (2004). Analysis and processing of lecture audio data: preliminary investigations. In *Proceedings of the Workshop on Interdisciplinary Approaches to Speech Indexing and Retrieval at HLT-NAACL 2004*, SpeechIR '04, pages 9–12, Boston, Massachusetts, USA.

Glass, J. R., Hazen, T. J., Cyphers, D. S., Malioutov, I., Huynh, D., and Barzilay, R. (2007). Recent progress in the MIT spoken lecture processing project. In *8th Annual Conference of the International Speech Communication Association (INTERSPEECH 2007)*, pages 2553–2556, Antwerp, Belgium. ISCA.

Goldman, J., Renals, S., Bird, S., de Jong, F., Federico, M., Fleischhauer, C., Kornbluh, M., Lamel, L., Oard, D. W., Stewart, C., and Wright, R. (2005). Accessing the spoken word. *Int. J. on Digital Libraries*, 5(4):287–298.

Hain, T., Burget, L., Dines, J., Garner, P. N., Grezl, F., Hannani, A. E., Huijbregts, M., Karafiat, M., Lincoln, M., and Wan, V. (2012). Transcribing meetings with the amida systems. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(2):486–498.

Hearst, M. (1993). TextTiling: A quantitative approach to discourse segmentation. Technical Report Sequoia 93/24, Computer Science Department, University of California, Berkeley, USA.

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

Hiemstra, D. (2001). *Using language models for information retrieval*. PhD thesis, University of Twente, The Netherlands.

Hsueh, P.-Y. (2008). Audio-based unsupervised segmentation of multiparty dialogue. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2008)*, pages 5049–5052, Las Vegas, Nevada, USA.

Hsueh, P.-Y. and Moore, J. D. (2006). Automatic topic segmentation and labeling in multiparty dialogue. In *Proceedings of the first IEEE/ACM workshop on Spoken Language Technology (SLT)*, Aruba.

Huang, X., Acero, A., and Hon, H.-W. (2001). *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development (1st edition)*. Prentice Hall PTR.

James, D. A. (1995). *The Application of Classical Information Retrieval Techniques to Spoken Document*. PhD thesis, University of Cambridge, UK.

James, D. A. (1996). A system for unrestricted topic retrieval from radio news broadcasts. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '96)*, pages 279–282, Atlanta, Georgia, USA.

Jelinek, F. (1976). Continuous speech recognition by statistical methods. *Proceedings of the IEEE*, 64(4):532–556.

Jiang, H. (2005). Confidence measures for speech recognition: A survey. *Speech Communication*, 45(4):455–470.

Johnson, S. E., Jourlin, P., Moore, G. L., Jones, K. S., and Woodland, P. C. (1998). Spoken Document Retrieval For TREC-7 at Cambridge University. In *Proceedings of the 8th Text Retrieval Conference (TREC-7)*, pages 138–147.

Jones, G. J. F. and Edens, R. J. (2002). Automated alignment and annotation of audio-visual presentations. In *Proceedings of the Sixth European Conference in Research and Advanced Technology for Digital Libraries (ECDL 2002)*, pages 276–291, Rome, Italy.

Jones, G. J. F., Foote, J. T., Spärck Jones, K., and Young, S. J. (1996). Retrieving spoken documents by combining multiple index sources. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR'96)*, pages 30–38, Zurich, Switzerland.

Jones, G. J. F., Zhang, K., Newman, E., and Adenike (2007). Examining the Contributions of Automatic Speech Transcriptions and Metadata Sources for Searching Spontaneous Conversational Speech. In *Proceedings of the Searching Spontaneous Conversational Speech (SSCS) workshop SIGIR*, Amsterdam, The Netherlands.

Juang, B.-H. and Furui, S. (2000). Automatic recognition and understanding of spoken language - a first step toward natural human-machine communication. *Proceedings of The IEEE*, 88:1142–1165.

Jurafsky, D. and Martin, J. (2000). *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition.* Prentice Hall.

Kamps, J., Pehcevski, J., Kazai, G., Lalmas, M., and Robertson, S. (2008). INEX 2007 evaluation measures. In *Focused Access to XML Documents*, pages 24–33. Springer.

Kaneko, T., Takigami, T., and Akiba, T. (2011). STD based on hough transform and SDR using STD results: Experiments at NTCIR-9 SpokenDoc. In *NTCIR-9 Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access*, pages 264–270, Tokyo, Japan.

Kekalainen, J. and Jarvelin, K. (2002). Using graded relevance assessments in IR evaluation. *Journal of the American Society for Information Science and Technology*, 53:1120–1129.

Kilgour, J., Carletta, J., and Renals, S. (2010). The Ambient Spotlight: Queryless desktop search from meeting speech. In *Proceedings of the 2010 international workshop on Searching Spontaneous Conversational Speech (SSCS) at ACM MM*, Florence, Italy.

Kolár, J. and Lamel, L. (2011). On development of consistently punctuated speech corpora. In *Proceedings of the 12th Annual Conference of the International Speech Communication Association (INTERSPEECH 2011)*, pages 833–836, Florence, Italy.

Kolár, J. and Lamel, L. (2012). Development and Evaluation of Automatic Punctuation for French and English Speech-to-Text. In *Proceedings of the 13th Annual Conference of the International Speech Communication Association (INTERSPEECH 2012)*, Portland, Oregon, USA.

Koumpis, K. and Renals, S. (2005). Content-based access to spoken audio. *IEEE Signal Processing Magazine*, 22(5):61–69.

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

Lanchantin, P., Bell, P., Gales, M. J. F., Hain, T., Liu, X., Long, Y., Quinnell, J., Renals, S., Saz, O., Seigel, M. S., Swietojanski, P., and Woodland, P. C. (2013). Automatic transcription of multi-genre media archives. In *Proceedings of the First Workshop on Speech, Language and Audio in Multimedia (SLAM@INTERSPEECH)*, volume 1012 of *CEUR Workshop Proceedings*, pages 26–31, Marseille, France. CEUR-WS.org.

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 *Proceedings of MediaEval 2011*, volume 807 of *CEUR Workshop Proceedings*, Pisa, Italy.

Larson, M., Soleymani, M., Eskevich, M., Serdyukov, P., Ordelman, R., and Jones, G. J. F. (2012). The community and the crowd: Multimedia benchmark dataset development. *IEEE MultiMedia*, 19(3):15.

Larson, M., Tsagkias, M., He, J., and De Rijke, M. (2009). Investigating the global semantic impact of speech recognition error on spoken content collections. In *Proceedings of the 31st European Conference on Information Retrieval (ECIR 2009)*, pages 755–760, Toulouse, France.

Lavrenko, V. and Croft, W. B. (2001). Relevance based language models. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '01)*, pages 120–127, New Orleans, Louisiana, USA.

Lee, A. and Kawahara, T. (2009). Recent development of open-source speech recognition engine julius. In *Proceedings of the 1st Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC2009)*, Sapporo, Japan.

Lee, D. and Lee, G. G. (2008). A Korean spoken document retrieval system for

lecture search. In *ACM SIGIR Workshop Searching Spontaneous Conversational Speech (SSCS)*, pages 73–74, Singapore.

Lee, L.-s. and Chen, B. (2005). Spoken document understanding and organization. *Signal Processing Magazine, IEEE*, 22(5):42–60.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Cybernetics and Control Theory*, 10(8):707–710.

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 (ACM SIGIR 2006)*, pages 673–674, Seattle, USA.

Lovins, J. B. (1968). Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics*, 11:22–31.

Luz, S. and Su, J. (2010). The relevance of timing, pauses and overlaps in dialogues: Detecting topic changes in scenario based meetings. In *Proceedings of INTERSPEECH 2010*, pages 1369–1372, Makuhari, Japan. ISCA.

Maekawa, K., Koiso, H., Furui, S., and Isahara, H. (2000). Spontaneous speech corpus of japanese. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC)*, Athens, Greece.

Malioutov, I. (2006). Minimum cut model for spoken lecture segmentation. Master's thesis, Massachusetts Institute of Technology, USA.

Malioutov, I. and Barzilay, R. (2006). Minimum cut model for spoken lecture segmentation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL 2006)*, pages 25–32, Sydney, Australia.

Mangu, L., Brill, E., and Stolcke, A. (2000). Finding consensus in speech recognition: word error minimization and other applications of confusion networks. *Computer Speech & Language*, 14(4):373–400.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press.

Min, J., Leveling, J., Zhou, D., and Jones, G. J. F. (2010). Document expansion for image retrieval. In *Proceedings of the Conference on Adaptivity, Personalization and Fusion of Heterogeneous Information (RIAO '10)*, pages 65–71, Paris, France.

Morgan, N., Baron, D., Bhagat, S., Carvey, H., Dhillon, R., Edwards, J., Gelbart, D., Janin, A., and Krupski, A. (2003). Meetings about meetings: Research at ICSI on speech in multiparty conversations. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal (ICASSP)*, Hong Kong, China.

Murveit, H., Butzberger, J., Digalakis, V., and Weintraub, M. (1993). Large-vocabulary dictation using sri's decipher tm speech recognition system: progressive search techniques. In *Proceedings of the 1993 IEEE international conference on Acoustics, speech, and signal processing (ICASSP'93)*, pages 319–322, Minneapolis, Minnesota, USA.

Ng, K. (2000). *Subword-based Approaches for Spoken Document Retrieval*. PhD thesis, Massachusetts Institute of Technology, USA.

Oard, D., Demner-Fushman, D., Hajic, J., Ramabhadran, B., Gustman, S., Byrne, W., Soergel, D., Dorr, B., Resnik, P., and Picheny, M. (2002). Cross-language access to recorded speech in the MALACH project. In *Proceedings of the Text, Speech, and Dialog Workshop*, Brno, Czech Republic.

Oard, D. W., Wang, J., Jones, G. J. F., White, R. W., Pecina, P., Soergel, D., Huang, X., and Shafran, I. (2006). Overview of the CLEF-2006 Cross-Language

Speech Retrieval Track. In *Evaluation of Multilingual and Multi-modal Information Retrieval, 7th Workshop of the Cross-Language Evaluation Forum (CLEF 2006)*, pages 744–758, Alicante, Spain.

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

Paice, C. D. (1990). Another stemmer. *SIGIR Forum*, 24(3):56–61.

Pecina, P., Hoffmannova, P., Jones, G. J. F., Zhang, Y., and Oard, D. W. (2007). Overview of the CLEF 2007 cross-language speech retrieval track. In *Advances in Multilingual and Multimodal Information Retrieval, 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007*, pages 674–686, Budapest, Hungary.

Ponte, J. M. and Croft, W. B. (1998). A language modeling approach to information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR'98)*, pages 275–281, Melbourne, Australia.

Popescu-Belis, A., Poller, P., Kilgour, J., Boertjes, E., Carletta, J., Castronovo, S., Fapso, M., Nanchen, A., Wilson, T., de Wit, J., and Yazdani, M. (2009). A multimedia retrieval system using speech input. In *Proceedings of 11th International Conference on Multimodal Interfaces and 6th Workshop on Machine Learning for Multimodal Interaction (ICMI-MLMI 2009)*, Beijing, China.

Porter, M. F. (1980). An Algorithm for Suffix Stripping. *Program*, 14(3):130–137.

Rabiner, L. and Levinson, S. (1981). Isolated and connected word recognition–theory and selected applications. *IEEE Transactions on Communications*, 29(5):621–659.

Renals, S., Hain, T., and Bourlard, H. (2007). Recognition and interpretation of

meetings: The AMI and AMIDA projects. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU '07)*, Kyoto, Japan.

Renals, S., Hain, T., and Bourlard, H. (2008). Interpretation of multiparty meetings: The AMI and AMIDA projects. In *Proceedings of IEEE Workshop on Hands-Free Speech Communication and Microphone Arrays (HSCMA 2008)*, pages 115–118, Trento, Italy.

Robertson, S. E. and Spärck Jones, K. (1976). Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27(3):129–146.

Rose, R. and Paul, D. (1990). A hidden Markov model based keyword recognition system. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP-90)*, volume 1, pages 129–132.

Rose, R. C. (1991). Techniques for information retrieval from speech messages. *Lincoln Lab. J.*, 4(1):45–60.

Rousseau, A., Bougares, F., Deléglise, P., Schwenk, H., and Estèv, Y. (2011). LIUM's systems for the IWSLT 2011 Speech Translation Tasks. In *Proceedings of International Workshop on Spoken Language Translation (IWSLT 2011)*, pages 79 –85, San Francisco, USA.

Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage.*, 24(5):513–523.

Sanderson, M. and Shou, X. M. (2007). Search of spoken documents retrieves well recognized transcripts. In *Proceedings of the 29th European Conference on Information Retrieval (ECIR 2007)*, pages 505–516, Rome, Italy.

Schäuble, P. and Wechsler, M. (1995). First experiences with a system for content based retrieval of information from speech recordings. In *Working notes of the Workshop on Intelligent Multimedia Information Retrieval at 14th International*

Conference on Artificial Intelligence (IJCAI-95), pages 59–69, Montreal, Quebec, Canada.

Sharp, B. and Chibelushi, C. (2008). Text segmentation of spoken meeting transcripts. *International Journal of Speech Technology*, 11:157–165.

Shen, W., Yu, R. P., Seide, F., and Wu, J. (2009). Automatic punctuation generation for speech. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)*, pages 586–589, Merano/Meran, Italy.

Shou, X. M., Sanderson, M., and Tuffs, N. (2003). The relationship of word error rate to document ranking. In *Proceedings of the AAAI Spring Symposium on Intelligent Multimedia Knowledge Manangement, Technical Report SS-03-08*, pages 28–33, Stanford, USA.

Singhal, A. and Pereira, F. (1999). Document expansion for speech retrieval. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '99)*, pages 34–41, Berkeley, California, USA.

Spärck Jones, K., Jones, G., Foote, J., and Young, S. (1996). Experiments in spoken document retrieval. *Information Processing & Management*, 32(4):399–417.

Tao, T., Wang, X., Mei, Q., and Zhai, C. (2006). Language model information retrieval with document expansion. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL '06)*, pages 407–414, New York, USA.

Tür, G. and De Mori, R., editors (2011). *Spoken Language Understanding*. Wiley, Chichester, UK.

Tür, G., Stolcke, A., Voss, L. L., Dowding, J., Favre, B., Fernández, R., Frampton, M., Frandsen, M. W., Frederickson, C., Graciarena, M., Hakkani-Tür, D., Kintz-

ing, D., Leveque, K., Mason, S., Niekrasz, J., Peters, S., Purver, M., Riedhammer, K., Shriberg, E., Tien, J., Vergyri, D., and Yang, F. (2008). The CALO meeting speech recognition and understanding system. In *Proceedings of IEEE Spoken Language Technology Workshop (SLT 2008)*, pages 69–72, Goa, India.

Utiyama, M. and Isahara, H. (2001). A statistical model for domain-independent text segmentation. In *Proceedings of the 39th Annual Meeting and 10th Conference of the European Chapter of the Association for Computation Linguistics*, pages 491–498, Toulouse, France.

Voorhees, E. M. and Harman, D. (1999). The Text REtrieval Conference (TREC): History and plans for TREC-9. *SIGIR Forum*, 33(2):12–15.

Wactlar, H., Kanade, T., Smith, M., and Stevens, S. (1996). Intelligent access to digital video: INFORMedia project. *Computer, IEEE Computer Society Press*, 29(5):46–52.

Wactlar, H. D., Christel, M. G., Gong, Y., and Hauptmann, A. G. (1999). Lessons learned from building a terabyte digital video library. *Computer*, 32(2):66–73.

Wagner, R. A. and Fischer, M. J. (1974). The string-to-string correction problem. *Journal of the Association for Computing Machinery*, 21(1):168–173.

Wartena, C. (2012). Comparing Segmentation Strategies for Efficient Video Passage Retrieval. In *10th International Workshop on Content-Based Multimedia Indexing (CBMI 2012)*, pages 1–6, Annecy, France.

Wartena, C. and Larson, M. (2011). Rich speech retrieval using query word filter. In *Working Notes Proceedings of the MediaEval 2011 Workshop*, volume 807 of *CEUR Workshop Proceedings*, Pisa, Italy.

Wessel, F., Schlter, R., Macherey, K., and Ney, H. (2001). Confidence measures for large vocabulary continuous speech recognition. *IEEE Transactions on Speech and Audio Processing*, 9(3):288–298.

White, R. W., Oard, D. W., Jones, G. J. F., Soergel, D., and Huang, X. (2005). Overview of the CLEF-2005 cross-language speech retrieval track. In *Proceedings of Accessing Multilingual Information Repositories, 6th Workshop of the Cross-Language Evalution Forum (CLEF 2005)*, pages 744–759, Vienna, Austria.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83.

Witbrock, M. J. and Hauptmann, A. G. (1997). Using words and phonetic strings for efficient information retrieval from imperfectly transcribed spoken documents. In *Proceedings of the Second ACM International Conference on Digital Libraries (DL '97)*, pages 30–35, Philadelphia, Pennsylvania, USA.

Zobel, J. (1998). How reliable are the results of large-scale information retrieval experiments? In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR'98)*, pages 307–314, Melbourne, Australia.

Appendix A

Publications List

The following publications led to this thesis or are based on the work appearing in this thesis:

- Datasets
  - **M. Eskevich**, G.J.F. Jones, M. Larson, R. Olderman. Creating a Data Collection for Evaluating Rich Speech Retrieval. In Proceedings of the 8th Language Resources and Evaluation Conference (LREC 2012). Istanbul, Turkey, 2012.
  - M. Larson, M. Soleymani, **M. Eskevich**, P. Serdyukov, R. Ordelman, G.J.F. Jones. The Community and the Crowd: Developing large-scale data collections for multimedia benchmarking. IEEE Multimedia, Special Issue. "Large-Scale Multimedia Data Collections", vol. 19, n. 3, 2012.
  - S. Schmiedeke, P. Xu, I. Ferrané, **M. Eskevich**, C. Kofler, M. Larson, Y. Estève, L. Lamel, G. J.F. Jones, T. Sikora. Blip10000: A social Video Dataset containing SPUG Content for Tagging and Retrieval. In Proceedings of the ACM Multimedia Systems Conference (MM Sys 2013), Dataset Track. Oslo, Norway, 2013.
- SCR Experiments and Analysis:
  - AMI corpus:

* G.J.F. Jones, **M. Eskevich**, A. Gyarmati. Towards Methods for Efficient Access to Spoken Content in the AMI Corpus. In Proceedings of the Workshop on Searching Spontaneous Conversational Speech at ACM Multimedia 2010, Florence, Italy, October 2010.
* **M. Eskevich**, G.J.F. Jones. Exploring Speech Retrieval from Meetings using the AMI Corpus. Computer Speech and Language, Special Issue. Information Extraction & Retrieval, 2014.

– Blip10000:

* M. Larson, **M. Eskevich**, R. Ordelman, C. Kofler, S. Schmiedeke, G.J.F. Jones. Overview of MediaEval 2011 Rich Speech Retrieval Task and Genre Tagging Task. In Working Notes of the MediaEval 2011, Pisa, Italy, 2011.
* **M. Eskevich**, G.J.F. Jones. DCU at MediaEval 2011: Rich Speech Retrieval (RSR). In Working Notes of the MediaEval 2011, Pisa, Italy, 2011.
* **M. Eskevich**, G.J.F. Jones, S. Chen, R. Aly, R. Ordelman and M. Larson. Search and Hyperlinking Task at MediaEval 2012. In Proceedings of the MediaEval 2012 Workshop, Pisa, Italy, 2012.
* **M. Eskevich**, G.J.F.Jones. DCU Search Runs at MediaEval 2012: Search and Hyperlinking Task. In Proceedings of the MediaEval 2012 Workshop, Pisa, Italy, 2012.
* **M. Eskevich**, G.J.F. Jones, M. Larson, C. Wartena, R. Aly, T. Verschoor, R. Ordelman. Comparing Retrieval Effectiveness of Alternative Content Segmentation Methods for Internet Video Search. In Proceedings of the 10th Workshop on Content-Based Multimedia Indexing (CBMI 2012), Annecy, France, 2012
* **M. Eskevich**, G.J.F. Jones, R. Aly, R. Ordelman, S. Chen, D. Nadeem, C. Guinaudeau, G. Gravier, P. Sébillot, T. De Nies, P. De-

bevere, R. Van de Walle, P. Galuščáková, P. Pecina and M. Larson. Multimedia Information Seeking through Search and Hyperlinking. In Proceedings of the 3rd ACM International Conference on Multimedia Retrieval (ICMR 2013), Dallas, Texas, USA.

  - NTCIR-9 and NTCIR-10:
    * **M. Eskevich**, G.J.F. Jones. DCU at the NTCIR-9 SpokenDoc Passage Retrieval Task. In Proceedings of the 9th NTCIR Workshop Meeting, Evaluation of Information Access Technologies: Information Retrieval, Question Answering, and Cross-Lingual Information Access, Tokyo, Japan, 2011.
    * **M. Eskevich**, G.J.F. Jones. DCU at NTCIR-10 SpokenDoc2 Passage Retrieval Task. In Proceedings of the 10th NTCIR Workshop Meeting, Evaluation of Information Access Technologies: Information Retrieval, Question Answering, and Cross-Lingual Information Access, Tokyo, Japan, 2013.

- Evaluation Metrics:
  - **M. Eskevich**, W. Magdy, G.J.F. Jones. New Metrics for Meaningful Evaluation of Informally Structured Speech Retrieval. 34th European Conference on Information Retrieval (ECIR 2012). Barcelona, Spain, 2012.

- Posters:
  - **M. Eskevich**. Improving Search for Unstructured Conversational Speech Archives. RuSSIR/EDBT Summer School "Web of Data", Saint Petersburg, Russia, 2011.
  - **M. Eskevich**, G.J.F. Jones. Investigating Next Generation Speech Search Applications. Workshop on Innovation and Applications in Speech Technology (IAST), Dublin, Ireland, 2012

Appendix B

NTCIR-9 Queries

In this Appendix we list all the 86 of the NTCIR-9 queries in Japanese with their approximate translations into English. They are organised into groups whether they appear to be more like a query for an ad-hoc search (37) or for a query answering system (49).

Query type
Ad-hoc search
4. 藤崎モデルを利用した研究について知りたい。 “I want to know about the research using the Fujisaki model.”
6. トウガラシを使った料理にはどんなものがありますか。 “What dishes do you have that use chilli pepper?”
11. 関西弁自体や関西弁を用いる人について触れている話を知りたい。 “I would like to know stories about kansai dialect or people who use kansai dialect.”
12. 話者の苦手なものについて触れている話。 “Stories about the narrator/speaker’s weaknesses.”
14. 音源定位を扱った研究を知りたい。 “I want to know (about) research on picking up the location of a sound source.”
16. 手法の一部に回帰分析を用いている研究を知りたい。 “I want to know research that contains a section that uses,regression analysis.”
25. 昼食に食べる料理は何があるか。 “What’s to eat for lunch?”
42. 検索質問の拡張方法をしりたい。 “I want to know how to extend the search question.”
43. 健康の維持や病気の予防に効果的な食べ物を知りたい。 “I want to know the food effective in the prevention of disease and maintenance of good health.”
44. ヒトゲノムとは何を示すのか知りたい、 またヒトゲノムに関連する実際の研究について知りたい。 “I want to know about actual research related to the human genome and also want to know what shows the human genome.”
45. 人口が経済や組織へ与える影響を知りたい。 また、その影響に対する政策の内容を知りたい。 “I want to know what influence the population has on organization and the economy. Also, I want to know what impact policy has on that influence.”
46. 子供はどのような遊びをするのか知りたい。 “I want to know what kind of games children play.”
47. 原子力発電以外の発電方法の種類と、その特徴を知りたい。 “I want to know some features of non-nuclear power.”
48. 現在抱えている環境問題に対するリサイクルの効果を知りたい。 また、実際ににどのような物がリサイクルされているのか知りたい。 “Also, I want to know what,those are what are actually recycled.”
49. オリンピックで活躍した陸上競技の選手に関する話題。 “Topics about athletes who took part in the Olympics.”
50. 日本に浸透した外国が発端の文化を知りたい。 また、日本と諸外国の考え方の違いを知りたい。 “Also, I want to know the difference in the,way of thinking of other countries and Japan.”
51. 将来の夢に関する話。 “Stories about future dreams.”
52. 国内で交通の面で不便である場所に関する情報がほしい。 “I want information about places in the country that are bad for traffic.”

Query type
Ad-hoc search
54. 地震に対する下準備や防災方法と、 実際に発生した時の対応方法に関する話題。 “Topics and disaster prevention and how to prepare for earthquakes below on how to respond when it actually occurred.”
55. 就職活動で犯した失敗や、就職後の苦悩に関する話題。 “A story of criminal failure when jobhunting and the resulting anguish.”
59. 日本以外の世界遺産にはどのようなものがあるのか知りたい。 また、遺産のある場所も知りたい。 “I want to know world heritage sites outside of Japan and their locations.”
60. リンゴを使った料理を知りたい。 “I want to know recipes using apples.” 58. 有名なクラシック音楽のタイトルと、その作曲者を知りたい。 “I want to know the title and the composer of the famous classical music.”
61. ウィザード・オブ・オズを用いてデータ収集を行った研究事例を知りたい。 “I want to know about case studies using Wizard of Oz.”
63. ヒトゲノム計画とはどんな計画ですか。 “What is your plan with the Human Genome Project.”
65. クロスバリデーションを使用して学習、評価実験を行っている研究発表。 “Research presentations that tutor cross validation and that are about evaluation experiments.”
67. 遺伝的プログラミング、遺伝的アルゴリズムとはどのようなものか知りたい。 “I want to know what it looks like genetic programming and genetic algorithm.”
69. 実際に起きた原子力関係の事故について知りたい。 “I want to know about the accident of nuclear power relationship that actually occurred.”
71. カレーの作り方を知りたい。 “I want to know how to make a curry.”
72. 色々な発明について知りたい。 できれば、誰が何をいつどのように発明したかを知りたい。 “,I want to know about different kinds of invention. If possible, I want to know who invented what and when.”
75. ダイエットの方法について知りたい。 “I want to know how to diet.”
77. パリ観光の魅力とは何ですか。 “What is the charm of Paris tourism.”
80. 健康的な睡眠の取り方を知りたい。 “I want to know how to take healthy sleep.”
81. 都市の騒音問題について知りたい。 また、主な騒音の原因について特に知りたい。 “I want to know about the noise problem of the city. Also I want to know in particular about the cause of the major noise.”
82. 強い将棋の戦略を知りたい。 また、将棋のプロはどのようになるのかを知りたい。 “I want to know the strategy of strong Japanese chess. Also I want to know how people have become Japanese chess professionals”

Query type
Ad-hoc search
83. バブルの頃の社会について知りたい。 バブル後にどう変わったのかも知りたい。 “I want to know about,society at the time of the economic bubble. I want to know what has changed post-bubble.”
85. ウサギに関する伝承を知りたい。 “I want to know folklore about rabbits.”
86. トマトが使われる料理を知りたい。 “I want to know recipes using tomatoes.”

Query type
Question answering system
1. 野鳥にはどういうものがいるか。 “What sort of things could be called ‘wild birds’?”
2. プロ野球選手の名前を知りたい。 “I want to know the name of the professional baseball player.”
3. 奈良県にはどんなものがありますか。 “What things are in Nara Pref?”
5. 図鑑にはどんな動物が載っているか教えてください。 “In the picture book, please tell me what animal is sleeping.”
7. 釣りで釣れる魚にはどんなものがあるか。 “Did you catch any fish (while fishing)?”
8. 釣りの種類にはどんなものがあるか。 “What kinds of fishing do you have?”
9. スウェーデン人はどんな人か知りたい。 “I want to know what kind of people Swedes are.”
10. ヨーグルトを使った料理を知りたい。 “I would like to know some recipes/dishes using yoghurt.”
13. 給食ではどんなものが出されるか知りたい。 “I want to know what is being served for lunch.”
15. 美術館の名前にはどんなものがあるか。 “What things are in the name of the museum?”
17. 情報処理学会とはどんなものか。 “What is the Information Processing Society of Japan?”
18. 過去に震災が起きた場所を知りたい。 “I want to know where the earthquake has,occurred in the past.”
19. 太陽系には地球以外にどういったものがあるか。 “Are there places in the solar system other than earth that,people have gone to?”
20. 大仏はどこにあるか。 “Where is the Great Buddha.”
21. メロンにはどんなものがあるか。 “What types of melon are there?”
22. マッサージは体のどの部分に行われますか。 “What part of the body is massage done to?”
23. ビタミンはどのようなものに含まれているか。 “What kind of thing contains vitamins?”

Query type
Question answering system
24. 歴代の都知事は誰なのか知りたい。 “I want to know who the successive governors are.”
26. デザートにはどんなものがあるか。 “What’s for dessert?”
27. 高血圧に対する対策は何があるか。 “What are some measures against hypertension.”
28. 山梨で有名なものは何か。 “What well-known things are there in Yamanashi?”
29. アイスやソフトクリームにはどこのものがあるのか。 “Do you have any ice or soft-serve ice cream?”
30. 花火大会は日本のどこでやっているか。 “Where in Japan are you doing fireworks display.”
31. サツマ芋はどのようにして食べるか。 “And how do eat Satsuma potato.”
32. セキュリティーにはどのようなものがあるか。 “What type of thing do you have for security?”
33. お守りにはどのようなものがあるか。 “What types of amulets do you have?”
34. マクドナルドはどこの国にあるか。 “What countries have McDonalds?”
35. アボリジニーはどこの国にいるか。 “What countries have aborigines?”
36. オーロラはどこで見られるか。 “Where can the aurora be seen from?”
37. 主食としてどんなものが食べられているのかを知りたい。 “I want to know what what kind of being eaten as a staple food.”
38. 和菓子の店はどんな場所にあるか。 “Is there any place to shop for sweets.”
39. 大気を汚染する原因は何か。 “What is the cause of the air pollution?”
40. 食中毒になりそうな食べもの、 あるいは食中毒にかかった人が食べたものは何か。 “What are the foods that caused food poisoning or that look to have caused food poisoning?”
41. バドミントンはどのようなところに取り入れられているか。 “Where can I play badminton?”
53. 日本で花火大会が開催される場所を知りたい。 “I want to know where the fireworks will be held in Japan.”
56. 今までの東京都知事の名前を知りたい。また、行った政策の詳細を知りたい。 “I want to know the name of the current governor of Tokyo. I’d,also like to know details of his policies.”
57. 日本国内にある空港の名前と場所を知りたい。 “I want to know the name and location of the airports in Japan.”

Query type
Question answering system
62. 自律神経失調の症状。 “Symptoms of autonomic imbalance.”
64. ロボットの名前の列挙。 “The number of the name of the robot.”
66. 言語処理分野の研究でカイ2乗統計量を使っているもの。 “Those that are using the chi-square statistic in the study of language processing field.”
68. どんなプログラミング言語があるか知りたい。 “I want to know what programming languages are used.”
70. 日本以外の城の名前のリストアップ。 “List of the names of the castle outside of Japan.”
73. おにぎりに入れる具材としてどのようなものがあるか知りたい。 “I want to know what kind of thing as the ingredients you put in the rice ball.”
74. TFIDFとはどのような尺度なのかを知りたい。 “I want to know what kind of a measure with TFIDF.”
76. イギリスに紅茶が伝わったのはいつ頃ですか。 またどのように伝わりましたか。 “When was black tea brought to the UK? How was it brought?”
78. あだ名にはどのようなものがありますか。 “What sort of nicknames do you have?”
79. 蚊に刺されないようにする方法を知りたい。 “I want to know how to prevent mosquito bites.”
84. カルシウムを摂取することにはどのような効果があるか知りたい。 またカルシウムが不足するとどうなるのか知りたい。 “I want to know what kind of effect that the ingestion of calcium. Also I want to know what will happen if the calcium is insufficient.”

Appendix C

AMI Corpus Queries

In this Appendix we list all the 10 development set and 25 test set queries that we have created on the basis of the AMI Corpus using the procedure described in Section 4.2.1.

<top>
<num> Number: 1
<title>
Findings
Whole house control will not be possible
Some extended electronic entertainment control should be possible, i.e. DVD, CD, stereo tuner, vcr
It takes too much time to learn how to use a new remote control (34%)
Remote controls are often lost somewhere in the room (50%)
<top>

<top>
<num> Number: 2
<title>
Personal Preferences
Concentrate on sizzle
Shorten learning curve
Simplicity is good - minimalism
Finding it is important
Should be attractive impulse purchase at 25 euros
We put the fashion in electronics!
<top>

Figure C.1: Development set queries of the AMI Corpus (1-2).

<top>
<num> Number: 3
<title>
Method
Gather suggestions
Accept/eliminate suggestions according to design and budget feasibility
Consult industrial engineer on cost and lead time for each feature (we need to go to market by Sept for the Christmas market)
Consult user interface person on friendliness of features and hierarchy of importance for inclusion
<top>

<top>
<num> Number: 4
<title>
Project evaluation
Project process
Satisfaction on for example
Room for creativity
Leadership
Teamwork
Means (e.g., whiteboard, digital pens, etc.)
New ideas found?
<top>

<top>
<num> Number: 5
<title>
Personal Preferences
To save money the components of the remote should be mass produced and the basic materials should be bought on mass
If we find another company who can produce the required chips, casing, LED and additional materials at a less expensive rate then we, ourselves, can we should go for it.
<top>

<top>
<num> Number: 6
<title>
Suggested Power Options
Solar cells
Hand dynamo?
Kinetic power shake to create power
<top>

Figure C.2: Development set queries of the AMI Corpus (3-6).

<top>
<num> Number: 7
<title>
Personal Preferences
Joystick to rest over joint of forefinger
Mid grip section to house function buttons, positioned for forefinger and middle finger
LCD to rest on wrist
Top FB is power button in basic, enter in menu
Bottom FB switches modes
USB to download programmes, or customise on computer
<top>

<top>
<num> Number: 8
<title>
New Requirements
No more use of the Teletext ? Internet.
Remote Control only used for the television ? too complex
Corporate image recognizable in the product (color, slogan).
<top>

<top>
<num> Number: 9
<title>
Method
Look at pre-existing models
Reuse the essential components
Incorporate new innovations
<top>

<top>
<num> Number: 10
<title>
Energy Source
As energy source we offer a basic battery or, more ingenious, a hand dynamo (like you'd find in 50-year-old torches), a kinetic provision of energy (such as in some modern watches that you shake casually to provide energy), or use of solar cells.
<top>

Figure C.3: Development set queries of the AMI Corpus (7-10).

<top>
<num> Number: 1
<title>
Current Component options
case material supplements
plastic
rubber such as used in anti-RSI stress balls
wood
titanium which we also use in our production of space designs
Unfortunately we can't use the titanium for the double curved cases and latex cases won't allow the use of solar cells as energy source
<top>

<top>
<num> Number: 2
<title>
Three Presentations
Conceptual specification of components properties and materials
Conceptual specification of user interface
Trend watching
<top>

<top>
<num> Number: 3
<title>
Decisions
Components concept
Energy
Chip on print
Case
User interface concept
Interface
Type
Supplements
<top>

Figure C.4: Test set queries of the AMI Corpus (1-3).

<top>
<num> Number: 4
<title>
Personal Preferences
Cheap model
Flat plastic case
Standard battery
Push buttons
Fancy model
Curved titanium
LCD screen
Multiple scroll buttons
Sample sensor sample speaker
<top>

<top>
<num> Number: 5
<title>
Recent Fashion Update
Fashion watchers in Paris and Milan have detected the following trends
This year fruit and vegetables will be the most important theme for cloths shoes and furniture
Also in contrast to last year the feel of material is expected to be spongy
<top>

<top>
<num> Number: 6
<title>
Method
We are going to look to whom we are going to sell the most remotes and adjust our remote the most to these people
Younger people 16-45 are interested in features
We posses about 2 3 of the market share
Elderly people 45 are not interested in features
We posses about 2/5 of the market share
<top>

Figure C.5: Test set queries of the AMI Corpus (4-6).

<top>
<num> Number: 7
<title>
Decisions
Energy kinetic
Chip on print regular
Cases plastic with rubber coating and interchangeable plates
User interface concept
Interface
Type command line interface
Supplements numbers channel changer volume
<top>

<top>
<num> Number: 8
<title>
Project Finance
Selling price 25 euro
Profit aim 50 M euro
Market range international
Production costs max 12-50 euro
<top>

<top>
<num> Number: 9
<title>
Real Reaction
Remote Control
Teaching Cont use plastic with rubber casing powered by kenetic energy
No decision on curvatures look like scroll but push button technology
Separate fashionable covers could be separate product
Yellow with black buttons with company slogan and image suggested
<top>

<top>
<num> Number: 10
<title>
Real Reaction
Remote Control
Problems Issues Production issues
Coordination of user technical marketing etc
Conflict of ideas with cost constraints
Time constraint of the meeting
<top>

Figure C.6: Test set queries of the AMI Corpus (7-10).

<top>
<num> Number: 11
<title>
Personal Preferences
We want something fancy and technological innovative
what we need is something from the future but not tacky
Fruit and veg for clothes and other stuff two options
Stay well away from fruit and veg
Incorporate the whole idea into the product e. g. a potato peeler
How practical is spongy material for an electrical product
How do LEDs fit in
Can stress it is long lasting
<top>

<top>
<num> Number: 12
<title>
Findings
We've got a market
Three out of four
RC s are ugly
four out of five wants to spend more money on them
Three out of four zaps a lot zap-buttons are used 168 times per hour
Power channel-selection volume and teletext-buttons are said to be relevant
Settings audio video sound are not relevant
<top>

<top>
<num> Number: 13
<title>
Tool Training
Try out whiteboard
Every participant should draw their favorite animal and sum up their favorite characteristics of that animal
<top>

<top>
<num> Number: 14
<title>
Personal Preferences
Lightening in the dark
Not too many buttons
A way to find it easily
<top>

Figure C.7: Test set queries of the AMI Corpus (11-14).

<top>
<num> Number: 15
<title>
Closing
Are the costs within the budget
Is the project evaluated
Don't forget to complete final questionnaire and meeting summary
Then Celebration
<top>

<top>
<num> Number: 16
<title>
Agenda
Opening
Project Manager secretary minutes
3 presentations
New project requirements
Decision on remote control functions
closing we have 40 minutes
<top>

<top>
<num> Number: 17
<title>
Findings 2
Most Relevant functions
Channel selection
Volume selection
Power
Teletext flipping pages included
One-time-use functions
<top>

<top>
<num> Number: 18
<title>
Findings
Many controls too complicated
Too many buttons can confuse user
Confusing labelling is bad eg PROG PRG
Simplicity good
but can lead to clunky design
Hard to access advanced functions ef slow motion
<top>

Figure C.8: Test set queries of the AMI Corpus (15-18).

<top>
<num> Number: 19
<title>
Closing
Next meeting starts in 30 minutes
Individual actions
ID the working design
UID the technical functions design
ME the user requirements specification
Specific instructions will be send to you by your personal coach
<top>

<top>
<num> Number: 20
<title>
Findings
The elder part of the audience doesn t care much for innovative features such as speech recognition or an LCD screen
People only use about 10 of the buttons and mostly zap
People often complain that they cant find there remote control 50 so we should build in a feature to support them
<top>

<top>
<num> Number: 21
<title>
Evaluation
Criteria rating with seven-point scale
Shape biomorphic 1
Size small 2
Color bright warm 1
Feel as soft as possible 3
Functionality people won't use it before they buy it paradoxically other features will be main selling points 5
<top>

Figure C.9: Test set queries of the AMI Corpus (19-21).

<top>
<num> Number: 22
<title>
History
Zenith engineer Eugene Polley invented the Flashmatic which represented the industry s first wireless TV remote
Introduced in 1955 Flashmatic operated by means of four photo cells one in each corner of the TV screen
The viewer used a highly directional flashlight to activate the four control functions which turned the picture and sound on and off and changed channels by turning the tuner dial clockwise and counterclockwise
While it pioneered the concept of wireless TV remote control the Flashmatic had some limitations
It was a simple device that had no protection circuits and if the TV sat in an area in which the sun shone directly on it the tuner might start rotating
<top>

<top>
<num> Number: 23
<title>
Evaluation
Criteria
I find this device really fancy
I find this device really handy
I find this device completely functional
I like the cool features of this device
Yes this MANDO banana is easy to use
Of course I would buy the MANDO banana for 25 euros if I needed a remote control
I would change my remote control for a MANDO banana
<top>

<top>
<num> Number: 24
<title>
1 The Shawshank Redemption
Directed by Frank Darabont
Gendre Drama
User Rating 9 0 10
Plot Outline The life of Andy Dufresne changes when he is convicted and jailed for the murder of his wife
Simply amazing
The best film of the 9
<top>

Figure C.10: Test set queries of the AMI Corpus (22-24).

<top>
<num> Number: 25
<title>
Equipment
Equipment we have
photocopier FAX printer
Each office has
A whiteboard
A bulletin board
A coatrack
Each person gets
A chair
A corner desk
A small filinf cabinet 3 drawers
One hanging shelf
<top>

Figure C.11: Test set queries of the AMI Corpus (25).

Appendix D

Details on crowdsourcing experiments methodology

The following paper gives further details on the topic covered in Section 4.3.

M. Eskevich, G.J.F. Jones, M. Larson, R. Olderman. Creating a Data Collection for Evaluating Rich Speech Retrieval. In Proceedings of the 8th Language Resources and Evaluation Conference (LREC 2012). Istanbul, Turkey, 2012.

Creating a Data Collection for Evaluating Rich Speech Retrieval

Maria Eskevich[1], Gareth J. F. Jones[1], Martha Larson[2], Roeland Ordelman[3]

[1] Centre for Digital Video Processing, School of Computing, Dublin City University, Dublin 9, Ireland
[2] Delft University of Technology, Delft, The Netherlands
[3] University of Twente, The Netherlands
{meskevich, gjones}@computing.dcu.ie, m.a.larson@tudelft.nl, roeland.ordelman@utwente.nl

Abstract

We describe the development of a test collection for the investigation of speech retrieval beyond identification of relevant content. This collection focuses on satisfying user information needs for queries associated with specific types of speech acts. The collection is based on an archive of the Internet video from Internet video sharing platform (blip.tv), and was provided by the MediaEval benchmarking initiative. A crowdsourcing approach was used to identify segments in the video data which contain speech acts, to create a description of the video containing the act and to generate search queries designed to refind this speech act. We describe and reflect on our experiences with crowdsourcing this test collection using the Amazon Mechanical Turk platform. We highlight the challenges of constructing this dataset, including the selection of the data source, design of the crowdsouring task and the specification of queries and relevant items.

Keywords: Speech Search, Speech Collection Creation, Speech Retrieval, Crowdsourcing

1. Introduction

The increasing capacity of digital storage media and advances in networking technologies for content delivery are resulting in an ever increasing expansion in the volume of audio and video data accumulated on the web and in offline archives. The nature of this data can vary significantly, from broadcast news and lectures to informal videos recorded for Internet TV channels by semi- and non-professionals. Whatever the form of the data, its potential can only be realised if users can locate relevant content in a timely and efficient manner. The diversity in online archives content potentially gives rise to multiple possible information needs and resulting formulations of search queries. Conventional research on speech retrieval has focused on locating content containing information relevant to a specific information need expressed in a text search query (Garofolo et al., 2000) (Pecina et al., 2008).

In this paper, we describe the design and construction of a speech search benchmarking collection that extends this goal to one when users are considered to be interested in not only informational content, but also in the speaker's intention while uttering their speech. The focus on speaker intention is motivated by the observation that the same words pronounced in different ways can have have different illocutionary meaning - one can promise or warn the listener. We describe our methodology for developing our test collection for investigating search involving speaker intention and our approach to preparing our ground truth. The test collection includes audio-visual spoken content, search queries and corresponding identified relevant data,

Developing meaningful queries and determining relevant segments in videos expressing speaker intention in an unbiased way is particularly challenging. One possible solution for this lies in crowdsourcing, the dissemination of a task to large numbers of human annotators, referred to as workers through an Internet platform (Surowiecki, 2004). This technique permits researchers to gather data from a diverse community of workers in return for micro-payments for their contributions. The potential of crowdsourcing has already been explored in a number of other speech and language applications, e.g. (Snow et al., 2008), (Callison-Burch and Dredze, 2010). Creating a test collection of the type we planned to develop is though muc more demanding of the crowdsource workers than the tasks undertaken in these previous studies.

The test collection described in this paper was developed for use in the Rich Speech Retrieval (RSR) task which formed part of the MediaEval 2011 Benckmark[1]. This RSR task is based on the observation that utterances are actually 'illocutionary speech acts'[2] carried out by speakers (Larson et al., 2011a). The queries are thus designed to search for instances of these speech acts which have been identified in our selected corpus of Internet video.

The paper is structured as follows: Section 2. describes the development of the RSR task we used as an example of dataset construction, Section 3. overviews relevant previous work in crowdsouring for collection of speech and language resources, Section 4. describes the preparation of the data and search collection creation, Section 5. describes task reward issues, Section 6. gives details of the results of the collection exercise, and Section 7. concludes and outlines directions for our future work.

2. Rich Speech Retrieval Task Data Preparation using Crowdsourcing

The focus of the MediaEval 2011 RSR task was to explore the effectiveness with which different types of 'illocutionary speech acts' can be located within audio-visual spoken data. For this we used the ME10WWW archive of semi-professional user generated video downloaded from the Internet video sharing platform blip.tv. The videos for this collection were collected for shows for which the link to

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

one of their episodes had been tweeted on the Twitter social network. Their licenses were checked to conform that they were Creative Commons. The dataset contains 1974 episodes (247 development and 1727 test) comprising a total of ca. 350 hours of data. The development set is small with respect to the test set and is not intended for training, but rather for parameter tuning. The episodes were chosen from 460 different shows, shows with less than four episodes were not considered for inclusion in the dataset. The ME10WWW dataset for the RSR 2011 task is accompanied by automatic speech recognition (ASR) transcripts (Lamel and Gauvain, 2008), which were generously provided by LIMSI (http://www.limsi.fr/) and Vocapia Research (http://www.vocapia.com/). In order to be included in the ME10WWW set, a video needed to have been transcribed by the ASR-system with an average word-level confidence score of > 0.7. The set is predominantly English with approximate 6 hours of non-English content divided over French, Spanish and Dutch. Further details of this video collection can be found in (Larson et al., 2011b).

The richness of the RSR task arises from the specific types of the queries that we are interested in. The information to be found has to be a combination of required audio and visual content and the speaker's intention. Five examples of basic speech acts types were chosen for this task: 'apology', 'opinion' from 'expressives' (speech acts that express on the speaker's attitudes and emotions towards the proposition, e.g. congratulations, excuses and thanks), 'definition' from 'assertives' (speech acts that commit a speaker to the truth of the expressed proposition) 'warning' from 'directives' (speech acts that are to cause the hearer to take a particular action, e.g. requests, commands and advice), 'promise' from 'commissives' (speech acts that commit a speaker to some future action, e.g. promises and oaths).

3. Crowdsouring in Development of Speech and Language Resources

3.1. Background and Relevant Existing Work

Crowdsourcing is a form of micro-outsourcing that allows tasks to be assigned to remote workers who receive a small financial compensation for their work. The importance of manually developed resources to support speech and language research and the cost of developing them means that crowdsourcing has become a topic of great interest in the development of resources for language technology research.

Looking at the general suitability of the use of untrained crowdsource workers in natural language tasks, Snow et al. (Snow et al., 2008) compared the work of domain experts and with that of non-experts recruited in a general crowdsourcing environment for a range of natural language labelling tasks, including recognising textual entailment and word sense disambiguation. The tasks were restricted to selection of multiple choice response or numeric input within a fixe range. Their results demonstrated that the non-expert crowdsource workers can produce work of a similar standard to expert workers. Callison-Burch and Dredze (Callison-Burch and Dredze, 2010) survey contributions to the NAACL-2010 workshop on crowdsourcing for speech and language resource, and highlight a number of important factors which should be taken into account when designing effective crowdsource tasks in this setting. These include issues of how to attract sufficient suitable workers to undertake the task, the level of payment that should be offered to a worker for undertaking a task, careful design of the task, that the instructions should be clear for the expected participants, and how to deal with the problem of workers who try to cheat on the task to earn payment without undertaking the work properly. It should be noted that here too the tasks examined were relatively straightforward such as involving the selection of appropriate labels from among a number offered or undertaking translation into a language in which the worker is fluent.

In the area of speech resource development Marge et al. (Marge et al., 2010) found that crowdsource workers are able to transcribe speech of varied qualities with reasonable accuracy. Evanini et al. (Evanini et al., 2010) investigated the more challenging task of transcribing non-native read-aloud and spontaneous speech, they found that even merging the results multiple workers produced errorful transcriptions particularly in the case of the spontaneous speech. Thus even in a clearly defined and apparently obvious task, crowdsourcing does not provide a simple solution to challenging tasks. Lane et al. (Lane et al., 2010) found some success exploring the related speech task of collecting spoken corpora using crowdsourcing, but identified issues in relation to the training of the speakers to undertake the task.

In the field of information retrieval, one of the common challenges in the development of test collections for system testing is establishing the relevance of available documents to a user search query. Crowdsourcing provides an intuitively appealing solution to this problem. In this scenario workers can be shown a query and asked whether specific documents are relevant to the query. An early study exploring this topic is described in (Alonso et al., 2008). This examined the important topics of establishing whether workers are actually qualified to carry out the task for which they are volunteering, and seeking to identify those not undertaking the work properly. This is a particular problem in relevance assessment of this nature since clearly the person requesting the work cannot manually check the accuracy of all submitted work. A further study on this topic by Grady and Lease (Grady and Lease, 2010) examined the topic of reward for work done. They examined the issue of worker pay, particularly considering the impact of offering a bonus to the worker for good work, where bonuses were manually assigned when checking the quality of work carried out. Interesting they observed that workers appeared to be attracted to do more work where a bonus was offered, and that they completed the work with greater accuracy on average.

From these existing studies it is clear that crowdsourcing can make a valuable and cost effective contribution to the development of language technology resources. However, workers can find even apparently simple tasks challenging and produce unsatisfactory work. All the tasks examined here are conceptually quite straightforward either relying on workers to use non task specific recognition skills, their own special linguistic knowledge, e.g. being bilingual, or

Table 1: Number of collected queries per speech act for MediaEval 2011 development and test sets

	Speech act type					Total
	Apology	Definition	Opinion	Promise	Warning	
Development Set	1	8	17	1	3	30
Test Set	1	17	21	5	6	50

transcribing or uttering some speech. There is no personal creativity required to perform any of these tasks. General factors include, the common observation that here is a persistent problem of some workers trying to cheat to receive payment without completing work properly. Also there are interesting questions requiring further exploration relating to the levels of pay offered for tasks, and the potential impact of bonus payments on loyalty and quality of work.

3.2. Amazon Mechanical Turk

Currently, the most widely-used platform for crowdsourcing is Amazon Mechanical Turk[3]. In the Mechanical Turk (MTurk) setting tasks are referred to as 'Human Intelligence Tasks' or HITs. To initiate a task, the requester uploads a HIT consisting of relevant instructions, questions, files, etc to be used by the workers while completing the HIT. When the workers have carried out HIT, the requester reviews the completed work and confirms payment to the worker with a previously set payment. If the requester is not satisfied with the work carried out, they can opt not to the worker. Potentially, the requester can also give the worker a bonus, as discussed previously, in order to both express appreciation for the quality of the work and to motivate the worker to continue.

4. Development of an Effective HIT

While we had a clear specification of the test collection that we wished to develop using MTurk, as highlighted in the previous section, earlier work using crowdsouring for the development of speech and language resources set workers much less complex tasks than we wished them to undertake. In many cases deliberately designing the task to be as simple as possible to reduce the effort involved for the worker, to maximise the potential number of workers interested and qualified to undertake the task and to minmise the chance of them making mistakes. Thus the creation of our test collection was actually exploring the research question of whether untrained MTurk workers can undertaken extended tasks which require them to be more creative than those examined previously.
For each HIT we required the worker to carry out the following activities:

- View an assigned video to attempt to locate the presence of speech act.
- Label the specific time at which the speech act begins and ends in the video.
- Accurately transcribe the words spoken within the time limits of the labeled speech act.
- Write a full sentence query which they believed would be able to refind this speech act, and write a short web style query to refind the speech act.

The task is thus much longer and more complex than tasks typically offered to workers, since it requires them to carry out multiple activities and also to be creative since they are asked to develop their own search queries as part of the HIT.
While assigning the HIT to a worker we did not require them to have any specific knowledge or experience of work with audio and video data. However we used internal Amazon MTurk platform information about the previous performance of this registered user in order to select those that are familiar with the system itself, and whose previous results satisfied other requesters. This measure is called HIT Approval Rate, and is a simple ratio of how many HITs submitted by each worker have been approved by the requesters. For our HIT we allowed only the workers with HIT Approval Rate greater than or equal to 90 to undertake the task.

4.1. Data Management

The videos in the blip TV dataset vary in length. We felt it unrealistic to expect workers to view extended videos while looking for speech acts. Also we observed a bias in our initial crowdsourcing trials with this data that workers tend to identify noteworthy segments in the first few minutes of a video. This may be caused by the fact that they are paid for each HIT, and are therefore interested in completing more HITs in time they have available. Thus for the original 247 and 1727 videos in the ME10WWW for development and test set respectively, we prepared 562 and 3278 starting points for longer videos at a distance of approximately 7 minutes apart. These starting points were then randomly allocated by the Amazon MTurk platform to be presented to the workers in each HIT. Even after presegmenting the videos into shorter parts, the workers rarely found noteworthy content later than the third minute from the start of playback point in the video.
The main technical challenge was that since MTurk does not support playback of multimedia files, the workers needed to watch videos stored on an external server. The path to the remote file was embedded within the html-code of the HIT page. Thus the video player used had to be compatible with different operating systems and browsers. Restrictions on this issue had to be made clear to workers in the general description of the task.

4.2. Data Collection Procedure

We used a three-stage approach for our collection procedure. First we prepared and uploaded a pilot version of the HIT and received 55 results, 34 of which we approved.

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

Figure 1: Amazon MTurk HIT example that was used to gather long and short queries (Questions 6 and 7) associated with certain speech acts (Question 1), time stamps (Questions 3 and 4) and transcript (Question 5) of the relevant content

These answers were not included in the final set, but provided us with valuable feedback to refine our HIT.

The initial HIT was found to contain too many concepts that workers did not understand clearly. The revised HIT thus avoided words such as 'transcripts', 'quote', 'categories' etc. The request to find the segment with a certain speech act was expressed indirectly, and was hard to understand. Initially the description of the task was: "Please watch the video and find a short portion of the video (a segment) that contains an interesting quote. The quote must fall into one of these six categories". Analysis of the results showed that workers were confused with the type of phrases they had to find, the concept of transcription was mixed with the general description of what was said during the video, the workers who were probably not familiar with the video player gave wrong time onsets and offsets for identified relevant speech events.

In our revised HIT, we attempted to make use of a concept with which general workers will be more familiar when working with the videos – sharing. The concept of sharing seemed to us to be part of the everyday experience of people who work with the Internet and would provoke a more natural human response setting. The new phrase of the HIT became: "Imagine that you are watching videos on YouTube. When you come across something interesting you might want to share it on Facebook, Twitter or your favorite social network. Now please watch this video and search for an interesting video segment that you would like to share with others because it is (an apology, a definition, an opinion, a promise, a warning)".

The other way to make the workers more familiar with the task was to provide full examples of how all the questions in one HIT can be answered for each speech act type. At the head of each HIT page we put a link to a webpage with an example video with all fields filled in and a dropout window on the page of the HIT itself with only the textual answers.

All these changes resulted in more appropriate answers to all the questions from a large majority of the workers. An

example of the HIT page is shown in Figure 1.
Additionally the initial trial HIT enabled us to set a suitable worker reward that both we as requesters and workers would be comfortable with. The setting of rewards is discussed in further detail in Section 5.

4.3. HIT Refinement

We ran the revised version of the HIT on the development set. An unexpected finding was that the difference in the types of speech acts, together with the limited time that workers are usually prepared to spend on one HIT caused a problem of unbalanced results. We found that it is much easier to assign something that was said by the speaker in the video to be his or her opinion than to find a warning or a promise. We think that this particular feature and the nature of the videos themselves, where we observed that there are not so many incidents of speakers making an apology or promising something in the videos, made our results unbalanced, and that this meant the number of 'opinions' was significantly higher than the number of the four other types. To avoid this situation for the video test set, we decided to run the HIT twice: one HIT with the option of 'rare' speech acts and one only for opinions. At the same time we added new questions about the speaker's appearance and behaviour to help us to detect workers who were not doing their work properly: 'Write one sentence describing the person that you see in this video. If there is more than one person, who is the person who seems to be the most important for the video'; 'Does this person have any particular mannerisms (gestures that they use, particular way of talking, nervous habits)? Please write one sentence to describe anything that you notice'. Our attempt to simply separate the HITs for 'rare' speech acts failed because workers seemed to be more amused by the new questions that had no meaning for our research, the information provided was not useful for our research, and apparently it was harder to find instances of the rare acts in the data.
Thus finally we decided to return to the original single HIT for all types of speech act with the revised wording of the instructions in collecting the queries, assuming that the lack of balance between acts might be just a feature of this dataset, and we used the same HIT questionnaire for the test set as for the development set.
In total we collected 30 queries for the development set and 50 for the test set, Table 1 shows the statistics of the collected speech acts. Examples of long queries that look like a natural sentences, short queries that correspond to the type of queries usually addressed to the Internet search engines, and transcripts of the relevant content are given in Table 2.

5. Reward Levels

The reward to a worker paid by a requester is generally set in the task description, and the workers take this value into consideration together with the general HIT description when choosing whether to undertake the task. Once some workers submit their work, statistics of the average reward per hour for this HIT are available for viewing by other potential workers to take into account when considering when to take on a HIT.
The availability of the option for the requester to change the reward amount when assessing the data submitted for a HIT by a worker gives the possibility to introduce the notion of a bonus (extra reward) or to decrease the reward if the requester finds that the work was not done correctly.
Our initial trial HIT enabled us to define a suitable reward that both we as requesters and workers would be comfortable with.
Initially we started with a reward of 0.11$ per HIT plus bonus per type of the illocutionary act (the sum varied depending on the rareness of the act). Due to the complicated and confusing formulation of the HIT, we received negative feedback from the workers. Apparently this task was inappropriately time consuming for the reward we set. Thus we worked on reformulation of the HIT to simplify it as described in Section 4.3. Also we added a clear statement in the HIT description that we are a non-profit organization, and we raised the reward to 0.19$ and made the workers themselves suggest their own bonus in the range from 0 to 0.21 $. Our motivation for allowing workers to choose their own bonus level was to demonstrate trust in them and appreciation of their work, which we conjectured would reinforce workers in carrying out work more thoughtfully and carefully. Interestingly, giving workers an opportunity to judge the difficulty of the task themselves resulted in useful answers with little evidence of greed (i.e., people didn't always choose the highest possible bonus). Workers were given a text box in which to provide justification for their requested bonus. Most of them took this opportunity to add a short comment. Sometimes the workers even explained that they were not sure of how well they had done on the task and therefore did not deserve the bonus for completing this HIT. Apart from spam submissions, we found bonus requests always to be reasonable.
In total, the cost of the completing the HITs was the following (10 % of all the rewards paid goes to the Amazon MTurk platform):

- price of the devset: (55*0.11 + 7*0.19 + 46*0.19) = 20$ + 16.12 = 36.12 + 10% = 40 $;
- price of the testset: (47.88 + 25.2 (approximately the amount of bonus money)) + 10% = 80.388 $.

6. Comments on the HIT Results

Since working with video as required by this HIT is not a common task for crowdsourcing workers, we wanted to support workers that took the effort to undertake our HIT. Thus we accepted reasonably good answers that could not be used in our test collection and even award a small bonus (0.02$) with an explanatory comment to the worker. Another reason for keeping the reward (even a small one), and not rejecting the work carried out by a worker, is that the MTurk platform monitors the level of rejection per worker in order to detect spam or any other inappropriate activity (HIT Approval Rate). Thus we did not want to decrease the HIT Approval Rate of workers who undertook our task and did a substantial amount of work, but could not make it correctly due to misunderstandings due to the nature of the task.

Table 2: Examples of 2 types of queries associated with speech acts and transcripts for the relevant segments

Speech act	Queries of 2 types and Transcript
Apology	Transcript: I'm here now with Terry Denison, who's the President of the Swim Coaches Association in Great Britain. Thanks for joining us on The Morning Swim Show. Oh, well, thank you for inviting me. Actually, I'm Chairman of the Swim Coaches Association.. it's a slightly.. Chairman Denison, I apologize.
	Long query: How does Anita Burns, host of the Open Mind Show, save face after the embarrassing comment she made during her interview with Victoria Edwards? Short query: Peter Busch president chairman Denison morning Swim Show apology
Definition	Transcript: Equality. How you wanna be equality for his people as far as material possessions and in verse fifteen he compared that to what was said in exodus. as it is written, He that gathered much had nothing over; and he that gathered little had no lack.
	Long query: Short video segment defining equality using a segment from a religious book Short query: Equality religious definition
Opinion	Transcript: Apple produces this new platform all of a sudden you know within roughly a short period that is twenty five thousand applications. Apple didn't write these applications other people did. You know you look at Twitter A Twitter is as minimal as service it doesn't offer very much and yet thousands of applications are out there adding value
	Long query: What makes Twitter more popular than Apple in terms of value Short query: Why Twitter and not Apple?
Promise	Transcript: They will launch a new effort to conquer a disease that has touched the life of nearly every American, including me... by seeking a cure for cancer in our time.
	Long query: Obama promises to find a cure for cancer! Short query: Obama healthcare promises
Warning	Transcript: And there are some here coming for their own purposes and for selfish reasons that are not for your highest good. Not everything out there out there is wonderful and good
	Long query: Woman warning that we should be aware of the intentions of the things going on in cosmos Short query: good in cosmos spoiled by selfish reason

We had several workers who completed several HITs for the development set, but did not participate in the HIT for the collection of the test set. This lack of overlap might cause certain differences in the way people formulated their queries and chose relevant segments.

In general, for the test set the number of accepted HITs with the speech act chosen was 58.1%, where 39.5% were suitable for use in the dataset and 18.6% were accepted, but not included (either some of the fields were missing or there were some issues with the work of the video). For the remaining 41.9% of HITs the worker indicated that they were unable to find an instance of any of the illocutionary acts, we found that 35% of these responses were reasonable, while the other 6.9% were disputable.

It is worth commenting that the data provided by the workers required an additional manual assessment by the requester because there were a number of spam entries. Some types of spam were easy to capture, even automatically, for example when all fields for the HIT were empty or contained the same word. However some spam workers were more creative and copied field by field the example we have provided in the HIT heading. In our HIT formulation there was the possibility to state that there is no speech segment that can be associated with any of listed speech acts and still get the basic reward. During the collection for the test set, only 16% of these answers seemed to be disputable, and thus could be classified as spam or improper work. In these cases the workers were not paid. This problem was also observed for confirmation of translations in (Callison-Burch and Dredze, 2010), an effective solution was to use images of the text which could not be copied, rather than using text itself. While manual checking by the requestor of the claim that there was no speech act present in the video shown for the HIT was practical for the small scale data collection undertaken here, it would quickly become prohibitive for larger collections. In these cases passing these video playback points to a second round of crowdsourcing might offer a means to check the judgement of the first worker.

Related to this issue, while we mainly only used one assessor to decide on the assign a speech act to each segment and create an appropriate query, this could be done multiple times for each segment. When the dataset was examined by participants in the MediaEval 2011 RSR , in more than 50 % of the cases there was a general consent on the information provided by the worker, however some cases were clearly disputable. Without any information about the workers background, language proficiency and the features in the audio or video that affected the assignment of the speech act to a certain utterance, it is not clear how they made their assignment. To better understand such cases, the same video segment could be given to multiple crowdsourcing workers during the assignment and query generation stage. Segments that all workers agree upon could be chosen for use in the retrieval collection. This would have the additional advantage that there would be multiple queries available for each segment, enabling more extensive RSR experimentation. Additionally, segments, assigned speech acts and associated queries, could be used as

a separate crowdsourcing task in order to get other workers opinions on the reliability of the initial workers judgments in the first round of experiment.

7. Conclusions and Future work

This paper has described our successful development of the test collection for the Rich Speech Retrieval task at MediaEval 2011. This work has demonstrated that is possible to use crowdsourcing workers to carry out more extensive and complex tasks in the creation of resources to support speech and language research than has previously been shown. Our experiences in developing the worker task demonstrate the importance of understanding the concepts and vocabulary with which workers are likely to be familiar and to ensure that the required task relates to their general life experiences. Related to the description of the actual task, crowdsourcing workers are currently generally not used to dealing with video and audio and thus tend to be confused by the technical terminology.

The requirement to fully understand the instructions and to successfully complete multiple stages in the HIT, and the somewhat subjective nature of some of the speech acts in the video data means that it may not be possible to reduce the high failure rate of the HIT. In this case while roughly 90 % of the workers were judged to seeking to fulfill the HIT to the best of their ability, and paid accordingly by the requester, with only 10 % not receiving payment, less than 50 % of the paid work was judged suitable for inclusion in the test collection. While the low cost of crowdsouring means that the amount of money wasted is not high, it would be preferable to make the HIT more efficient. Seeking to do this could form the basis of further investigation.

One disadvantage of using this approach is that the crowdsourcing platform is not specifically tuned for video processing, thus we had to use an external video player. Therefore some technical problems that the workers had (the video was not displayed or it was too slow) are hard to control, and it is impossible to detect whether they are caused by the interaction of the platform with external software, or the workers Internet connection affects the video display.

We found that the choice of award level for demanding tasks of the type specified here was very important. Setting the award too low in our initial trial HIT was very unpopular, but this problem was easily addressed when the reward amount was raised, and workers were found to generally be honest in their self assessment of the quality of their work for the HIT and the reward that they deserved.

We presented videos that are longer than 7 minutes to the workers several times, each time starting the playback at a distance of approximately the same length in order to get the queries from all of the data and not only the beginning of the files. However even with this setting, as noted earlier, workers tend to watch only a maximum of the first 3 minutes from the start of the playback which biases our results. Using a smaller window between the playback start points might be a solution to this problem. Although this change is not completely straightforward due to the presence of music and other non-speech sounds that has to be taken into account while assigning the position of playback start points within each file.

In future work we plan to collect more retrieval queries with speech act information for this dataset through crowdsourcing. We assume that the retrieval process might benefit when queries of different speech act types are processed differently. However, the number of queries of different types in the current test collection is not sufficient to draw conclusions in this regard from experiments. We will investigate whether the creation of a set of HITs to collect the query set, and then checking their reliability through crowdsourcing could form a basis for the creation of a large retrieval collection for future investigation in the domain of rich speech retrieval.

8. Acknowledgments

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

9. References

Omar Alonso, Daniel E. Rose, and Benjamin Stewart. 2008. Crowdsourcing for relevance evaluation. *SIGIR Forum*, 42(2):9–15.

Chris Callison-Burch and Mark Dredze. 2010. Creating speech and language data with Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk (CSLDAMT 2010)*, pages 1–12, Stroudsburg, PA, USA. Association for Computational Linguistics.

Keelan Evanini, Derrick Higgins, and Klaus Zechner. 2010. Using Amazon Mechanical Turk for transcription of non-native speech. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk (CSLDAMT 2010)*, pages 53–56. Association for Computational Linguistics.

John S. Garofolo, Cedric G. P. Auzanne, and Ellen M. Voorhees. 2000. The TREC spoken document retrieval track: A success story. In *Proceedings of RIAO 2000*, pages 1–20.

Catherine Grady and Matthew Lease. 2010. Crowdsourcing document relevance assessment with mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk (CSLDAMT 2010)*, pages 172–179. Association for Computational Linguistics.

Lori Lamel and Jean-Luc Gauvain. 2008. Speech processing for audio indexing. In *Advances in Natural LanguageProcessing*, pages 4–15. Springer Berlin / Heidelberg.

Ian Lane, Alex Waibel, Matthias Eck, and Kay Rottmann. 2010. Tools for collecting speech corpora via mechanical-turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk (CSLDAMT 2010)*, pages 184–187. Association for Computational Linguistics.

Martha Larson, Maria Eskevich, Roeland Ordelman, Christoph Kofler, Sebastian Schmiedeke, and Gareth J. F. Jones. 2011a. Overview of Mediaeval 2011 Rich Speech Retrieval task and genre tagging task. In Martha Larson, Adam Rae, Claire-Hélène Demarty, Christoph Kofler, Florian Metze, Raphaël Troncy, Vasileios Mezaris, and Gareth J. F. Jones, editors, *Proceedings of the MediaEval 2011 Workshop*, volume 807. CEUR-WS.org.

Martha Larson, Mohammad Soleymani, Pavel Serdyukov, Stevan Rudinac, Christian Wartena, Vanessa Murdock, Gerald Friedland, Roeland Ordelman, and Gareth J. F. Jones. 2011b. Automatic tagging and geotagging in video collections and communities. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval (ICMR 2011)*, pages 51:1–51:8. ACM.

Matthew Marge, Satanjeev Banerjee, and Alexander I. Rudnicky. 2010. Using the Amazon Mechanical Turk for transcription of spoken language. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2010)*, pages 5270–5273. IEEE.

Pavel Pecina, Petra Hoffmannová, Gareth Jones, Ying Zhang, and Douglas Oard. 2008. Overview of the CLEF 2007 Cross-Language Speech Retrieval Track. In Carol Peters, Valentin Jijkoun, Thomas Mandl, Henning Müller, Douglas Oard, Anselmo Peñas, Vivien Petras, and Diana Santos, editors, *Advances in Multilingual and Multimodal Information Retrieval*, Lecture Notes in Computer Science, pages 674–686. Springer Berlin / Heidelberg.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2008)*, pages 254–263, Stroudsburg, PA, USA. Association for Computational Linguistics.

James Surowiecki. 2004. *The Wisdom of Crowds*. New York, Random House Inc.