Examining the Contributions of Automatic Speech Transcriptions and Metadata Sources for Searching Spontaneous Conversational Speech

Gareth J. F. Jones Ke Zhang Eamonn Newman Adenike M. Lam-Adesina
Centre for Digital Video Processing
Dublin City University
Dublin 9, Ireland
{gjones,kzhang,enewman,adenike}@computing.dcu.ie

ABSTRACT

The searching spontaneous speech can be enhanced by combining automatic speech transcriptions with semantically related metadata. An important question is what can be expected from search of such transcriptions and different sources of related metadata in terms of retrieval effectiveness. The Cross-Language Speech Retrieval (CL-SR) track at recent CLEF workshops provides a spontaneous speech test collection with manual and automatically derived metadata fields. Using this collection we investigate the comparative search effectiveness of individual fields comprising automated transcriptions and the available metadata. A further important question is how transcriptions and metadata should be combined for the greatest benefit to search accuracy. We compare simple field merging of individual fields with the extended BM25 model for weighted field combination (BM25F). Results indicate that BM25F can produce improved search accuracy, but that it is currently important to set its parameters suitably using a suitable training set.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Indexing Methods*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Algorithms, Experimentation

Keywords

searching spontaneous speech transcriptions, metadata, data fusion, field combination

1. INTRODUCTION

Spontaneous speech forms a natural and often almost unconscious means of communicating information between individuals. Increasing archives of digitally recorded spontaneous speech are creating new opportunities to access information contained in this data. However, retrieving relevant content presents significant challenges. Previous research in spoken document retrieval (SDR) for broadcast news, notably the TREC-8 and TREC-9 tasks, has demonstrated that, when handled appropriately, there is little difference in retrieval effectiveness between errorful transcriptions generated using automatic speech recognition (ASR) and a near accurate[1] manual transcription [3]. However, much of this data is read speech and well defined distinct document units are generally easily identifiable. Data nearer to spontaneous speech was used in the Video Mail Retrieval using Voice (VMR) project [5], where there was degradation in retrieval performance for automatically indexed data compared to manual transcriptions, but in this case the manual transcriptions were completely accurate. While the documents in the VMR collection generally comprise spontaneous speech, they are still distinct individual documents.

Spontaneous conversational speech, where document boundaries are often not well defined, raises a number of new issues for search. The CLEF Cross-Language Speech Retrieval (CL-SR) uses data from the Malach oral history collection to explore retrieval of spontaneous speech with significant conversational elements in the context of cross-language information retrieval [9]. These collections can of course also be used to explore monolingual search without the additional complexities associated with cross-language search. An interesting feature of this collection is that ASR document transcriptions are accompanied by several automatically and manually derived metadata fields. Results from CLEF workshops held in 2005 and 2006 show that retrieval effectiveness using only the ASR fields is poor, while using metadata gives much better performance. It is not however clear exactly why this is the case, this topic is explored in more detail in Sections 2 and 3. Retrieval is clearly shown to be improved by combining metadata fields, with manual metadata being considerably more useful than automatic metadata. While the utility of field combination is clear, it is important to consider how these fields should best be combined for best results, we explore this issue in Section 4 with specific reference to the BM25 model, based on the analysis in [8], and report experimental results using the CLEF CL-SR collection in Section 6. Note that while metadata is clearly important for retrieval of spontaneous speech in the

[1]Word Error Rate (WER) $\approx$ 10%.

case of these CLEF collections, it is not possible to explore whether it might also improve retrieval effectiveness results for the earlier SDR tasks reviewed above beyond their already high values, since comparable metadata fields do not exist for these collections.

The remainder of this paper is organised as follows: Section 2 discusses the nature of spontaneous conversational speech, Section 3 examines searching this spontaneous speech and associated metadata, Section 4 explores issues in field combination and the BM25F model, Section 5 outlines the CLEF CL-SR test collections, Section 6 gives our experimental results and analysis, and finally Section 7 concludes the paper.

2. SPONTANEOUS VS SCRIPTED SPEECH

Unlike more deliberately generated written text communication or speech read from a script, when speaking spontaneously a person will often convey many details in an informal and unstructured way, and frequently make considerable use of the context in which they are speaking and the background of the audience which is being addressed, whether it be an individual, a business meeting, a class of students or a general interest grouping.

The degree of genuine spontaneity will depend on the circumstances in which they are speaking, and their experience in ensuring that what they say is unambiguous and will not return to haunt them in the future. Contrast the implications of a slip of the tongue in a social gathering between close friends or a business meeting with regular colleagues, and a live radio or TV interview of a leading politician or a contract negotiation meeting between companies. In the former cases perhaps a simple clarification or apology will often suffice if a slip is made, or perhaps no one will even notice and the exchange can proceed without interruption, in the latter cases there may be significant long term implications of using a certain expression or even implying something unintentionally. While the political interview is spontaneous in the sense that it is not scripted, the interviewer will often have an agenda of points that they wish to raise and include sufficient context in their questions and responses to inform the listener of the topic under discussion, and the politician will typically respond carefully, for the reasons stated above. In a social gathering or local business meeting such contextual information will generally be missing from the exchanges, since the participants are familiar with each other or the subject under discussion and many details to not need to be stated. Essentially there is a large degree of *tacit knowledge* at play in such exchanges.

Words spoken may make reference to what is known to the participants to establish or maintain common understanding of the point under discussion, but important words related to the topics may often be new or very specific ones important in conveying new information. Such words will often be outside the vocabulary of an ASR system meaning that they cannot be recognised correctly, and will therefore not appear in such automatically generated transcriptions and in consequence not be available for search. In the case of ASR transcriptions of spontaneous speech we might well expect them to contain words with low average specificity in terms of identifying relevant documents, i.e. recognised words may appear in a higher average proportion of document transcriptions, making it harder to rank relevant documents reliably. We demonstrate that this can indeed be the case in Section 6.

For more structured interviews the need to establish the context for listeners will mean that a greater number of topic related words appear and that these additional words are more common in the language as a whole. These more common topics words are more likely to be within the vocabulary of an ASR system, particularly if it has been adapted to the domain of interest. Obviously examining this hypothesis fully would require access to suitable corpora of accurately transcribed speech. However, if it is found to be even partially correct, the success of TREC SDR may be to some degree attributable to these words which can be recognised correctly, as well as redundancy and term co-occurrence effects. Searching spontaneous conversational speech may thus be an intrinsically much more difficult task.

3. SEARCHING SPONTANEOUS SPEECH

These observations potentially have significant implications for searching of spontaneous conversational as speech. If the words are not articulated between participants while expressing an opinion, developing an idea or clarifying some point, since they are already common knowledge, then searching an audio recording to find material containing content pertaining to these details is clearly going to present problems, since many of the obvious search words are just not present in the speech. This problem would be significant in itself if the details of the conversations were accurately transcribed. However, the volume of speech means that it is only practical to perform transcription using ASR which inevitably introduces introduces errors arising from various sources. Thus the issue of the absence of important context descriptive content in conversational speech is compounded by the presence of errors in the transcription. As has been observed previously [3], the issue of transcription errors has not proven to be a significant problem for searching spoken segments which can be broken into distinct documents, such as the easy segmentation of a news broadcast, and which are scripted to explain the context of the material covered, again as exemplified by broadcast news stories. However, conversational speech represents a new search problem combining the previously described problems of the absence of contextual review, ASR errors and also uncertainty in topic boundaries, and indeed even the scope of topics within the data to which boundaries might be assigned. Where there is a lower density of topic specific words being spoken, recognising individual spoken words correctly becomes more important. This lack of redundancy means that failure to correctly recognise individual useful words may have apparently disproportionately significant implications for retrieval.

In order in facilitate effective search of this errorfully transcribed data where topic boundaries are unclear and which lacks articulation of much of the associated knowledge, it would seem obvious to suggest that the content should be annotated with terms useful for improving search reliability. The question then arises how should such annotation, or descriptive metadata, be assigned to the speech transcription? One option obviously is to enter this manually, although this will often be extremely expensive, and will only be justified in limited cases. The other option is to seek to assign

metadata automatically or possibly semi-automatically. The availability of suitable metadata will depending on the type of data under consideration.

In the field of education there is growing interest in recording of lectures. These can then be made available for download for students for private study to reinforce lectures or distance learning. Beyond this basic use, lectures recordings are also potentially a very valuable new resource, since they are often sources of the lecturers tacit knowledge of a subject which they fail to include in written materials associated with the course, or which arise unexpectedly during the lecture, possibly promoted by questions from the audience. Whilst a student taking a particular course will be able to identify the lecture recording that they wish to access, as such archives grow it will clearly become impractical to search lecture archives manually. This will be true for small archives in the case of distance learning students or those searching a remote archive, when the student doesn't know exactly where the information they are interested in is located. Thus we should seek to make lectures searchable. Importantly making them searchable also significantly increases their value as a knowledge source for students wishing to learn about a subject. The first stage in making a lecture recording searchable is to transcribe the content using ASR. However, even after the correct lecture has been located, viewing a complete lecture takes a considerable amount of time, and efficiency in locating relevant sections of a lecture can be improved by adding structure to the lecture. Associated with a lecture there will often be a set of electronic slides.

In previous work we demonstrated that where such slides are available, even highly errorful lecture transcriptions can be segmented and assigned to their related slide with a high degree of reliability [4]. Associating relevant manually created metadata with each section of a noisy lecture transcription has several positive advantages. Since they are created manually, the contents of the slides are accurate, and since they are slides designed to support a lecture presentation, they are likely to contain concise statements of the key points to be raised in the lecture, and to do so using carefully selected vocabulary used to describe the topic under discussion. By contrast the ASR transcription of the lecture will contain mistakes and will almost certainly fail to recognise important domain specific words which are outside the vocabulary of the ASR system. In addition, the lecturer may fail to use accepted domain specific vocabulary in their description while extemporizing on the subject under discussion[2]. Thus annotating the transcription with the slides can improve the indexing and search of this content. Annotating spontaneous speech in this way is only possible if there is high quality descriptive content available that can be associated with the transcription. The structure of lecture presentation means that the problem is generally one of alignment within a limited search space. Other environments will constitute a much more challenging metadata association task.

While the contents of a formal lecture are generally spontaneous, they are not often truly conversational, unless the lecturer chooses to engage in extensive interaction with the class. Within education a small group tutorial forms a better example of a spontaneous conversational speech environment. Such sessions may possibly be even more valuable than formal lectures. The discussions will be largely unstructured with many unanticipated comments from the tutor and the students, with much greater potential for the expression of ideas that are not available in formal instruction associated with the course. This environment introduces the problems associated with searching spontaneous conversational speech discussed earlier. A key question if the speech is to be augmented with metadata for searching, where might this metadata come from? Research at IBM has explored the automated delivery of information associated with a meeting [1]. The *Meeting Miner* system performs live ASR on the audio stream emerging from a meeting, and analyses the resulting transcription to form questions or queries to archives related to the meeting, and returns items from the archive to the participants in an attempt to provide them with additional information that they may find useful to enhance their participation in the meeting. Information gathered in this way might potentially be used to annotate the meeting transcription, to more fully describe the topic under discussion in the meeting and thus potentially facilitate improved search. The key question here is whether materials can be chosen with sufficient selectivity and reliability to give improved search.

[2]This of course assumes that the lecturer is not reading from a script!

4. INFORMATION RETRIEVAL AND FIELD COMBINATION

Assuming that annotations can be suitably selected, there is the further important question of how the ASR transcription might be combined with metadata fields to provide most effective search. Two methods are typically used to process documents with multiple fields in retrieval. The simplest approach is simply to merge all the data for a document into a single vector losing the document structure, and then perform standard information retrieval. The alternative is to perform separate retrieval runs for the individual search fields, and then form a sum of the resulting ranked lists to produce a single combined document list for output. In this latter method, often referred to as *data fusion* the lists may be weighted prior to merging.

In this section we examine these methods in more detail in the context of the BM25 retrieval model [7] based on the review of this topic and proposed a simple multi-field extension model (BM25F) appearing in [8].

BM25 is a very successful weighting scheme based on the probabilistic model of information retrieval. The model was developed for standard single field documents such as those used in early TREC ad hoc search tasks. The standard model does not allow for exploitation of the structure of multi-field documents. However, as illustrated later, this approach can lead to problems in term weighting when we attempt to take account of the field structure in multi-field documents, due to the nonlinear treatment of within document term frequency ($tf(i,j)$) in the BM25 function.

4.1 The Problem

Consider an unstructured document j belonging to a collection J, where j can be regarded as a vector
$j = \{tf(1,j), tf(2,j), \ldots, tf(V,j)\}$ where $tf(i,j)$ is the *term*

frequency of the i term in j, and V is the total vocabulary. Documents can be scored against a query using a ranking function such as BM25, where BM25 is defined as follows,

$$cw(i,j) = \frac{tf(i,j) \times (k_1 + 1)}{k_1((1-b) + b \times ndl(j)) + tf(i,j)} cfw(i),$$

where

$$cfw(i) = \log \frac{N - n(i) + 0.5}{n(i) + 0.5}, \quad (1)$$

$cw(i,j)$ is the combined term weight of i in j, N = total number of documents in the collection, $n(i)$ = number of documents in the collection containing term i, $cfw(i)$ = collection frequency weight, $ndl(j) = dl(j)/\text{ave } dl$ = normalised document length, $dl(j)$ = length of document j, ave dl = average document length across the collection, and k_1 and b are scalar parameters. The standard document matching score $ms(j,q,J)$ is computed by summing the $cw(i,j)$ of terms matching a query q also represented by a vector and assumed to be unweighted $q = \{q(1), q(2), \ldots, q(V)\}$.

Consider a collection with a set of field types $T = \{1, \ldots, f, \ldots, K\}$, e.g. $f = 1$ ASR transcription, $f = 2$ assigned keywords, etc, and assum that the fields are non-repeatable and non-hierarchical.

A structured document $\mathbf{j}$ can be written as a vector of fields: $\mathbf{j} = \{j[1], j[2], \ldots, j[k], \ldots, j[K]\}$. Each $j[k]$ can be seen as a vector of term frequencies $(tf(i,j[k]))_{i=1,\ldots,V}$ similar to a standard unstructured document. $\mathbf{j}$ is thus a matrix, note any field may be empty for an individual document. Let $\mathbf{J}$ refer to the collection of structured documents. In order to weight the fields differently, define the field weight vector of each document as $\mathbf{v} \in R^K$. Without loss of generality, set one field weight, e.g. the ASR transcription, equal to 1.

When scoring a structured document for query q we want to take account of the document contents and the collection, but also the field structure and the relative weight vector $\mathbf{v}$. The problem is therefore how to extend a standard ranking function $ms(j,q,J)$ into a new function $ms(\mathbf{j},q,\mathbf{J},\mathbf{v})$. The extension model proposed in [8] basically assumes that similar words appear in different fields, although probably with different distributions.

Most modern term weighting functions, including BM25, have a nonlinear $tf(i,j)$ component. This is desirable since the information gained on observing a term the first time in a document is greater than that of each subsequent occurrence. In BM25 the term frequency saturates after a few occurrences, which is fine for simple single field short documents, such as published new stories, for which it was originally developed, but may not be so for more complex "structured" documents. The rate at which the saturation point is reached is controlled by the k_1 factor, and this needs special consideration for such documents.

The simple linear summation of scores across multiple fields breaks the nonlinear $tf(i,j)$ relation. For example, for a query term in a document with metadata ASR $tf(i,j[2]) = 2$ and $tf(i,j[1]) = 1$. For a standard unstructured document these will be combined to give an overall $tf(i,j) = 3$ in a single BM25 combined weight for this term i in document j. If we weight the metadata $v[f] = 2$ and the ASR $v[f] = 1$. This should boost the weight of this term somewhat overall in the matching score of the document, but not in a simple linear fashion. The linear combination of scores in simple data fusion would give a rather higher value than this, equivalent to an effective $tf(i,j)$ contribution of $2 \times f(BM25_{metadata}(tf(i,j[1]) = 1) + f(BM25_{ASR}(tf(i,j[2]) = 2)$, i.e. almost double the expected BM25 $tf(i,j)$ function value for a single field document. This would mean that a document matching a single query term over several fields could score much higher than a document matching several terms in one field only.

4.2 Developing a Solution

If all the field weights v_f are set to 1, it is reasonable that the document and retrieval result should revert to the unstructured case (equivalent to merging all the fields). However, this is not the case with a non-linear tf function with linear summation of the field scores, i.e.

$$ms(j,q,J) \neq \sum_f ms(j[f],q,J)$$

Instead, we get a score that is very hard to interpret and no longer satisfies the properties of the original ranking function. In this case, setting weights becomes a hard problem.

BM25 requires the two parameters k_1 and b to be tuned for each collection to which it is applied. k_1 controls the nonlinear $tf(i,j)$ effect, b the effect of length normalization. The simple linear sum of scores method requires separate parameters to be set for each field. The values of a field weight vector $\mathbf{v}$ would also have to be set empirically, $K-1$, since one field can be set to 1. Thus for BM25 the total number of tuning parameters to be set is $2K + (K-1) = 3K - 1$.

The method proposed in [8] is based on weighting term frequency combination at indexing time. In doing this it seeks to modify standard ranking functions to exploit multiple weighted fields, while satisfying the following requirements:

- **preserve term frequency non-linearity** which has been shown repeatedly to improve retrieval performance.
- **give a simple interpretation** to collection statistics and to document length incorporating field weights.
- **revert to the unstructured case** when field weights are set to 1.

The method combines the term frequencies of the different fields by forming a linear combination weighted by the corresponding field weights,

$$\mathbf{j}' = \sum_{f=1}^{K} v_f . j[f]$$

and $\mathbf{J}'$ is a new collection of documents. Note that $\mathbf{j}'$ and $\mathbf{J}'$ are both dependent on the values in the field weight vector $\mathbf{v}$.

Documents are then scored using the resulting term frequencies,

$$ms_2(\mathbf{j}, q, \mathbf{J}, \mathbf{v}) = ms(\mathbf{j}', q, \mathbf{J}')$$

In this scenario the term weighting and scoring functions are applied only once to each document.

From the earlier example, combining the term frequencies and field weights would give $2 + 2 \times 1 = 4$, resulting in a slight boost to the weight of the term in each field, while term dependence is maintained. The resulting boost is sufficiently small that matching several terms remains more significant than matching the same individual term in several fields. This is equivalent to mapping the structured document collection into a new unstructured collection with modified term frequencies.

Although developed for BM25, this method is generally applicable for different ranking functions for non-structured documents. However, the benefits of using it may vary for different functions.

A few issues of interpretation need to be considered in the case of the extended multi-field BM25 model.

Document Length. There are various different ways of counting the document length. The simplest is to count the number of words in the document, considering only those words that are indexed. Thus the length of the document is the sum of the term frequencies. This definition applies naturally to the modified documents of J': the modified term frequencies are simply summed.

k_1 *and* b. Since the merging method substantially changes the $tf(i, j)$ values, it can also be expected to change the optimal value of k_1. [8] proposes a method for estimatiing k_1 and b based on values derived empirically for an unweighted merged collection. However, in experiments we found this approach to be unreliable and instead set them empirically for the each modified weighted collection itself.

5. CLEF CL-SR TEST SET

This section summaries the design and features of the CLEF CL-SR test collections, further detail is contained in the original track report [9]. The collection is based on digitized interviews with Holocaust survivors, witnesses and rescuers made by the Survivors of the Shoah Visual History Foundation (VHF). A very large collection (116,000 hours) of interviews was collected. One 10,000 hour subset of this collection was extensively annotated. A project funded by the U.S. National Science Foundation focused on Multilingual Access to Large Spoken Archives (MALACH) has produced ASR systems for this collection to foster research on access to spontaneous conversational speech [2].

5.1 Document Test Set and Related Metedata

The objective of a ranked retrieval system is to sort a set of "documents" in decreasing order likelihood of relevance. This makes the implicit assumption that clearly defined document boundaries exist. The nature of oral history interviews means that document boundaries are less clearly defined. The average VHF interview lasts more than 2 hours. It is not realistic to browse spoken units of this size spoken. Therefore it is more useful to retrieve relevant passages rather than entire interviews. The annotated 10,000 hour subset of the VHF collection is provided manually segmented by subject matter experts into topically coherent segments. Segments from these recordings were selected as the "documents" for the CLEF 2005 and CLEF 2006 CL-SR evaluations.

The document set used for the CLEF evaluations was selected as follows. Roughly 10% of the dataset, comprising 403 interviews (totaling roughly 1,000 hours of English speech) were selected. Of these interviews, portions of 272 were digitized and processed by two ASR systems for the CLEF 2005 CL-SR test collection. A total of 183 of these are complete interviews; for the other 89 interviews ASR results were available for at least one, but not all, of the 30-minute tapes on which the interviews were originally recorded. Finally, some further sections involving brief discussion of visual objects were eliminated from the collection. The resulting test collection comprised 8,104 segments from 272 interviews totaling 589 hours of speech. Thus each segment ("document") has an average duration of about 4 minutes (503 words) of recognized speech. A collection of this size is very small from the perspective of contemporary text information retrieval experiments, such as those as TREC, but is comparable to the 550 hour broadcast news collection used in the TREC 8 and TREC 9 SDR evaluations [3]. For the retrieval evalation each segment was uniquely identified by a DOCNO based on the recording from which it was taken.

For each segment a number of fields, including the ASR transcriptions, were created by VHF subject matter experts while viewing the interviews. The following fields were included in the test collection:

- NAME: contains the names of persons other than the interviewee that are mentioned in the segment.
- MANUALKEYWORDS: The MKW field contains thesaurus descriptors selected manually from a large thesaurus that was constructed by VHF. Two types of keywords are present, but not distinguished: (1) keywords that express a subject or concept; and (2) keywords that express a location, often combined with time in one pre-coordinated keyword. On average about 5 manually thesaurus descriptors were manually assigned to each segment, at least one of which was typically a pre-coordinated location-time pair (usually with one-year granularity)
- SUMMARY: contains a three-sentence summary in which a subject matter expert used free text in a structured style to address the following questions: who? what? when? where?

The following fields were generated fully automatically by systems that did not have access to the manually assigned metadata for any interview in the test collection. These fields can therefore be used to explore the potential of different techniques for automated processing:

- ASRTEXT fields contain words produced by an ASR system. The speech was automatically transcribed by ASR systems developed at the IBM T. J. Watson Research Center. For CLEF 2005, two ASR transcriptions were generated. The ASRTEXT2004A field contains a transcription using the best available ASR system, for which an overall mean word error rate (WER) of 38% and a mean named entity error (NEER) rate of 32% was computed over portions of 15 held-out interviews. The recognizer vocabulary for this system was primed on an interview-specific basis with person names, locations, organization names and country names mentioned in an extensive pre-interview questionnaire. The ASRTEXT2003A field contains a transcription generated using an earlier system for which a mean WER of 40% and a mean NEER of 66% was computed using the same held-out data. The ASRTEXT2006A ASR field was created for CLEF 2006 with mean word error rate of 25%. This was not available for all segments, where the ASRTEXT2004A field was inserted instead to form the ASRTEXT2006B field, further details are contained in [6].

- Two AUTOKEYWORD fields contain thesaurus descriptors, automatically assigned by using text classification techniques. The AUTOKEYWORD2004A1 (AKW1) field contains a set of thesaurus keywords that were assigned automatically using a k-Nearest Neighbor (kNN) classifier using only words from the ASRTEXT2004A field of the segment; the top 20 keywords are included. The classifier was trained using data (manually assigned thesaurus keywords and manually written segment summaries) from segments that are not contained in the CL-SR test collection. The AUTOKEYWORD2004A2 (AKW2) field contains a set of thesaurus keywords that were assigned in a manner similar to those in the AKW1, but using a different kNN classifier that was trained (fairly) on different data; the top 16 concept keywords and the top 4 location-time pairs (i.e., the place names mentioned and associated dates) were included for each segment.

5.2 Topics and Relevance Assessment

For the CLEF 2005 CL-SR task, a total of 75 requests felt to be representative of the form and subjects real search requests were selected from those created by users of the VHF collection. These were formed into standard TREC style topic statements consisting of a title, a short description and a narrative. Only topics for which relevant segments exist can be used as a basis for comparing the effectiveness of ranked retrieval systems. The developers sought to choose a set of topics and interviews for which the number of relevant segments was likely to be sufficient to yield reasonably stable estimates of mean average precision (30 relevant segments was chosen as the target, but considerable variation was allowed). A total of 12 topics were excluded, 6 because the number of relevant documents turned out to be too small to permit stable estimates of mean average precision (fewer than 5) or so large (over 50% of the total number of judgments) that the exhaustiveness of the search-guided assessment process used was open to question. The remaining 6 topics were excluded because relevance judgments were not ready in time for release as training topics and they were not needed to complete the set of 25 evaluation topics. The 63 topics developed in CLEF 2005 were thus available as a training set for CLEF 2006. 30 additional topics were created for the CLEF 2006 task. These were combined with 12 topics developed in 2005, but for which relevance data was not released, to form a test topic set of 42 topics. Following analysis of the results of participants submission 33 topics from the 42 topic released as the test set were selected as the 2006 evaluation set. Full details of the topics and relevance assessment procedures adopted are given in [9] and [6].

6. EXPERIMENTAL INVESTIGATION

In this section we give experimental retrieval results for the individual metadata fields of the CLEF CL-SR task and give some analysis of these results, and then report results for experiments combining ASR transcriptions and metadata fields. The basis of our experimental system is the City University research distribution version of the Okapi system [7]. The documents and search topics are processed to remove stopwords from a standard list of about 260 words, suffix stripped using the Okapi implementation of Porter stemming and terms are indexed using a small standard set of synonyms. None of the indexing procedures were adapted for the CLEF CL-SR test collections. All experiments are for the 63 English language training topics from CLEF 2006 using the combined TD topic fields[3]. $k1$ and b were tuned empirically for each experiment. Standard Okapi pseudo relevance feedback (PRF) [7] is used in all cases with an empirically determined upweighting of the original topic terms in each case. Results here thus represent an upper bound on expected performance for this system. The following metrics are shown: Recall in terms of total number of relevant documents retrieved for topics, standard TREC mean average precision (MAP), and precision at rank cutoffs of 5, 10 and 30.

6.1 Individual Field Retrieval Runs

Table 1: Retrieval results for individual document fields with CLEF 2005 CL-SR test topics.

	Recall	MAP	P5	P10	P30
MKW	2274	0.225	0.444	0.381	0.296
Summary	2157	0.234	0.422	0.384	0.285
ASR2006B	1488	0.071	0.215	0.200	0.131
AKW1	1451	0.047	0.149	0.138	0.106
AKW2	625	0.039	0.102	0.094	0.064

Table 1 shows retrieval results for individual fields[4]. Looking at these results for individual fields we can observe a number of interesting points. The good result for the Summary field is perhaps not surprising since these descriptions are constructed manually by domain experts. However, the result for MKW is only slightly lower. Our indexed MKW fields had an average of about 22 terms, similar to the number of terms in each of the AKW fields. This indicates that if a set of keywords related to the specific contents of a document can be assigned, then useful retrieval performance can

[3]The CLEF 2006 test set was not used since it is in use as test data in CLEF 2007

[4]No result is shown for the Name field since it is empty for many documents

be achieved without the need for extensive manual descriptions. Retrieval performance based on the ASR and AKW1 and AKW2 fields is much lower. Without access to full accurate transcriptions of the speech and AKWs assigned based on such transcriptions, it is not clear to what extent poor retrieval performance is due only to errors in the ASR transcriptions, and consequentially the assigned keywords. Or the extent to which the failure of important words to be articulated in the speech at all, means that even with perfect transcription relevant documents cannot be reliably retrieved at high ranks. However, even without this information we can perform some interesting analysis of spoken transcriptions in relation to indexing and search.

Table 2: Term occurrence statistics for TREC 8 and TREC 9 SDR Text and Speech collections.

	Text	Speech
No. of Unique Terms	78611	23316
Terms $n(i) = 1$	46626	4444
Terms $n(i) > 1$	31985	18872

One interesting feature to consider is the coverage of the vocabulary appearing in spoken documents against the vocabulary of the ASR system. There is no ground truth of the contents of CLEF CL-SR collections. However, we performed an analysis of the vocabulary of the spoken document collection used for the TREC-8 SDR task [3]. This data set comprises around 22,000 broadcast news documents. A baseline ASR transcription is provided along with a rough manual transcription of the data. The results of this analysis are shown in Table 2. It can be seen that the total number of unique terms appearing in the ASR transcription is about one third of those appearing in a manual transcription, while the vocabulary of the manual transcriptions is somewhat inflated by the presence of typos which will not be present in the ASR transcription, there is a clear trend. While the frequency of many of these additional terms in the manual transcription is very low, the data associated with speech segments must be mapped to within-vocabulary words, and evidence suggests that this set of words is drawn from a subset of the recognition vocabulary of the ASR system. This means that the frequency of recognised words will be higher in the ASR transcriptions than in accurate transcriptions. While the OOV rate is overall probably less than 10% for the ASR system in this news domain, a great many rare words are missing from the transcription, either because they are outside the vocabulary, or because the ASR system is "reluctant" to use them, possibly because of problems in statistical estimation in the language model associated with rare words in the training set. Whatever the reason for this, their absence from the transcription means they are not available for search.

The BYBLOS recognition system used to generate this transcription is quite well suited to the data to be recognised. Given the training of the ASR used to generate the CL-SR transcriptions described in Section 5.1, while the TREC SDR corpus is read rather than spontaneous speech, a similar trend is likely to occur for the ASR output of spontaneous speech in terms of vocabulary coverage in terms of vocabulary coverage.

Table 3: Average $cfw(i)$ and topic coverage values for CLEF 2006 CL-SR Title field. Total no of non-stopword terms = 74.

Field	Mean	Std Dev	Terms Present
MKW	5.24	1.76	48
Summary	6.47	1.89	66
ASR2006B	5.35	1.57	66
AKW1	4.17	2.02	47
AKW2	4.23	2.64	39

Table 4: Results for combination of ASR2006B with various metadata fields.

		Recall	MAP	P5	P10	P30
+AKW1	Unwgt	1584	0.077	0.248	0.219	0.144
	Wgt	1641	0.086	0.254	0.237	0.156
+AKW2	Unwgt	1665	0.086	0.238	0.210	0.149
	Wgt	1663	0.088	0.244	0.211	0.140
+AKW1	Unwgt	1717	0.092	0.241	0.233	0.157
+AKW2	Wgt	1778	0.097	0.264	0.221	0.164
+MKW	Unwgt	2129	0.225	0.417	0.370	0.273
	Wgt	2334	0.255	0.432	0.419	0.313
+SUMM	Unwgt	2166	0.213	0.415	0.363	0.270
	Wgt	2252	0.242	0.454	0.405	0.292

Given that we expect many rare search terms will be missing from the ASR transcriptions, we might expect that the terms which do appear will have lower discriminative ability, i.e. lower than expected $cfw(i)$ values. Table 3 shows mean and standard deviation cfw(i) values for the document fields calculated for non-stopwords in the Title field of the CLEF 2006 CL-SR topics with non-zero $cfw(i)$ values. It can be seen that mean $cfw(i)$ values for the ASR fields are lower than for the Summary field, both of which have the same coverage of the search terms. The keyword fields have significant numbers of topic search terms missing. We can again see that the mean $cfw(i)$ value for manual MKW fields is higher than those for the AKW fields. While the differences in $cfw(i)$ values may not appear large, these actually correspond to very large variations in the numbers of document in which a search term appears. Thus we can see that manual fields have greater discrimination than the automatically generated ones.

6.2 Field Combination Experiments

We now report results for merging of the ASR transcription field with metadata fields as described in Section 4. Table 4 shows combination of the automatically generated ASR2006B field with AKW1, AKW2, AKW1 and AKW2, MKW and Summary fields. Two combination schemes are compared in this experiment: simple merging of the fields and weighted field merging using BM25F. The field weights for the weighted runs and BM25 parameters were based on training using the CLEF 2005 data sets. Fields weights are based on the relative average precisions for the individual fields on the training set. Weights were set as follows: ASR2006B *times* 2, AKW1 *times* 1, AKW2 *times* 1, MKW *times* 4, and Summary *times* 4. A number of observations

can be made about these results. Use of BM25F by weighting the fields improves retrieval performance with respect to all metrics in nearly all cases. Comparing with the results for the individual fields in Table 1 it can be seen that the weighted combination results are in all cases better than those of any one of the individual component fields. Similar comparison reveals simple merge of ASR2006A with either the MKW or Summary field reduces performance compared to the individual manual fields, while simple combination of the automated fields still produces an improvement in effectiveness compared to individual fields, albeit a small one than with the weighted combination. The improvement in effectiveness for MKW and Summary when using weighted combination with ASR2006B is interesting since it indicates that while the transcription is noisy, it is still able to contribute useful information does not appear in the manual fields. These results should be treated with some caution since the parameters have been optimised for the individual runs on the test collection. We will be conducting further evaluations of field combination scenarios as part of our participation in the CLEF 2007 CL-SR track, and it will be interesting to see whether the trends retrieval effectiveness observed in this paper in are preserved for a set of search topics for which the algorithms have not been tuned.

7. CONCLUSIONS

Searching spontaneous conversation speech is a challenging problem raising more significant research challenges than earlier work on retrieval from read speech news collections. This paper has explored some of these problems, and examined the potential utility of related metadata to spoken content to enhance search effectiveness. We then examined the issue of field combination in multi-field documents. Experimental results using the CLEF CL-SR data sets illustrate that combination of ASR transcripions with metadata fields can enhance retrieval effectiveness. Further work is required in examining data combination for search, if performance on unseen search topics is to be made reliable. Examination of $cfw(i)$ values for automatically and manually fields show that automatic fields have lower term specificity indicating that this is one of the reasons for poor document ranking using these features.

Overall the results indicate that spontaneous speech search can benefit from the use of high quality metadata. Generating manual metadata is time consuming and expensive, although as demonstrated in our experiments it can be much more effective that automatically generated material. A research challenge then is to improve the quality of automatically generated metadata. In some domains, such as education, useful metadata is often easily available and relatively simple to associate with spoken content, in other domains, automatically locating and assigning precise metadata to associate with spoken segments for search will prove very challenging.

8. ACKNOWLEDGEMENT

Work partially supported by European Community under the Information Society Technologies (IST) programme of the 6th FP for RTD - project MultiMATCH contract IST-033104. The authors are solely responsible for the content of this paper. It does not represent the opinion of the European Community, and the European Community is not responsible for any use that might be made of data appearing therein.

9. REFERENCES

[1] E. Brown, S. Srinivasan, A. Coden, D. Ponceleon, J. W. Cooper, and A. Amir. Towards speech as a knowledge resource. *IBM Systems Journal*, 40(4):985–1001, 2001.

[2] W. Byrne, D. Doermann, M. Franz, S. Gustman, J. Hajic, D. Oard, M. Picheny, J. Psutka, B. Ramabhadran, D. Soergel, T. Ward, and W.-J. Zhu. Automatic recognition of spontaneous speech for access to multilingual oral history archives. *IEEE Transactions on Speech and Audio Processing, Special Issue on Spontaneous Speech Processing*, 12(4):420–435, 2004.

[3] J. Garofolo, G. Auzanne, and E. Voorhees. The trec spoken document retrieval track: A success story. In *RIAO 2000 Conference: Content-Based Multimedia Information Access*, pages 1–20, Paris, France, 2000.

[4] G. J. F. Jones and R. J. Edens. Automated alignment and annotation of audio-visual presentations. In *ECDL '02: Proceedings of the 6th European Conference on Research and Advanced Technology for Digital Libraries*, pages 276–291, London, UK, 2002. Springer-Verlag.

[5] G. J. F. Jones, J. T. Foote, K. S. Jones, and S. J. Young. Retrieving spoken documents by combining multiple index sources. In *19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 96)*, pages 30–38, Zurich, Switzerland, 1996.

[6] D. W. Oard, J. Wang, G. J. F. Jones, R. W. White, P. Pecina, D. Soergel, X. Huang, and I. Shafran. Overview of the clef-2006 cross-language speech retrieval track. In *CLEF 2006: Workshop on Cross-Language Information Retrieval and Evaluation*, Alicante, Spain, 2007.

[7] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3. In *Third Text REtrieval Conference (TREC-3)*, pages 109–126.

[8] S. E. Robertson, H. Zaragoza, and M. Taylor. Simple BM25 extension to multiple weighted fields. In *13th ACM International Conference on Information and Knowledge Management*, pages 42–49, Washington D.C., U.S.A., 2004.

[9] R. W. White, D. W. Oard, G. J. F. Jones, D. Soergel, and X.Huang. Clef 2005 cross-language speech retrieval track overview. In *CLEF 2005:Workshop on Cross-Language Information Retrieval and Evaluation*, pages 744–759, Vienna, Austria, 2006.