THE THISL BROADCAST NEWS RETRIEVAL SYSTEM

Dave Abberley (1), David Kirby (2), Steve Renals (1) and Tony Robinson (3)

(1) University of Sheffield, Department of Computer Science, UK
(2) BBC, Research and Development, UK
(3) SoftSound, UK

ABSTRACT

This paper described the THISL spoken document retrieval system for British and North American Broadcast News. The system is based on the ABBOT large vocabulary speech recognizer, using a recurrent network acoustic model, and a probabilistic text retrieval system. We discuss the development of a realtime British English Broadcast News system, and its integration into a spoken document retrieval system. Detailed evaluation is performed using a similar North American Broadcast News system, to take advantage of the TREC SDR evaluation methodology. We report results on this evaluation, with particular reference to the effect of query expansion and of automatic segmentation algorithms.

1. INTRODUCTION

THISL is an ESPRIT Long Term Research project in the area of speech retrieval. It is concerned with the construction of a system which performs good recognition of broadcast speech from television and radio news programmes, from which it can produce multimedia indexing data. The principal objective of the project is to construct a spoken document retrieval system, suitable for a BBC newsroom application. Additionally, we have constructed systems based on North American broadcast news, and a French language version is being developed. In this paper we shall describe the development of both the British and American English systems. Although British English is the main target for our demonstrator, working in American English enables us to evaluate the system performance through the TREC spoken document retrieval track.

The THISL system uses the ABBOT large vocabulary continuous speech recognition (LVCSR) system [1] and well-understood probabilistic text retrieval techniques. Section 2 discusses the overall approach, with the collection of the application specific acoustic and textual data discussed in section 3. Section 4 outlines the ABBOT LVCSR system, with particular reference to the estimation of acoustic, pronunciation and language models for British English broadcast news. Section 5 describes the text retrieval methods that we used, with particular attention to the use of query expansion and the development of automatic algorithms to segment streams of broadcast audio into "documents" suitable for text retrieval. The resulting system has been evaluated in the framework of the TREC-7 Spoken Document Retrieval (SDR) track, discussed in section 6. The overall implementation of the THISL system is described in section 7, including issues pertaining to the user interface.

This work was supported by ESPRIT Long Term Research Project THISL (23495).

2. APPROACH

There are two principal approaches to the task of spoken document retrieval. The *phone-based* approach processes the audio data with a lightweight speech recognizer to produce either a phone transcription or some kind of phone lattice. This data may then be directly indexed or used for word spotting. The *word-based* approach applies a complete large vocabulary speech recognition system to the audio track to produce a word-level transcription; at this point the problem may be treated as standard text retrieval (modulo speech recognizer errors).

In the THISL project we have adopted a word-based approach to spoken document retrieval, similar to that employed by several other groups (eg [2, 3]). This approach requires more computation than phone-based approaches, since a full large vocabulary decoding needs to be applied to the entire archive. However, it enables the constraints of the pronunciation dictionary and language model to be applied: text retrieval is more robust when applied to words rather than phone n-grams. Aside from computational considerations, the most frequently cited drawback of this approach is the problem of out-of-vocabulary words. We do not believe that this is a significant problem, and is certainly outweighed by the advantages of the word-based approach. Indeed, of the ad-hoc topics used in the past five TREC evaluations (TRECs 3–7), 9 out of 900 query words were out of vocabulary relative to the 65,000 word vocabulary used in the experiments reported in this paper. This 1% out-of-vocabulary rate corresponds with what is typically observed when recognizing broadcast news data.

3. DATA COLLECTION

To cover a reasonably wide range of conditions, speakers and topics, acoustic and textual data for training the British English version was gathered from a variety of BBC News and Current Affairs programmes. In to-

tal about 50 hours of recorded programmes were transcribed, the majority of which were from television and radio news bulletins but with about 15% from other programmes of a political or financial nature. Transcriptions were carefully checked to ensure they accurately represented the acoustics, as is standard practice. However, we departed from the normal practice of adding fine granularity timing information, say at the end of each sentence or speaker turn, as we found that this was particularly labour intensive. The timing of major changes in acoustic condition were noted but otherwise we only added synchronization marks every five minutes and we further developed our speech alignment software to take the coarse timing information and provide word and phone alignments.

Textual data was acquired from a wider range of sources although still centred on news. Access to the BBC News text database provided material from March 1997 onwards and this was again supplemented with material from related programmes. In total these sources provided about 6.4 million words.

4. SPEECH RECOGNITION USING ABBOT

We have used the ABBOT LVCSR system developed at the Universities of Cambridge and Sheffield [1] and further developed by SoftSound. ABBOT differs from most other state-of-the-art LVCSR systems in that it has an acoustic model based on connectionist networks [4]; in ABBOT, we typically use two recurrent networks trained on forward-in-time and backward-in-time data (PLP front-end). In this application we use a 64K word pronunciation dictionary, together with a trigram language model.

ABBOT has several characteristics that make it suitable for spoken document retrieval applications including realtime (or close to realtime) performance, decoders with low latency and a simple architecture. In particular, we have evaluated systems on broadcast news tasks using only context-independent connectionist acoustic models.

We outline here the development of ABBOT for British English broadcast news; the North American broadcast news system is described in [5].

Acoustic Models: Acoustic models were trained on most of the transcribed corpora. In order to reduce the manual effort in checking transcriptions we filtered the training data using a measure of the confidence that the alignment was in fact the true transcription. The confidence measure chosen was simply the average log probability of the labelled phone class, although there is scope for use of other measures [6].

Pronunciation Dictionary: We used a pronunciation dictionary similar to BEEP[1]. As an extension, we included common acronyms and case dependent entries. This facilitated the process of checking the dictionary for accurate pronunciations and allows us to conduct IR experiments comparing the two schemes.

[1] ftp://www-svr.eng.cam.ac.uk/pub/comp.speech/dictionaries/beep.tar.gz

Language Models: For the North American Broadcast News system, language model construction was straightforward, involving the estimation of n-gram language models from text data provided for ARPA/NIST evaluations. There is currently less processed data for the British English system. The trigram language models use some of the North American text data, together with British English newspaper and newswire data (about 4 million words from Sep–Dec 1998)), transcriptions and scripts from BBC news and current affairs output (about 6 million words from Mar 1997 – Sep 1998) and transcriptions from CNN output (about 8 million words from Sep–Dec 1998).

Search: The LVCSR search space is huge. In the ABBOT system we have adopted stack decoding search strategies embodied in the NOWAY [7] and CHRONOS [8] decoders. These search algorithms are able to make direct use of the posterior probability estimates produced by the neural network acoustic model by pruning all those phones which have an estimated local posterior probability below a threshold. We have further developed the CHRONOS decoder for this search task to achieve:

Real-time recognition Using a 450MHz Pentium-II running UNIX we average real-time decoding with a typical memory usage of under 256Mb. This is important for this task as our final system targets about 1000 hours of audio.

Whole show decoding The efficient memory usage of CHRONOS allows decoding of hour-long shows and so enables the use of online acoustic normalisation as an alternative to the more common per-segment normalisation techniques.

Confidence measures When decoding continuous audio the error rate varies by more than an order of magnitude. We have integrated the confidence measures of [6] to allow weighting of the Term Frequency component in text retrieval.

Cross sentence decoding In common with most implementations, our language model contains a special symbol, <s>, to indicate a sentence boundary. Giving this symbol an acoustic realisation of a short period of silence allows the decoder to hypothesise sentence boundaries, and so fit the desired functionality of multiple sentence decoding.

Speech Recognition Results:

Our primary objective is fast, efficient information retrieval. Speech recognition performance is weakly correlated with this goal and in this section we give the word error rate (WER) for various configurations of our system. In many cases we are prepared to accept an increase in WER in order to maximise the overall system performance.

Table 1 shows the WER of the system evaluated on two news broadcasts, the BBC Nine O'Clock news from 8 May 1998 and the BBC One O'Clock news from 9 February 1999. The baseline system was set up to run in real-time, it used the language model described above

with online acoustic normalisation instead of segmentation, used cross sentence decoding and employed confidence based training of the acoustic models. The baseline WER is higher than that reported for North American broadcast news system [5], in part because we decode complete broadcasts and score against single hypothesis transcriptions. The four times real-time system shows that we make at least 2% more errors in order to run at the speed we desire. Selecting the training data used, using a more appropriate language model and performing cross-sentence decoding all improved the error rate slightly.

System	WER
baseline system	36.6%
4x real-time	35.9%
all training data	37.1%
with North American LM	37.4%
without cross-sentence	36.9%

Table 1: Word error rates for variations on the baseline British English system.

5. TEXT RETRIEVAL

The information retrieval component of THISL is based on the bag-of-words probabilistic model. Each document — which is produced by a speech recognizer — is preprocessed using a stop list and the Porter stemming algorithm, and may be represented as a bag of processed terms. We use the Okapi term weighting function [9] to match a term t against a document d:

$$CW(t,d) = \frac{CFW(t) * TF(t,d) * (K+1)}{K((1-b)+b*NDL(d))+TF(t,d)}, \quad (1)$$

where $TF(t,d)$ is the frequency of term t in document d and $NDL(d)$ is the normalized document length of d

$$NDL(d) = \frac{DL(d)}{\overline{DL}}. \quad (2)$$

$DL(d)$ is the length of document d (ie the number of unstopped terms in d). $CFW(t)$ is a term that measures what proportion of the collection t appears, and is referred to as the collection frequency weight:

$$CFW(t) = \log\left(\frac{N}{N(t)}\right), \quad (3)$$

where N is the number of documents in the collection and $N(t)$ is the number of documents containing term t.

The parameters b and K in (1) control the influence of document length and term frequency in the weighting function. These are set empirically; for our spoken document retrieval work we typically use values such as $b = 0.5$ and $K = 1.0$.

A query is also represented as a bag of (stopped and stemmed) terms. The overall match between a document and a query is obtained by summing (1) over all terms in the query. The collection may then be ranked with respect to relevance to a particular query.

5.1. Query Expansion

Under the bag of words model, if a relevant document does not contain the terms that are in the query, then that document will not be retrieved. The aim of query expansion is to reduce this query/document mismatch by expanding the query using words or phrases with a similar meaning or some other statistical relation to the set of relevant documents. This procedure may have even greater importance in spoken document retrieval, since the word mismatch problem is heightened by the presence of errors in the automatic transcription of spoken documents.

An obvious danger in using relevant documents retrieved from a database of automatically transcribed spoken documents is that the query expansion may include recognition errors (eg [2]). One way of avoiding this problem is by using a secondary corpus of documents from a similar domain that do not contain recognition errors. For our application an obvious choice for such a corpus is contemporaneous newswire or newspaper text. This secondary collection is ranked with respect to the query. A query expansion algorithm may then be applied using this ranking to find those terms in the secondary collection that have the largest mutual information (or related statistic) with the query terms.

In interactive systems, where there is a human in the loop, it is possible to definitely mark documents as relevant or non-relevant, and such documents can be used as training data for a *relevance feedback* query expansion approach. In the purely automatic case, in which no relevance judgements are available, it is assumed that the top nr documents are relevant. This process is sometimes termed *pseudo-relevance feedback*. We have used such an algorithm, based on the local context analysis algorithm of Xu and Croft [10]. In this algorithm, the query expansion weight for a term given a query and the secondary collection is based on the nr top ranked documents in the secondary collection:

$$QEW(Q,e) = \sum_{t \in Q} \log\left(\frac{\log(AF(e,t)) * CFW(e)}{\log(nr)} + \delta\right) * CFW(t). \quad (4)$$

The potential query expansion terms e are simply those terms in the relevant documents. The term $AF(e,t)$ measures the term frequency correlation of two terms e and t across collection of documents d_i:

$$AF(e,t) = \sum_{i=1}^{nr} TF(e,d_i) * TF(t,d_i). \quad (5)$$

The nt possible expansion terms with the largest weights are then added to the original query, weighted as $1/rank$. Note that this algorithm is not discriminative, since it does consider (pseudo-)non-relevant documents.

In practice the values of nr and nt are maximum limits, since we threshold so that only those documents with a score greater than 0.8 times the score of the top-ranked document are considered, and only those terms

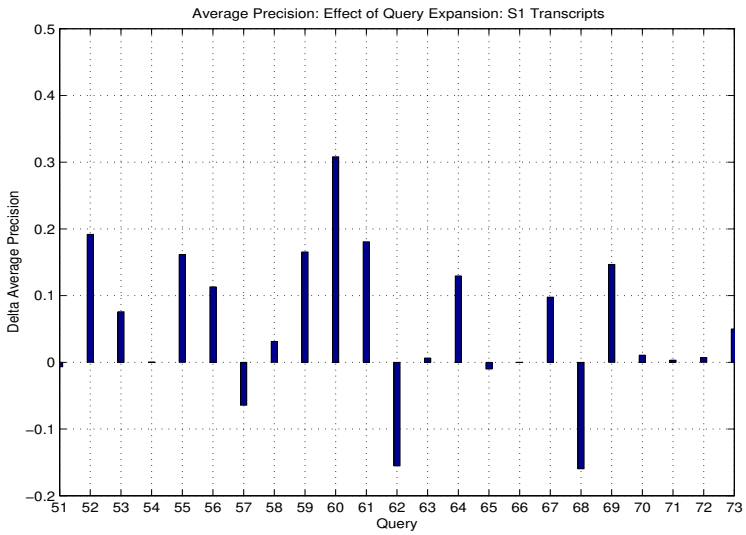

Figure 2: Query-by-query effect of query expansion in terms of change in average precision compared with no query expansion.

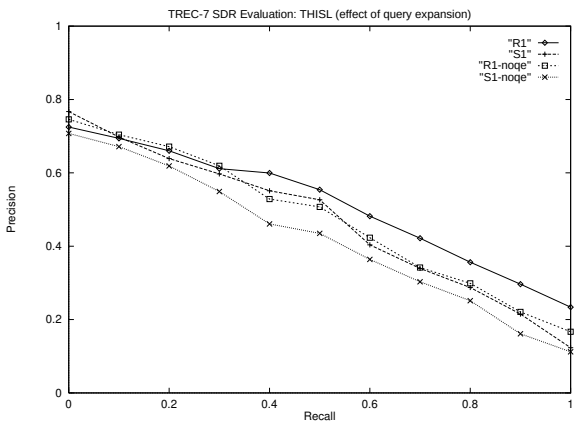

Figure 1: Effect of query expansion on recall-precision for TREC-7 SDR using reference transcriptions (R1) and the output of the ABBOT recognizer (S1).

with $QEW(Q,e)$ greater than an empirically-determined threshold are added.

We experimented with this query expansion algorithm on the TREC-7 SDR corpus. Figure 1 shows the effect on the interpolated recall-precision curve for the reference and speech recognition conditions; the query-by-query effect over the 23 queries in the TREC-7 SDR evaluation is shown in figure 2.

5.2. Segmentation

Speech rarely arrives with marked segment boundaries, as is well known to speech recognition researchers. Although controlled evaluations, such as TREC SDR, have included hand segmentation of news broadcasts into stories, this feature is typically not available for most applications. The corpus we have collected for the BBC application is recorded off air, and some segmentation is necessary to develop an SDR system.

To enable the objective evaluation of different automatic segmentations, we have used the TREC SDR corpus since relevance judgements are available. Since this is a segmented corpus, some adaptations were necessary to enable automatic segmentation experiments. To simulate the unsegmented condition all segmented stories were abutted and segment boundaries removed. This has the side effect of removing the "gaps" due to unrecognized material such as adverts and sports news. Automatically segmented documents may be characterised by a time index (eg of the segment mid-point); to enable the TREC relevance judgements to be used, these time indexes are converted to the original document IDs at evaluation time.

There has been a substantial amount of work in automatically segmenting documents for text retrieval. Callan [11] and Kaszkiel and Zobel [12] have investigated so-called *passage* retrieval in which documents are broken down into passages typically using document markup or windows of a fixed number of words. Algorithms that automatically segment documents into semantically separate topics have also been investigated recently [13, 14]. Benefits of passage-based retrieval include the retrieval of the most relevant portions of longer documents, the avoidance of document-length normalization problems and the possibility of more user-friendly interfaces that return the most relevant portion of a document. It has also been claimed that passage retrieval can improve average precision, since it returns short passages with the highest query word density. The principal problems with passage retrieval are the segmentation algorithm, and also the possibility of a dramatic increase in the number of "documents" (ie passages) in the collection.

The situation for spoken data is somewhat different to that for text. Without some kind of prosodic analysis any kind of "document markup" must be at a much coarser level. Also, the average topic length may be

much shorter in broadcast news, compared with many text documents.

We have investigated two straightforward approaches to automatic segmentation using windows based on time and number of words. In both cases we have used rectangular windows, of varying lengths and varying degrees of overlap. Initial experiments were carried out using the TREC-7 SDR system, without query expansion. In this case, our standard hand-segmented system resulted in an average precision of 0.4062. Figure 3 shows the average precision for varying window lengths and overlaps, using rectangular windows based on fixed time intervals (left) and fixed word lengths (right). The maximum average precision for both systems is similar, 0.3720 and 0.3757 respectively. This occurs with a relatively short window length (30s and 80 words respectively) and with an overlap of around 50%. The dependence of average precision on window length and overlap seems much smoother for the time-based window.

The above experiments were repeated using the best window parameters and query expansion from a LA Times/ Washington Post corpus contemporaneous with the Broadcast News data. The corpus was manually segmented into stories (documents). Each query was expanded with the top 10 terms from the top 8 documents as described in Section 5.1. Query expansion increased the average precision for the hand-segmented data to 0.4598, with similar improvements for the automatically segmented cases (see Document query expansion in Table 2).

The final experiment was to apply automatic segmentation to the query expansion corpus using an 80 word window with 50% overlap. This resulted in much shorter query expansion documents than before. Consequently the number of relevant documents for query expansion was increased from 8 to 50. This method worked well, with average precision rising dramatically to 0.5024. In addition, the average precision of the time-window segmented documents was almost as high as the best result for manual segmentation. The average precision of the word-window segmented documents did not improve, however (see Passage query expansion in Table 2).

6. EVALUATION

Direct evaluation of the THISL system for BBC news is difficult since the TREC methodology of pooled relevance assessments is difficult to implement for a single system, thus making accurate recall statistics difficult to obtain. Furthermore, since the BBC data has only been transcribed by a medium word error rate recognizer, large-scale relevance assessments are likely to be labour-intensive. A twofold strategy is possible: evaluation of the equivalent North American broadcast news system within TREC, and precision-oriented evaluation of the BBC system (eg using precision at 1,2,5,10). Precision results for the BBC system will be reported at the workshop; table 2 reports results from the TREC-7 SDR evaluation, including using automatic segmentation. Note that the parameters for the segmentation window were developed on the evaluation set. We note that automatic segmentation only results in a 10% relative reduction in average precision compared with the hand-segmentation. Also, using passage retrieval (with an 80 word window, with 50% overlap) on the secondary query expansion collection results in a small improvement in average precision, compared with using the document boundaries given in that collection. This is consistent with the results reported in [10].

Query Expansion	Segmentation	Average Precision
No	Manual	0.4062
No	Time	0.3720
No	Words	0.3757
Document	Manual	0.4598
Document	Time	0.4226
Document	Words	0.4254
Passage	Manual	0.5024
Passage	Time	0.4577
Passage	Words	0.4170

Table 2: Average precision for the TREC-7 SDR evaluation data. Conditions included no query expansion and query expansion using a contemporaneous newswire corpus (LA Times and Washington Post 1997–98) with either passage or document segmentation; and manual segmentation (provided by NIST/LDC) and automatic segmentation based on fixed rectangular time (30s, 60% overlap) or word (80 words, 50% overlap) windows.

7. DEMONSTRATION SYSTEM

The current THISL system for BBC news uses the speech recognition and information retrieval strategies discussed above. Query expansion is performed using a secondary collection derived from the British Press Association newswire, and we use a time-based rectangular window for automatic segmentation.

In addition to the standard keyboard interface for submitting queries, we are also experimenting with a spoken query interface. To assess the system in a practical situation, easy user access would be required from office environments. This dictated that access from general purpose desktop PCs would be essential and hence a Web browser would be the most cost-effective interface for keyboard initiated queries. Although this approach would not permit the spoken query interface to be assessed to the same extent, this latter aspect could be assessed in more controlled conditions at a dedicated workstation. Consequently, the THISL system itself is most conveniently implemented on a dedicated central system with the core functions of recording, speech recognition and indexing carried out automatically on at least a daily basis.

The size of the database, and the fact that it would be updated with new programmes each day, required care-

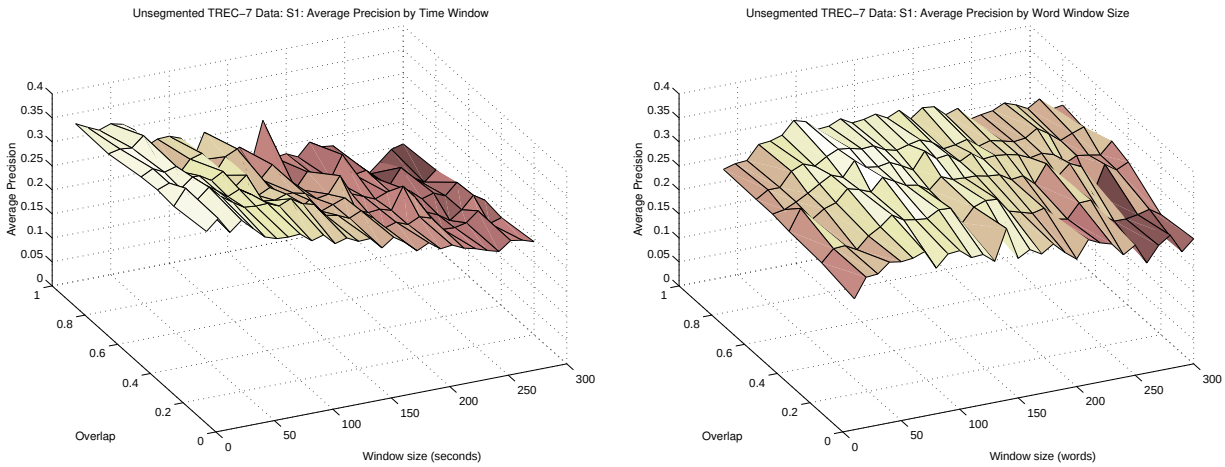

Figure 3: Effect on average precision of automatic segmentation window length.

ful consideration of the amount of data that could be processed and handled in practice. It was decided to build the main English language database by taking six main BBC News broadcasts each day: three each from television and radio channels. This amounts to about 2.5 hours of audio and, although by no means the full output from a newsroom, should cover all the major breaking stories each day. Over the course of the project, the resultant database size will be large enough to assess the effectiveness of speech retrieval but not too onerous to manage. The prototype demonstration at the workshop will be based on an archive of over 500 hours of BBC news output.

8. REFERENCES

[1] T. Robinson, M. Hochberg, and S. Renals, "The use of recurrent networks in continuous speech recognition," in *Automatic Speech and Speaker Recognition – Advanced Topics* (C. H. Lee, K. K. Paliwal, and F. K. Soong, eds.), ch. 10, pp. 233–258, Kluwer Academic Publishers, 1996.

[2] J. Allan, J. Callan, W. B. Croft, L. Ballesteros, D. Byrd, R. Swan, and J. Xu, "INQUERY does battle with TREC-6," in *Proc. Sixth Text Retrieval Conference (TREC-6)*, pp. 169–206, 1998.

[3] S. E. Johnson, P. Jourlin, G. L. Moore, K. Sparck Jones, and P. C. Woodland, "The Cambridge University Spoken Document Retrieval System," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 1999. to appear.

[4] H. Bourlard and N. Morgan, *Connectionist Speech Recognition—A Hybrid Approach*. Kluwer Academic, 1994.

[5] G. Cook, K. Al-Ghoneim, D. Ellis, E. Fosler-Lussier, Y. Gotoh, B. Kingsbury, N. Morgan, S. Renals, T. Robinson, and G. Williams, "The SPRACH system for the transcription of broadcast news," in *Proc. DARPA Broadcast News Workshop*, 1999. To appear.

[6] G. Williams. and S. Renals, "Confidence measures derived from an acceptor HMM," in *Proc. ICSLP*, (Sydney), pp. 831–834, 1998.

[7] S. Renals and M. Hochberg, "Start-synchronous search for large vocabulary continuous speech recognition," *IEEE Trans. Speech and Audio Processing*, in press.

[8] T. Robinson and J. Christie, "Time-first search for large vocabulary speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, (Seattle), 1998.

[9] S. E. Robertson and K. Sparck Jones, "Simple proven approaches to text retrieval," Tech. Rep. TR356, Cambridge University Computer Laboratory, 1997.

[10] J. Xu and W. B. Croft, "Query expansion using local and global document analysis," in *Proc. ACM SIGIR*, 1996.

[11] J. P. Callan, "Passage-level evidence in document retrieval," in *Proc. ACM SIGIR*, pp. 302–309, 1994.

[12] M. Kaszkiel and J. Zobel, "Passage retrieval revisted," in *Proc. ACM SIGIR*, 1997.

[13] M. A. Hearst, "TextTiling: Segmenting text into multi-paragraph sub-topic passages," *Computational Linguistics*, vol. 23, pp. 33–64, 1997.

[14] J. Yamron, I. Carp, L. Gillick, S. Lowe, and P. van Mulbregt, "A hidden Markov model approach to text segmentation and event tracking," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, (Seattle), 1998.