THE THISL SYSTEM FOR INDEXING AND RETRIEVAL OF BROADCAST NEWS

Steve Renals
Dave Abberley
Dept. of Computer Science
University of Sheffield
United Kingdom
{d.abberley,s.renals}@dcs.shef.ac.uk

David Kirby
BBC Research & Development
david.kirby@rd.bbc.co.uk

Tony Robinson
SoftSound
United Kingdom
tony.robinson@softsound.com

http://www.dcs.shef.ac.uk/research/groups/spandh/projects/thisl/

Abstract - This paper describes the THISL news retrieval system which maintains an archive of BBC radio and television news recordings. The system uses the ABBOT large vocabulary continuous speech recognition system to transcribe news broadcasts, and the thislIR text retrieval system to index and access the transcripts. Decoding and indexing is performed automatically, and the archive is updated with three hours of new material every day. A web-based interface to the retrieval system has been devised to facilitate access to the archive.

INTRODUCTION

THISL is an ESPRIT Long Term Research project concerned with developing key technologies for spoken document retrieval [1, 2]. A major goal of the project is to produce a prototype *news-on-demand* system suitable for a BBC newsroom application.

At the time of writing, the database consists of 750 hours of radio and television news material from more than 2000 different broadcasts covering the period from January 1998 to May 1999. Data collection is ongoing and the corpus is growing at the rate of about 3 hours per day.

The following sections describe the components of the THISL news retrieval system. The ABBOT large vocabulary continuous speech recognition system [3] is used to transcribe news broadcasts. The transcriptions are then segmented automatically and indexed by the thislIR text retrieval system [4]. thislIR can then use this index to produce a list of news clips in response to a query. thislIR can be accessed via a simple web-based interface.

THE ABBOT LVCSR SYSTEM

The ABBOT LVCSR system is a hybrid connectionist/HMM system which uses neural network acoustic models and a stack decoding search strategy. For this work, the CHRONOS decoder [5] has been developed to allow memory efficient decoding of whole broadcasts.

Acoustic Modelling

For the acoustic model, ABBOT employed two recurrent neural networks trained on forward-in-time and backward-in-time PLP feature vectors. A combined context-independent phone probability vector for each frame was produced by merging the output vectors of the individual networks.

The acoustic models were trained on about 50 hours of BBC radio and TV current affairs broadcasts. The programmes were transcribed manually but fine granularity timing information (at the end of each sentence or speaker turn, say) was not available as it proved too labour intensive to produce. Speech alignment software was developed to take the coarse timing information and provide the necessary word and phone alignments.

In order to reduce the manual effort in checking transcriptions, the training data was filtered 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].

Language Modelling

A trigram language model was used. The trigrams were estimated from scripts and transcriptions from BBC news and current affairs output (about 6 million words from March 1997 – September 1998) bulked out with US broadcast news text data (100 million words) provided for ARPA/NIST evaluations [7]. This obviously gives the language model a heavy bias towards US English, and it is hoped that this can be redressed in the near future by incorporating UK English newspaper and newswire data.

An extensible vocabulary system, in which the language model and dictionary are updated on a daily basis, is currently being investigated. This will reduce the problems caused by new out of vocabulary words which are likely to appear regularly in news broadcasts.

Decoding

The CHRONOS decoder [5] has been developed to allow the decoding of an entire show in real-time. Using a 450MHz Pentium-II running Solaris, it was possible to achieve real-time decoding with a typical memory usage of under 256Mb. Decoding speed becomes important when dealing with several hundred hours of audio data. The efficient memory usage of CHRONOS allows decoding of hour-long shows and so enabled the use of online acoustic normalisation as an alternative to the more common per-segment normalisation techniques.

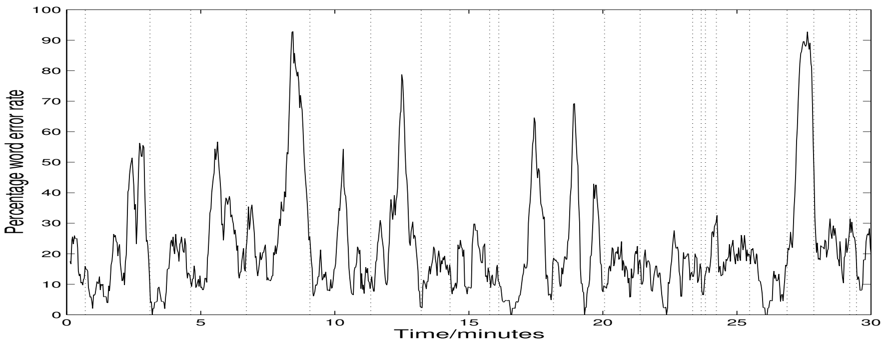

Figure 1: Word error rate over time for Radio 4 News of 10 Feb 1999.

Preliminary results indicate word error rates of about 35% for television news and 23% for radio news broadcasts. The difference in the two figures is because radio news relies much more on read speech in studio conditions which is easier to recognise. These results were obtained on six half-hour shows from a held-out subset of the training data. This is in line with results obtained on a similar system trained on North American Broadcast News for the TREC-7 evaluation [4].

Figure 1 plots the error rate throughout a show measured using a 15 second rectangular window. The dashed lines mark the story boundaries: note that as news topics are introduced by the newsreader the WER at the start of a topic is often relatively low. Also note that there is a very large variation in WER within a topic. This has implications for unsegmented information retrieval and related areas where it is desirable to concentrate on the sections where the speech recognition system is performing well.

THE thislIR TEXT RETRIEVAL SYSTEM

The thislIR text retrieval system is a 'textbook' information retrieval system based on the bag-of-words probabilistic model. Each document is preprocessed using a stop list and the Porter stemming algorithm, and may be represented as a bag of processed terms. The Okapi term weighting function [8] is used to match a term t against a document d:

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

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

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

$DL(d)$ 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:

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

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 determined empirically and in this work values of $b = 0.5$ and $K = 1.0$ have been used.

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.

Automatic Story Segmentation

Before indexing can take place, the decoded broadcasts have to be segmented into stories. To this end, automatic segmentation schemes have been investigated. Each broadcast was segmented into 'stories' comprising a fixed word count or time window. Differing degrees of window overlap were also tried. Experiments with the THISL TREC-7 system found that the best performance was obtained using 30 second time windows with a 12 second overlap which produced an average precision of 0.3720 compared with a value of 0.4062 for manually segmented data [9].

A side-effect of the automatic segmentation scheme is that adjacent overlapping segments are likely to produce similar scores. Consequently, the list of retrieved documents will contain many segments from the same news item. In an attempt to combine these story fragments, any overlapping segments occurring in the list of retrieved stories are combined into one, larger story.

INTERFACE

The demonstration system, THISLDemo, uses a web-based interface written as a Perl/CGI script. This has the advantage of making the system extremely portable as it can be run from a web-browser such as Netscape.

Figure 2 shows the results section of THISLDemo. The user enters a text query referring to a news item together with, optionally, restrictions on the news programme and time period from which the results will be presented. The upper part of Figure 2 shows the results produced by the thislIR server in response to the query. A list of shows and dates is produced together with information about the IR score and length of the clip retrieved. The user can then select one of these stories in order to view the transcript and play an audio file of the clip (lower part of Figure 2).

ACKNOWLEDGMENTS

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

References

[1] J. Hirschberg, S. Whittaker, D. Hindle, F. Pereira, and A. Singhal, "Finding information in audio: a new paradigm for audio browsing and retrieval," in *Proc. ESCA ETRW Workshop Acessing Information in Spoken Audio*, (Cambridge), pp. 117–122, 1999.

[2] H. D. Wactlar, T. Kanade, M. A. Smith, and S. M. Stevens, "Intelligent access to digital video: Informedia project," *IEEE Computer*, vol. 29, pp. 46–53, May 1996.

[3] 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.

[4] S. Renals, D. Abberley, G. Cook, and T. Robinson, "THISL spoken document retrieval at TREC-7," in *Proc. Seventh Text Retrieval Conference (TREC-7)*, 1999.

[5] 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.

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

[7] G. D. Cook and A. J. Robinson, "The 1997 Abbot system for the transcription of broadcast news," in *Proceedings of the 1998 Broadcast News Transcription and Understanding Workshop*, 1998.

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

[9] D. Abberley, D. Kirby, S. Renals, and T. Robinson, "The THISL broadcast news retrieval system," in *Proc. ESCA ETRW Workshop Acessing Information in Spoken Audio*, (Cambridge), pp. 14–19, 1999.

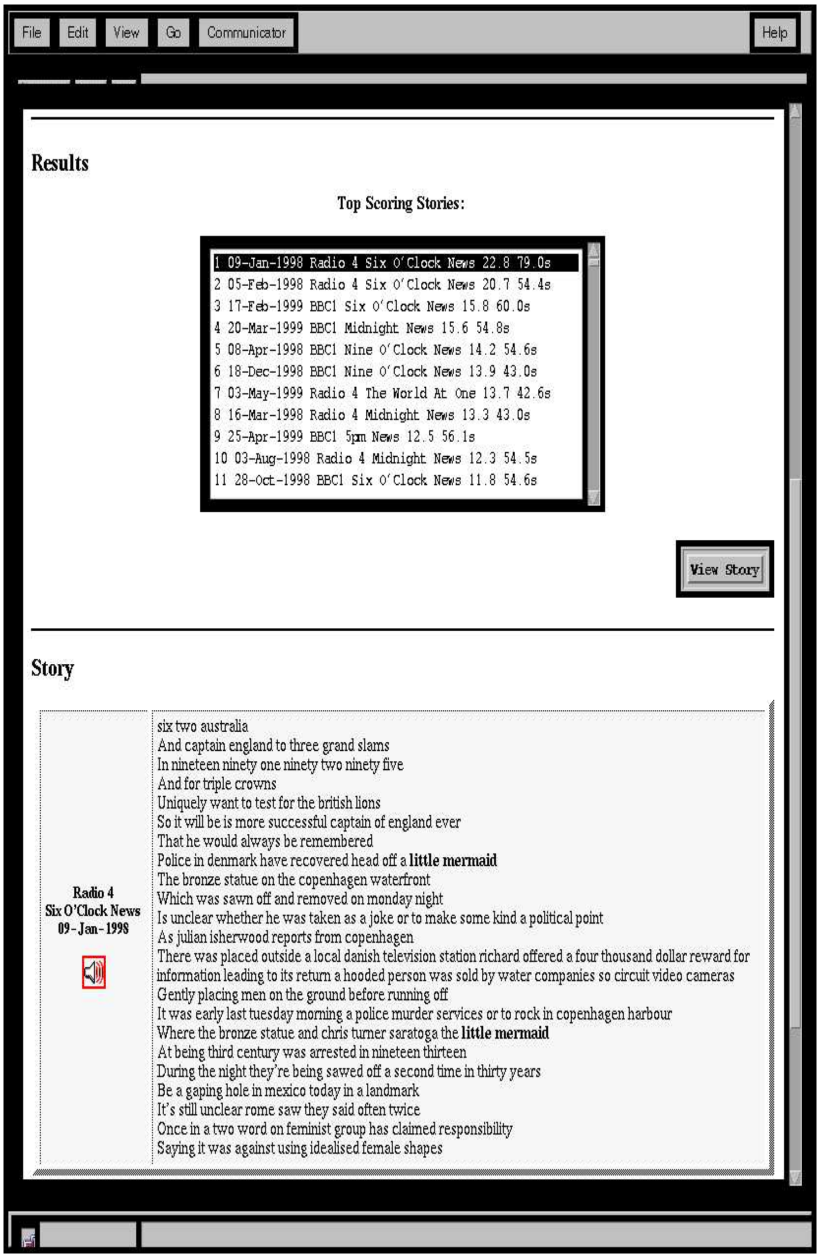

Figure 2: THISLDemo: response to the query "What happened to the little mermaid ?", showing the speech recognition transcription for the top-scoring story. The button on the left allows the corresponding audio clip to be played.