RETRIEVAL OF BROADCAST NEWS DOCUMENTS WITH THE THISL SYSTEM

Dave Abberley (1), Steve Renals (1) Gary Cook (2) and Tony Robinson (2,3)

(1) Department of Computer Science, University of Sheffield, UK
(2) Department of Engineering, University of Cambridge, UK
(3) SoftSound, UK
Email: {s.renals,d.abberley}@dcs.shef.ac.uk , {ajr,gdc}@eng.cam.ac.uk

ABSTRACT

This paper describes the THISL system that participated in the TREC-7 evaluation, Spoken Document Retrieval (SDR) Track, and presents the results obtained, together with some analysis. The THISL system is based on the ABBOT speech recognition system and the thislIR text retrieval system. In this evaluation we were concerned with investigating the suitability for SDR of a recognizer running at less than ten times realtime, the use of multiple transcriptions and word graphs, the effect of simple query expansion algorithms and the effect of varying standard IR parameters.

1. INTRODUCTION

THISL is an ESPRIT Long Term Research project that is investigating the development of a news-on-demand system using speech recognition, natural language processing and text retrieval. The main goal of the project is to develop a system, directed mainly toward UK English speech, for a BBC newsroom application; the TREC/SDR evaluation gives us a good opportunity to evaluate our current system on a closely related task.

The THISL spoken document retrieval system is based on the ABBOT large vocabulary continuous speech recognizer [1] and a probabilistic ranked text retrieval system. The large vocabulary speech recognizer is used to transcribe the broadcast audio, thus transforming the problem into one of text retrieval.

In this evaluation we were concerned with the following questions:

- Is a recognizer running in substantially less than ten times real-time suitable for spoken document retrieval?
- Can the use of multiple transcriptions or word graphs of documents be used to increase robustness and decrease the effect of recognition errors?
- Can query expansion be used to improve recall and precision?
- What is the effect of differing stop lists and application of stemming?

The system we used in this year's evaluation differs somewhat from the system we used in the TREC-6 SDR track [2]. In particular, we have replaced the PRISE text retrieval system with a locally implemented probabilistic system. We no longer use wordspotting to deal with out-of-vocabulary terms in queries, since experience has indicated that this is not a serious problem with speech recognizers that use a vocabulary of around 65,000 words (in this evaluation it turned out that there were three query words that were out-of-vocabulary with respect to our recognizer).

2. SPEECH RECOGNITION

2.1. ABBOT

ABBOT is a hybrid connectionist/HMM system [3] that differs from traditional HMMs in that the posterior probability of each phone given the acoustic data is directly estimated at each frame, rather than the likelihood of a phone (or state) model generating the data. Posterior probability estimation is performed by a connectionist network (or set of networks) trained to classify phones. In the ABBOT system, a recurrent network [4] is used. Direct estimation of the posterior probability distribution using a connectionist network is attractive since fewer parameters are required for the connectionist model (the posterior distribution is typically less complex than the likelihood) and connectionist architectures make very few assumptions on the form of the distribution. Additionally, this approach enables the use of an efficient search algorithm that uses a posterior probability-based pruning [5] and is able to provide useful acoustic confidence measures [6].

The speech recognition system used by the THISL group in the TREC-7 SDR track was a version of that used by the

This work was supported by ESPRIT Long Term Research Projects THISL (23495) and SPRACH (20077).

related CU-CON group in the 1997 ARPA CSR Hub 4 evaluation [7].

2.2. TRAINING

2.2.1. ACOUSTIC MODEL TRAINING

The acoustic model used in the THISL system consisted of two recurrent networks with 53 context-independent phone classes (plus silence). One network estimated the phone posterior probability distribution for each frame given a sequence of 12th order perceptual linear prediction features [8]. The other network performed the same distribution estimation with features presented in reverse order (since recurrent networks are time-asymmetric) and the two probability estimates were averaged in the log domain. Each network contained 384 state units, resulting in a total of about 350 000 acoustic model parameters, trained on the SDR acoustic training data. About 76 hours of the 100 hours of SDR acoustic training data is transcribed. After computing the average log likelihood per frame during a Viterbi alignment, a further 16 hours of this data was discarded as being below an empirically chosen log likelihood threshold, resulting in a transcribed set of acoustic training data of about 60 hours duration.

The final system used 697 context-dependent phone models, the acoustic context classes being arrived at via a decision tree algorithm. A context class network was used for each context-independent phone class, which (when combined with the context-independent phone probabilities) produced a context-dependent phone probability [9].

2.2.2. LANGUAGE MODEL TRAINING

A backed-off trigram language model was estimated from the following text sources:

- 1997 Hub4 LM text data (broadcast news transcriptions to 1996) (132M words);
- 1995 Hub4 non-financial newswire texts (108M words);
- 1995 Hub3 financial newswire texts (45M words);
- The transcripts of the SDR acoustic training data (0.8M words);
- 1995 Marketplace acoustic transcripts (0.05M words).

The 65,532 word vocabulary used all the words from the transcription of the acoustic training data, plus the most frequent remaining words extracted from the broadcast news text corpus (ignoring common misspellings and obvious text processing errors). The resultant language model contained 7.1 million bigrams and 24.0 million trigrams. We did not use the more recent SDR LM data for language modelling, although some of this data was used for query expansion (section 7).

2.3. RESULTS

Using the noway start-synchronous decoder [5] the system ran in about seven times real time on an Ultra-1/167MHz (512-1024 Mb RAM), with the computation split approximately equally between the recurrent network-based acoustic model and the LVCSR search algorithm. Running at this speed required more pruning than would be employed in a CSR evaluation, and the estimated relative search error resulting from incorrect pruning of the search space was 10–20%.

The overall average word error rate (WER) of the THISL speech recognition system in this evaluation was 35.9%. We can also use an error metric conditioned on the text retrieval system, the *term error rate* (TER) [10], which is given by the following formula:

$$TER = \frac{\sum_{t \in T} \mid R(t) - H(t) \mid}{|T|} \times 100\% \tag{1}$$

where $R(t)$ and $H(t)$ represent the number of occurrences of *term t* in the reference and hypothesised transcripts respectively. The set of terms T is calculated after the transcripts have been stopped and stemmed but without taking account of term order. Thus TER gives a more accurate measure than WER of the erroneous terms which will be processed during IR. Additionally, calculating WER is meaningless for merged transcripts (section 5), but TER still provides some information about transcript quality. In conjunction with our submitted system, using a 379 word stop list and Porter stemming the THISL speech recognition system returned a TER of 52.2%.

3. TEXT RETRIEVAL

In last year's SDR evaluation we used the PRISE text retrieval system developed by NIST. This year we used a locally implemented system. This was essentially a "textbook TREC system", using a stop list, the Porter stemming algorithm and the Okapi term weighting function. Specifically we used the term weighting function $CW(t,d)$ for a term t and a document d given in [11]:

$$CW(t,d) = \frac{CFW(t) * TF(t,d) * (K+1)}{K((1-b) + b * NDL(d)) + TF(t,d)}. \tag{2}$$

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

$$NDL(d) = \frac{DL(d)}{\overline{DL}}, \tag{3}$$

where $DL(d)$ is the length of document d (ie the number of unstopped terms in d). $CFW(t)$ is the collection frequency weight of term t and is defined as:

$$CFW(t) = \log\left(\frac{N}{N(t)}\right) \tag{4}$$

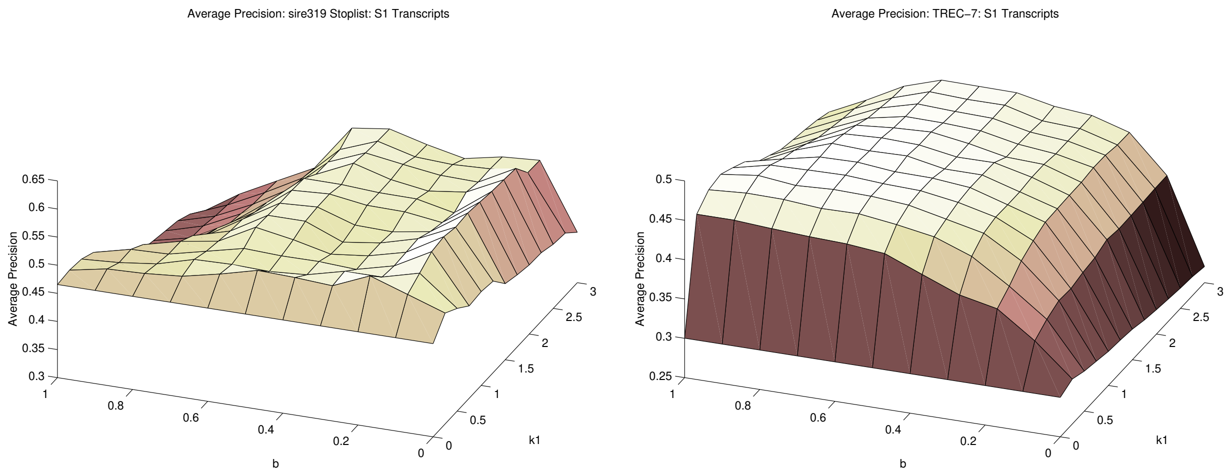

Figure 1: Plot of average precision against term weighting parameters b and K for TREC-7/SDR local development queries (left), and TREC-7/SDR evaluation queries (right).

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 (2) control the effect of document length and term frequency as usual.

Prior to the evaluation, we conducted a variety of experiments, using a development set of 16 queries devised and judged for relevance locally. These experiments were designed to investigate:

- the effect of varying term weighting parameters, stop lists and stemming (section 4)
- the use of multiple transcriptions arising from the component networks of the ABBOT acoustic model (section 5);
- the use of word graph representations of spoken documents (section 6);
- the behaviour of query expansion (section 7).

We note that these development queries were more similar to the TREC-6/SDR queries than the TREC-7/SDR queries, having an average of 4.5 relevant documents per query (found by manual operation of PRISE). This compares with the evaluation queries which had an average of 17 relevant documents per query.

4. TEXT RETRIEVAL PARAMETERS

4.1. TERM WEIGHTING PARAMETERS

Since the document collection and queries are a little different to the TREC ad-hoc task, we decided to investigate the effect of varying the parameters b and K in the term weighting function (2). The results for the development set are shown on the left of figure 1. After the evaluation we produced a similar graph for the TREC-7 SDR evaluation queries (figure 1, right). We note that for the development queries there is a ridge of high average precision along $K = 0.25$, which corresponds to a decrease in the significance of TF compared with CFW. This is not present in the evaluation queries. There is another a maximum around $(b,K) = (0.5,1.0)$, for both sets of queries, which (fortunately) were the parameter settings used for all our submitted runs.

The reason for the different behaviour of the two query sets is not clear. Although it may be due to the relatively small task size (around 3000 spoken documents), we also note that our local development queries had many fewer relevant documents per query compared with the evaluation queries (4.5 vs. 17). Support for the latter hypothesis is given by the fact that the parameter landscape for the known-item TREC-6/SDR queries (ie 1 relevant document per query) is most similar to the development set.

4.2. STOP LISTS

We conducted experiments using hand constructed stop lists including the 23 word stop list that is standard with PRISE, the 319 word stop list used by the University of Glasgow, the 429 word stop list in [12], and a locally developed 379 word stop list based on the Glasgow stop list with extra words added following analysis of previous TREC queries. As control experiments we used stop lists comprising the most frequent n words, and also no stop list. Results on our development set of queries are shown in figure 2, and a graph of term error rate vs. stop-list size is shown in fig-

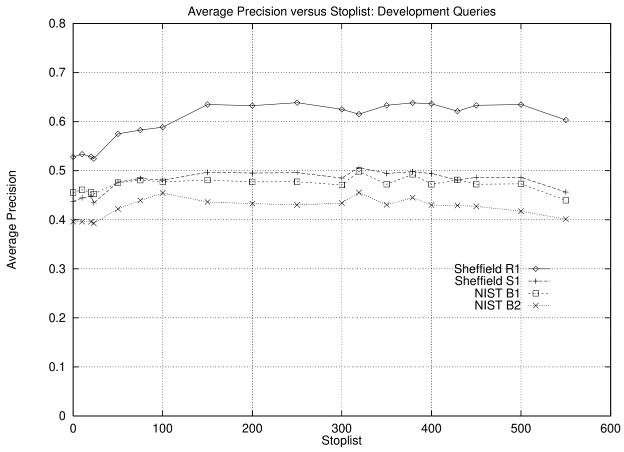

Figure 2: Effect of stop list on average precision using local TREC-7 development queries for R1, S1, B1 and B2 conditions, using Porter stemming. Stop lists of size 23, 319, 379 and 429 were hand-constructed the others comprised the most frequent *n* words.

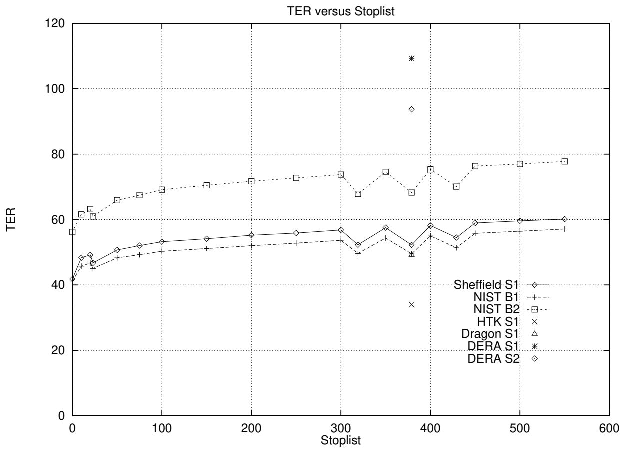

Figure 3: Effect of stop list on term error rate for S1, B1 and B2 recognizers, using Porter stemming. The hand constructed stop-lists of 319, 379 and 429 words can be clearly identified. TER for Dragon, CUHTK and DERASRU recognizers with the 379 word stop list are also shown.

ure 3. We note that hand-constructed stop lists perform a little better than the similarly sized "top-*n*" stop lists.

4.3. STEMMING

We also evaluated the effect of stemming by running after the evaluation with and without the Porter stemming algorithm. These results are shown in table 1.

	Average Precision	
System	With Stemming	Without Stemming
R1	0.4886	0.4179
S1	0.4599	0.3774
B1	0.4355	0.3570
B2	0.3529	0.2570

Table 1: Effect of stemming (Porter algorithm) on average precision to TREC-7 SDR queries (post-evaluation experiment). Experiments used a 379 word stop list and query expansion.

5. MULTIPLE TRANSCRIPTIONS

A number of participants at the TREC-6 SDR track (eg, [13, 14]) took advantage of the availability of multiple sets of speech recognition transcriptions and merged them to produce improved information retrieval performance. This method was successful because although speech recognizers make errors, different speech recognizers are likely to make different errors. Thus if an important query word has been missed by one recognizer, another one might recognize it correctly so that it does not get omitted from the index.

As mentioned in section 2, the ABBOT acoustic model is based on multiple recurrent networks, which are averaged together at the acoustic frame level. However, it is possible to run separate decodings based on the individual recurrent networks and to merge them together at the transcription level. Experiments were run on the TREC-6 known-item retrieval task using the 379 word stop list but no query expansion. Table 2 shows the results in terms of word error rate (WER), term error rate (TER) and the various TREC-6 IR performance measures.

The table indicates that merging the RNNs at the acoustic probability level (S1) produces better WER/TER and IR performance than either of the individual networks. Despite the inevitably higher TER, merging multiple transcripts seems to produce slightly better IR results than taking their union. The detrimental effects of merging may be partially offset by term frequency weighting. In these experiments, neither merging technique produced clearly better IR performance than the single best set of transcripts (S1), except for the percentage of queries for which the answer was not found.

The results from these experiments are somewhat inconclusive: it is possible that multiple transcripts could be used to enhance retrieval performance but these benefits have yet to be demonstrated unequivocally, and must be offset against the considerable extra resources required to produce the multiple transcriptions (which is why the experiments were not repeated on TREC-7 data).

Transcripts	WER	TER	Mean Rank	Mean Reciprocal	Percentage at Rank 1	Percentage Not Found
R1	–	–	5.85	0.8509	78.7%	0.0%
S1	38.8%	55.4%	11.72	0.7776	74.5%	2.1%
Forward net	43.2%	63.3%	14.33	0.6996	61.7%	2.1%
Backward net	41.7%	61.4%	17.96	0.7091	63.8%	4.3%
Merged fwd+bwd	–	135.9%	14.51	0.7414	68.1%	0.0%
Union fwd+bwd	–	90.3%	18.45	0.7477	68.1%	0.0%
Merged S1+fwd+bwd	–	228.5%	14.40	0.7793	72.3%	0.0%
Union S1+fwd+bwd	–	95.9%	19.77	0.7434	68.1%	0.0%

Table 2: Use of multiple transcriptions derived from ABBOT on the TREC-6 known-item retrieval task. R1 are the reference transcripts, S1 are the transcripts produced by ABBOT using frame-level merging. Forward and backward are the decodings produced by the nets in isolation. The term 'merged' implies the concatenation of two or more sets of transcripts whereas the term 'union' implies the union of sets of transcripts — multiple occurrences of the same term are discarded.

6. WORD GRAPHS

As a side effect of large vocabulary decoding, it is possible to produce word graphs (lattices). A word graph consists of set of nodes, each labelled with a reference time, and a set of links. Each link connects two nodes and is labelled with a word, together with information such as the acoustic score of that word accounting for the acoustic data that covers the time span between the nodes. Each link in the word graph corresponds to a word that was hypothesised during the search that could contribute to a complete word path within the graph. Thus a word graph efficiently represents the entire valid search space considered by the speech recognition decoder. On average, the word graphs produced by ABBOT contain about twenty times as many words as in the most probable transcription.

We treated a word graph as we would a single transcription in text retrieval, representing a document as a bag of word graph links. However since word graphs tend to be bushier in regions of acoustic confusion, the contribution of link i to the corresponding term frequency was based on the reciprocal of the graph density $(1/GD_i)$ for i. The graph density GD_i of graph link i is defined as the average number of links in the graph that account for each frame covered by link i.

The term frequencies that arise from representing documents as bags of graph links are less sharp than those that arise from 1-best transcriptions, since more terms are present in the document. Two ways to sharpen the term frequencies arising from graphs are by merging with the most probable transcription or by weighting the lattice links by an acoustic score. In this paper we have only tried the former.

We ran a number of experiments using word graphs on our development queries, using the Glasgow 319 word stop list. Results indicated that best performance resulted from parameter values $b = 0.6$ and $K = 2.0$. Recall and precision curves are shown in figure 4 (left). These results did not indicate that word graphs gave improvements in recall and precision, so we did not use them in our submitted evaluation system.

After the evaluation we ran the SDR evaluation queries against indexed word lattices, using the same parameters as before. Recall and precision curves are shown in figure 4 (right). The performance of the word graph representation is substantially worse than the one-best transcriptions. Since the evaluation queries were quite different to our development queries we reran a search in (b, K) space, but no other settings of these parameters were significantly better.

7. QUERY EXPANSION

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. Such a process may have increased 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. This was an experience reported by the INQUERY group in the TREC-6 SDR evaluation [15]. To avoid this problem we retrieved relevant documents from another collection of newswire text. The query expansion algorithm was then applied to the top n documents retrieved from that collection. The resulting expanded query was then applied to the collection of spoken documents.

We used an algorithm based on the local context analysis algorithm of Xu and Croft [16]. The initial query Q is

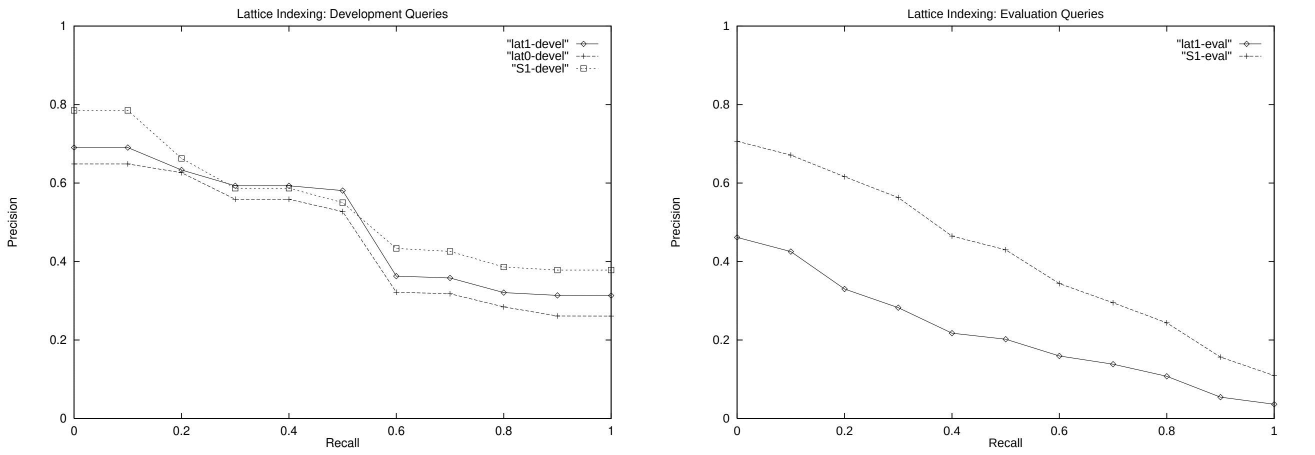

Figure 4: Recall-precision plot of SDR development queries (left) and evaluation queries (right) with documents represented as word graphs (lat0), most probable transcriptions (S1) and merged word graphs and transcriptions (lat1).

applied to the secondary query expansion collection. The nr top ranked documents are regarded as relevant; the algorithm is not discriminative so no non-relevant documents are required. A query expansion weight, $QEW(Q,e)$ is defined as follows:

$$QEW(Q,e) = \sum_{t \in Q} \log\left(\frac{\log(AF(e,t)) * CFW(e)}{\log(nr)} + \delta\right) * CFW(t). \quad (5)$$

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:

$$AF(e,t) = \sum_{i=1}^{nr} TF(e,d_i) * TF(t,d_i). \quad (6)$$

The nt possible expansion terms with the largest weights are then added to the original query, weighted as $1/rank$.

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

In the SDR evaluation we used the June 1997–February 1998 LA Times/Washington Post portion of the SDR LM text corpus as the query expansion database. This corpus contains about 13 million words and about 22,000 documents. The parameters nr and nt are clearly dependent on the size of the query expansion collection. Experiments to investigate the dependence on these parameters were carried out on our local development queries, and the results are shown in figure 5. From this we chose parameter values $(nr,nt) = (8,10)$. Figure 6 shows the performance of query expansion using a newswire corpus versus expanding on the target recognizer transcripts. Note that expanding on the recognizer transcripts is worse than no query expansion.

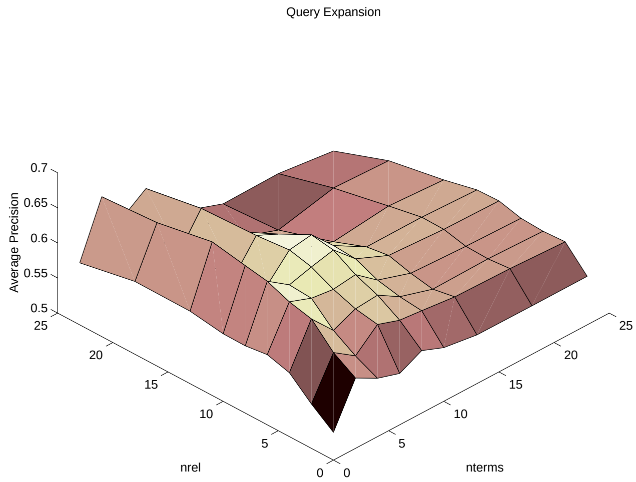

Figure 5: Effect of the query expansion parameters nr (maximum number of relevant documents to consider) and nt (maximum number of terms to add) on the average precision for S1, using 379 word stop list applied to our local development queries. The LA Times/Washington Post portion of the SDR language model corpus was used as the query expansion collection.

8. EVALUATION RESULTS

The text queries were preprocessed before being input to the system, to remove punctuation, convert to lower case and to

Condition	WER	TER	Retrieved	Relevant	Rel. Retrieved	AveP	R-P
R1	–	–	17613	390	364	0.4886	0.4583
S1	35.9%	52.2%	18312	390	360	0.4599	0.4485
B1	35.2%	49.5%	18093	390	355	0.4355	0.4562
B2	47.8%	68.3%	18671	390	354	0.3529	0.3347
CR-CUHTK	24.8%	34.0%	18105	390	365	0.4711	0.4469
CR-DERASRU-S1	66.2%	109.3%	17844	390	334	0.3780	0.4164
CR-DERASRU-S2	61.5%	93.7%	17973	390	344	0.4047	0.4016
CR-DRAGON-S1	29.8%	49.2%	18252	390	361	0.4613	0.4372

Table 4: Summary of results in different conditions. WER is word error rate, TER is term error rate (defined in section 2), AveP is the average precision and R-P is the R-precision.

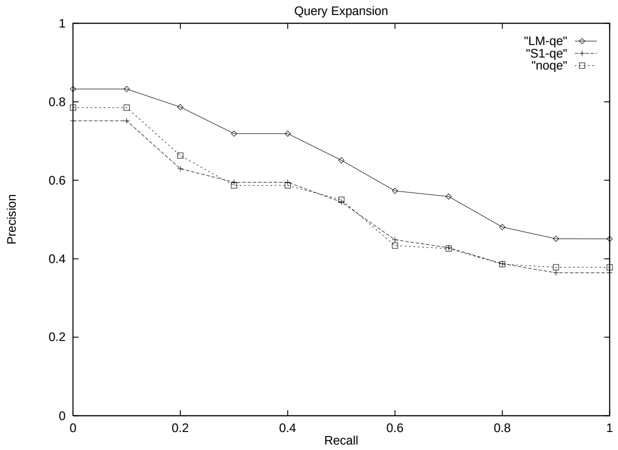

Figure 6: Effect of query expansion using newswire text (LM-qe) and recognizer transcripts (S1-qe) compared with no query expansion (noqe) on development queries.

expand abbreviation words to cover alternative transcription possibilities (eg, "AIDS" was expanded to "aids" and "a. i. d. s.").

No multiwords or phrases were used in the recognition or retrieval process. There were three OOV query words: *Montserrat*, *Trie* and *vs.* (versus). Our TREC-6 word spotting system [2] for OOV word restoration was not used in the TREC-7 system. Hopefully query expansion partially offset some of the problems caused by the OOV words.

Our submitted system used the 1-best transcriptions together with the query expansion algorithm outlined above. The 379 word stop list was used, together with the Porter stemming algorithm. The tunable parameters (set according to local development data) are given in table 3.

We ran using the following different transcripts:

R1 Reference transcripts (low WER)

B1 Medium error baseline transcripts (NIST running CMU Sphinx) (35% WER)

Parameter	Value
b	0.5
K	1.0
QE-b	0.5
QE-K	0.25
QE-nt	10
QE-nr	8

Table 3: Parameter settings for TREC/SDR submitted runs.

B2 High error baseline transcripts (NIST running CMU Sphinx) (49% WER)

S1 THISL speech recognition (36% WER)

CR-CUHTK Cambridge University (HTK) speech recognition (25% WER)

CR-DERASRU-S1 DERA/SRU speech recognition (66% WER)

CR-DERASRU-S2 DERA/SRU speech recognition (61% WER)

CR-DRAGON-S1 Dragon speech recognition (30% WER)

The results are summarized in table 4. The recall-precision curves resulting from these runs are shown in figure 7. Figure 8 shows the effect of query expansion on recall and precision for the R1 and S1 conditions. Results for the other speech recognizers are not shown to avoid cluttering the graph, but the effect of query expansion follows a similar trend for those.

Figure 9 shows the relative change due to query expansion for each of the twenty-three queries. As can be seen, query expansion resulted in an improvement or no significant change in average precision for most queries. An example of a query for which the query expansion algorithm proved effective:

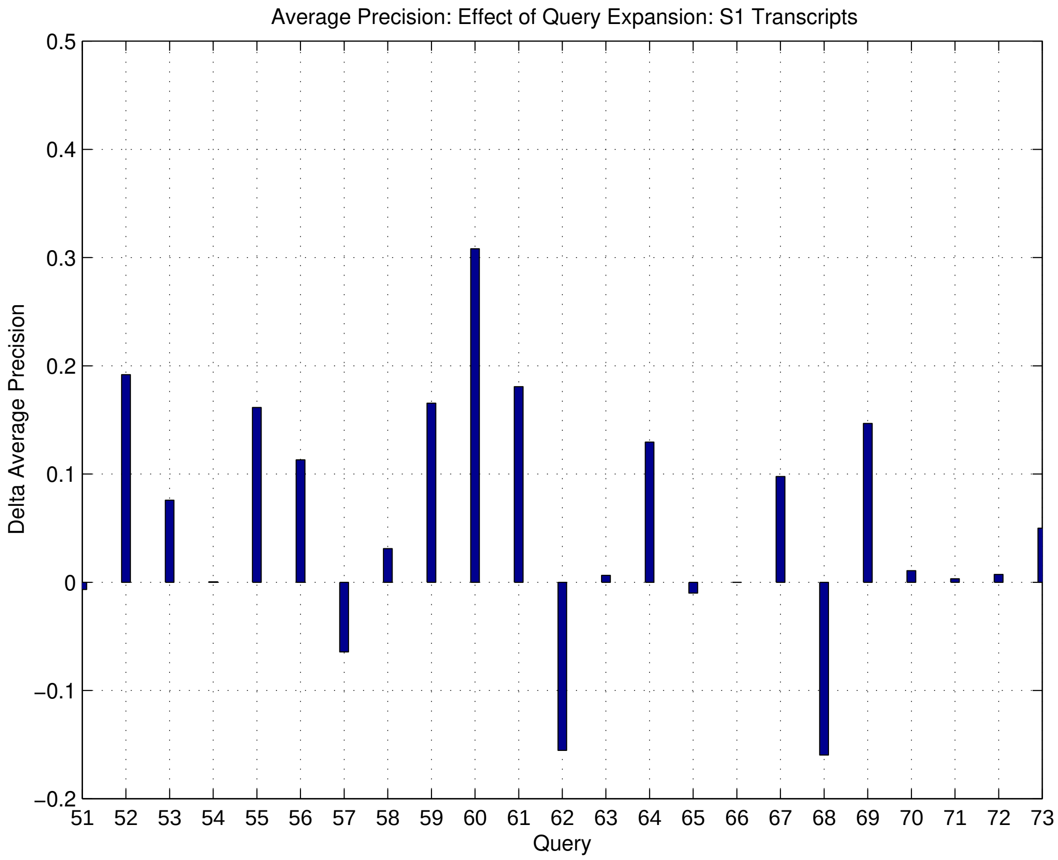

Figure 9: Query-by-query effect of query expansion in terms of change in average precision compared with no query expansion.

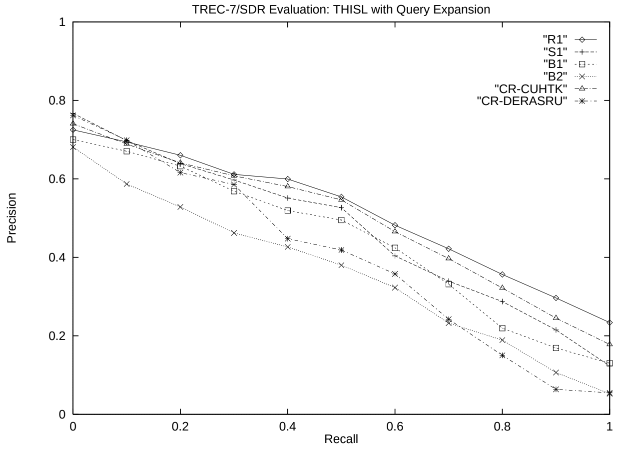

Figure 7: Recall-precision curves of the THISL system running on various transcripts submitted for TREC-7/SDR.

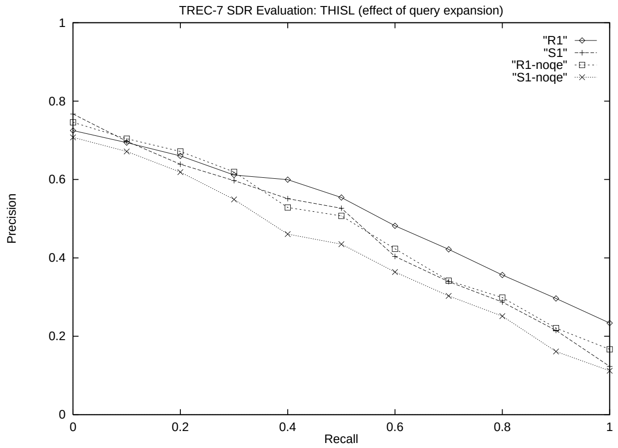

Figure 8: Effect of query expansion on recall-precision for evaluation R1 and S1 conditions (post-evaluation experiment).

60: What information is available on the activities and motivation of intrusive photographers, i.e., the so-called paparazzi?

Original Query: activ avail paparazzi photograph intrus motiv call (AveP = 0.5630)

Expansion Terms: spencer ritz gambino merced editor trespass tabloid (AveP = 0.8589)

A query for which query expansion failed was the following:

62: Find reports of fatal air crashes.

Original Query: air fatal crash (AveP = 0.3520)

Expansion Terms: auto aviat safeti vehicl occup bag jour util (AveP = 0.1893)

9. CONCLUSIONS

The major conclusions that we have drawn from these experiments are:

- Query expansion using a secondary collection derived from newswire data from a similar time period gives a consistent relative improvement in average precision of around 10%.
- Although speech recognizer word error rate does have an effect on recall and precision of the retrieval performance, there is not a clear linear relationship. It seems to be the case that varying retrieval strategy has a much greater effect than improving the recognizer.
- Our first attempts at including word graph and multiple transcription information have not resulted in improvements in recall and precision.
- Using a 100 hour audio archive, spoken document retrieval using a relatively high WER speech recognizer has around 5% lower average precision compared with the reference transcriptions.

10. 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] D. Abberley, S. Renals, G. Cook, and T. Robinson, "The 1997 THISL spoken document retrieval system," in *Proc. Sixth Text Retrieval Conference (TREC–6)*, pp. 747–752, 1998.

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

[4] A. J. Robinson, "The application of recurrent nets to phone probability estimation," *IEEE Trans. Neural Networks*, vol. 5, pp. 298–305, 1994.

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

[6] G. Williams and S. Renals, "Confidence measures derived from an Acceptor HMM," in *Proc. Int. Conf. Spoken Language Processing*, 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] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Amer.*, vol. 87, pp. 1738–1752, 1990.

[9] D. J. Kershaw, M. M. Hochberg, and A. J. Robinson, "Context-dependent classes in a hybrid recurrent network-HMM speech recognition system," in *Advances in Neural Information Processing Systems*, vol. 8, MIT Press, 1996.

[10] 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.

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

[12] C. Fox, "Lexical analysis and stoplists," in *Information Retrieval: Data Structures and Algorithms* (W. B. Frakes and R. Baeza-Yates, eds.), ch. 7, pp. 102–130, Prentice Hall, 1992.

[13] F. Crestani, M. Sanderson, M. Theophylactou, and M. Lalmas, "Short queries, natural language and spoken document retrieval: Experiments at Glasgow University," in *Proc. Sixth Text Retrieval Conference (TREC-6)*, pp. 667–686, 1998.

[14] A. Singal, J. Choi, D. Hindle, and F. Pereira, "AT&T at TREC-6: SDR track," in *Proc. Sixth Text Retrieval Conference (TREC-6)*, pp. 227–232, 1998.

[15] 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.

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