Improving Retrieval on Imperfect Speech Transcriptions

P. Jourlin†, S.E. Johnson‡, K. Spärck Jones† & P.C. Woodland‡

†Cambridge University Computer Laboratory
Pembroke Street, Cambridge, CB2 3QG, UK.
Email: {pj207,ksj}@cl.cam.ac.uk

‡Cambridge University Engineering Department
Trumpington Street, Cambridge CB2 1PZ, UK.
{sej28,pcw}@eng.cam.ac.uk

Abstract

This paper presents the results from adding several forms of query expansion to our retrieval system running on several sets of transcriptions of broadcast news from the 1997 TREC-7 spoken document retrieval track.

1 Introduction

Retrieving documents which originated as speech is complicated by the presence of errors in the transcriptions. If some method of increasing retrieval performance despite these errors could be found, then even low-accuracy automatically generated transcriptions could be used as part of a successful spoken document retrieval (SDR) system. This paper presents results using four query expansion techniques described in [3] on 8 different sets of transcriptions generated for the 1997 TREC-7 SDR evaluation.

The baseline retrieval system and the techniques used for query expansion are described in section 2, the transcriptions on which the experiments were performed in section 3 and results and further discussion are given in section 4.

2 Retrieval Systems

2.1 Baseline System (BL)

Our baseline system uses most of the strategies applied in our 1997 TREC-7 SDR evaluation system [1]. Compound word processing was applied for geographical names such as `New York` and `United Kingdom`. A list of 400 words were defined for stopping and abbreviations such as "C. N. N." were made into single words. Porter's algorithm was used for stemming along with a list of exceptions and a synonym map for place names (such as U.S.A. ≡ U.S.). The index file was then generated containing the term frequencies, collection frequencies and document lengths and used for retrieval with the part-of-speech weighted query. A ranked list of documents was thus produced using the standard combined weight formulae. Further details of this system can be found in [3].

2.2 Geographic Semantic Posets (GP)

It was assumed that if a user wants to retrieve documents about a general entity, then documents which are about a more specific entity which is seen as a *part-of* it may also be relevant. For example, if someone is trying to find information about events in the U.S., occurrences of 'California' within the documents should not be ignored. Since locations are very common in requests in the broadcast news domain it was decided to form a partially ordered set (poset) of location information and use this to attach to each request location word the set of words which express its sub-locations. The frequency within a document of a location word is then the sum of occurrences of its sub-locations (which includes itself), whilst its collection frequency is the number of documents in which at least one of its sub-locations occur. An example of a geographic semantic poset is given in Figure 1.

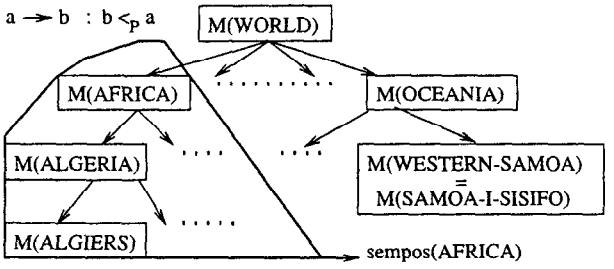

Figure 1: Example of Geographic Semantic Poset

2.3 WordNet Hyponym Posets (WP)

Adding semantic entities which are *part-of* more generalised entities is not restricted purely to location information. Providing a term has only one possible sense in the document file, this approach can be used on any kind of term. A list of unambiguous nouns was obtained from WordNet1.6 and a noun hyponym semantic poset was constructed using the *is-a* relation. For example, malaria *is-a* disease, so the query term `disease` would be expanded to include `malaria`. Note that words which have more than one possible sense were ignored during the expansion process.

2.4 Parallel Blind Relevance Feedback

A parallel corpus of 18628 documents from manually transcribed broadcast news shows was assembled. This cor-

pus spanned January to May 1997 and thus pre-dated the TREC-7 test data. Retrieval was performed on this corpus and the top 15 documents were assumed to be relevant. From these documents the 5 terms with the highest Offer Weight were automatically extracted and added to the query. Since the parallel corpus supposedly contains no transcription errors it was hoped that this process would recover terms missing from the automatic transcriptions, thus increasing average precision. The parallel corpus offers more robust Offer Weight estimation since it is much larger than the test collection and by increasing precision at low recall levels, subsequent blind relevance feedback could also potentially be improved for all the transcriptions.

2.5 Blind Relevance Feedback (BRF)

Blind relevance feedback on the actual test corpus was also included, adding the term with the highest Offer Weight retrieved from the top 5 documents.

3 Transcriptions Used

The experiments reported in this paper use the reference (manually generated) transcriptions, two baseline transcriptions generated by NIST using the CMU recogniser and our own HTK transcriptions. These constituted the mandatory runs in the 1997 TREC-7 SDR evaluation. Results are also reported for the transcriptions from Dragon, AT&T, Sheffield University and DERA which we used as cross-recogniser conditions in the evaluation.

Traditionally word error rate (WER) has been used to report the performance of a speech recogniser. However, since this requires an alignment of the transcriptions and thus is word-order dependent, it does not seem appropriate in a retrieval context when word order is not important. To overcome this problem a term error rate (TER) has been introduced [2] which does not depend on word order. It is also possible to calculate the TER after preprocessing to take into account the effects of stopping, stemming etc. It has been shown that this processed term error rate (PTER) can offer a better predictor of retrieval performance than WER [1] and therefore the transcriptions described in this paper are compared by PTER (averaged over stories). The error rates of the recognisers are given in Table 1.

	HTK	ATT	Dragon	Base1	Sheff	Base2	DERA
WER	24.8	31.0	29.8	34.6	35.8	47.1	61.5
TER	35.7	40.7	42.0	50.1	49.1	69.8	90.0
PTER	34.6	39.7	41.6	48.5	50.4	68.9	93.0

Table 1: Error rates for the transcriptions

4 Results and Discussion

The average precision for the expansion techniques described in section 2 on the transcriptions described in section 3 are given in Table 2 and the results are shown in Figure 2.

It is clear that adding geographic semantic posets, noun-based semantic posets from WordNet and parallel blind relevance feedback (PBRF) increases performance for all transcription error rates. PBRF is especially beneficial at high

	Ref	HTK	AT&T	Dragon
1=BL (7.04)	49.11	47.30	44.84	44.27
2=1+GP (7.04)	51.55	49.77	47.47	46.08
3=2+WP (7.04)	52.33	50.75	48.39	46.59
4=3+PBRF(12.04)	53.59	51.73	50.64	48.99
5=4+BRF (13.04)	55.88	55.08	53.48	50.86

	Base1	Sheff	Base2	DERA
1=BL (7.04)	42.95	44.27	33.95	38.70
2=1+GP (7.04)	45.09	46.17	35.71	39.74
3=2+WP (7.04)	46.53	46.84	36.26	40.47
4=3+PBRF(12.04)	48.03	49.26	40.13	44.22
5=4+BRF (13.04)	51.96	51.97	39.73	44.15

Table 2: Average Precision on the TREC-7 SDR test collection with mean number of query terms in parenthesis

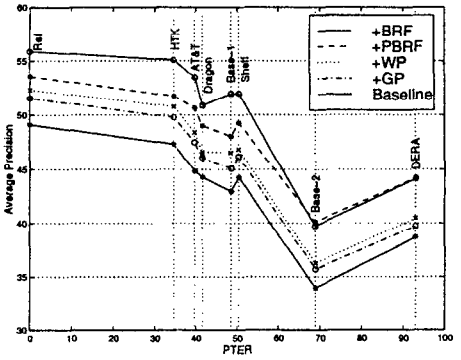

Figure 2: Average precision for different expansion techniques across a wide range of PTER

error rates which confirms the theory that it is compensating for transcription errors. Note also that the difference between the reference and the automatically derived transcriptions is reduced for all but the most accurate set of transcriptions after PBRF has been added.

In these experiments blind relevance feedback on the document set is beneficial for the lower error rate transcriptions, but is ineffective at higher error rates. The decision to include blind relevance feedback in a system should therefore be influenced by the accuracy of recognition process.

The expansion techniques presented in this paper have been shown to not only reduce the difference in performance between manually and automatically generated transcriptions, but also to increase retrieval performance by more than 14% relative on all the sets of transcriptions.

References

[1] S E Johnson, P Jourlin, G L Moore, K Spärck Jones & P C Woodland. *Spoken Document Retrieval for TREC-7 at Cambridge University.* Proc. TREC-7, 1999

[2] S E Johnson, P Jourlin, G L Moore, K Spärck Jones & P C Woodland. *The Cambridge University Spoken Document Retrieval System.* Proc. ICASSP'99, Vol I, pp. 49-52

[3] P Jourlin, S E Johnson, K Spärck Jones & P C Woodland. *General Query Expansion Techniques for Spoken Document Retrieval.* Proc ESCA Workshop on Extracting Information from Spoken Audio, pp. 8-13