

GENERAL QUERY EXPANSION TECHNIQUES FOR SPOKEN DOCUMENT RETRIEVAL

Pierre Jourlin[†], Sue E. Johnson[‡], Karen Spärck Jones[†] and Philip 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 some developments in query expansion and document representation of our Spoken Document Retrieval (SDR) system since the 1998 Text REtrieval Conference (TREC-7).

We have shown that a modification of the document representation combining several techniques for query expansion can improve Average Precision by 17% relative to a system similar to that which we presented at TREC-7 [1]. These new experiments have also confirmed that the degradation of Average Precision due to a Word Error Rate (WER) of 25% is relatively small (around 2% relative). We hope to repeat these experiments when larger document collections become available to evaluate the scalability of these techniques.

1. INTRODUCTION

Accessing information in spoken audio encompasses a wide range of problems, in which spoken document retrieval has an important place. A set of spoken documents constitutes the file for retrieval, to which the user addresses a *request* expressing an *information need* in natural language.

This original sequence of words is transformed by the system into a set of *query terms* which are used to retrieve documents which may or may not meet the user's information need. A good SDR system retrieves as many *relevant* documents as possible whilst keeping the number of *non-relevant* retrieved documents to a minimum. For this work we take text-based queries and use an Automatic Speech Recognition (ASR) system to produce word-based transcriptions for the documents.

Following earlier scattered studies, the TREC-7 SDR evaluation provided further support for the claim that conventional Information Retrieval (IR) methods are applicable to automatically transcribed documents. Our retrieval system was run on 7 different sets of automatically transcribed broadcast news texts with a WER varying from 24.8% to 66.2%. The corresponding range of Average Precision was 45% to 35% [1].

The difference in Average Precision between the best ASR-based system and the manual reference transcriptions was only 5% relative for our retrieval engine. Therefore we concluded that improving IR performance would probably be more profitable than improving ASR performance. We have therefore focussed our research on general IR techniques, such as query expansion, implemented within the Probabilistic Retrieval Model (PRM) [2].

The formal framework for this research is presented in sections 2 and 3, with the experimental procedure and system description in section 4. Results are given in section 5 and conclusions are drawn in section 6.

2. A BRIEF DESCRIPTION OF THE PRM

The PRM framework [2] is not too prescriptive on document representation. Here we address the relation between the notion of query and document index *terms* and the more ordinary notion of *words*. In section 3, we show how more complex relations can be established to enrich the document representation.

The PRM is based on the idea that documents are ranked by the retrieval engine in order of decreasing estimated probability of relevance $P_Q(R|D)$.¹ The relevance R is taken to be a basic, binary criterion variable. D is a random variable taking values in the document universe Ω_D .

For a given document collection, Ω_D is a set of possible events, each event corresponding to the occurrence of a particular document and document representation. The query Q is used in the creation of the document representation and therefore is necessary to define Ω_D .

Suppose, for the moment, that the query terms are just plain words. By assuming all query words are independent, a document event might be represented as the set of couples $(w, wf(w, d))$ for all query words w , where the word frequency $wf(w, d)$ is the number of occurrences of w in document d . By way of illustration, a small but complete retrieval example could be given by :

This work is in part supported by an EPSRC grant on Multimedia Document Retrieval reference GR/L49611.

¹To estimate $P_Q(R|D)$, additional information outside the document universe, such as the document length, may be used.

Query : “information retrieval”
 1st doc. : “information retrieval is no easy task”
 2nd doc. : “speech is an information rich medium”

$$\begin{aligned}
 e_1 &= \{(information, 1), (retrieval, 1)\} \\
 e_2 &= \{(information, 1), (retrieval, 0)\} \\
 \Omega_D &= \{e_1, e_2\} \\
 \Omega_R &= \{yes, no\} \\
 &P(R = yes | D = e_1) > P(R = yes | D = e_2)
 \end{aligned}$$

The original query word frequencies within a document therefore provide basic predictors of relevance. However, we should bear in mind that the query words are derived from the user’s original *request*, which in turn conveys a need which could have been expressed differently. In addition, as queries are just word sets, the same set could have been extracted from different text requests. In the next section we review various ways of modifying the document universe: some are well established, others less so, but we believe they are worthy of further study. More specifically, we introduce *semantic posets* as an appropriate characterisation for particular forms of modification.

3. MODIFICATIONS OF THE DOCUMENT UNIVERSE

3.1. Compound Words

Context can change the *meaning* of words dramatically. One method of taking context into account is to treat a given sequence of words as an irreducible *atomic* semantic unit. The atomic terms in this approach are either individual words or multi-word sequences that are treated as explicit and undecomposable. Some proper names (e.g. New York) may be such compound words.

The compound word vocabulary Φ can then be added to the single-word vocabulary (extracted from the document collection) to give the new atomic term vocabulary V .

Both the original queries and documents are segmented so that the longest possible sequence of words in V is always preferred. For example, suppose Φ consists of the sequences **new-york-city** and **new-york**, then the sentence “New York City is in the state of New York but York is in the county of North Yorkshire” produces “**new-york-city** is in the state of **new-york** but **york** is in the county of north yorkshire”².

A new document universe is now defined in a similar way as before, but with the notions of *word* and *word frequency* replaced by those of *atomic term* $at \in Q'$ and *atomic term frequency* $atf(at, d)$, where Q' is the query formed from V .

These new atomic terms should act as better relevance predictors. For example, a document about *New*

²as opposed to “new **york** city ...”

York is not likely to be relevant to a query about *York* and vice versa.

Such compound words should only be used when there are no alternative ways of expressing the same concept using some or all of the constituent words. Thus *information retrieval* should not be a compound word as we may have the alternative *retrieval of information* or simply *retrieval* alone.

3.2. Removing Stop Words

Non-content words (e.g. the, at, with, do, ...) are generally of no retrieval value [3]. Most IR systems define a set S of these *stop words* and remove them from both the queries and the documents.

The new document universe is defined with a set of query atomic terms $at \in (Q'' = Q' - S)$ and an atomic term frequency function :

$$atf(at, d) = \begin{cases} 0 & \forall at \in S \\ \text{number of occurrences of } at \text{ in } d & \forall at \notin S \end{cases}$$

3.3. Stemming

Stemming [4] allows the system to consider the words with a (real or assumed) common root as a unique semantic class. For an atomic term at_i , a corresponding set of atomic terms $st(at_i)$ exists which share the same stem (e.g. $st(\text{trains}) = \{\text{train, trainer, trained, training, trains...}\}$).

We define a *term* t as a set of *atomic terms*. The term frequency $tf(t, d)$ is then defined as :

$$tf(t, d) = \sum_{at \in t} atf(at, d)$$

The corresponding events making up the document universe are therefore defined as :

$$e_i = \bigcup_{at \in Q''} \{(st(at), tf(st(at), d_i))\}$$

An example at this stage would look like :

Query : “Trains in New York”
 1st doc. : “There is a train in New York”
 2nd doc. : “The trainer is training in New York”
 $st(\text{trains}) = \{\text{trainer, train, training}\}$
 $st(\text{new-york}) = \{\text{new-york}\}$

$$\begin{aligned}
 e_1 &= \{(st(\text{trains}), 1), (st(\text{new-york}), 1)\} \\
 e_2 &= \{(st(\text{trains}), 2), (st(\text{new-york}), 1)\} \\
 &P(R = yes | D = e_1) < P(R = yes | D = e_2)
 \end{aligned}$$

3.4. Semantic Posets

It is also possible to use a list of equivalence classes of terms to allow more complex associations. We assume that the user’s original query words refer to semantic units rather than just words. Therefore, words

which share the same meaning should be considered as equivalent. A simple equivalence list can be used to process synonyms in a similar way to how the stemming procedure deals with the different forms of individual words.

In addition, we can also assume that if the user is interested in a general semantic entity, then s/he is also interested in more specific entities which are seen as *part of* it.

For instance, the word `Europe` may refer to the class containing the names of all European countries, regions and cities, whilst the word `England` only refers to the class containing the English county and city names.

Several attempts to deal with this particular kind of semantic structure have been described in the literature (e.g. [5, 6]). However, the corresponding experiments have shown very little improvement in IR performance.

We find it convenient to represent this behaviour of term classes by considering a semantic Partially Ordered Set (poset) [7] which contains the *meaning* $M(at_i)$ of each atomic term at_i .

The equivalence relation for poset P , $=_P$, could be taken from a synonym thesaurus and the strict partial ordering $<_P$ relation from a hyponym thesaurus. An atomic term at is considered more specific than at' if $M(at) \leq_P M(at')$. The two thesauri are kept consistent by ensuring the properties of posets are not broken.

We define the function *sempos* which assigns the set of equivalent or more specific atomic terms to a given atomic term at :

$$sempos(at) = \bigcup_{at' : M(at') \in P} \{at'\} : M(at') \leq_P M(at)$$

The document universe is then defined from the events :

$$e_i = \bigcup_{at \in Q''} \{(st(sempos(at)), tf(st(sempos(at)), d_i))\}$$

Figure 1 shows an example of a poset representing geographic locations and sub-locations using a tree structure to show the partial ordering relation.

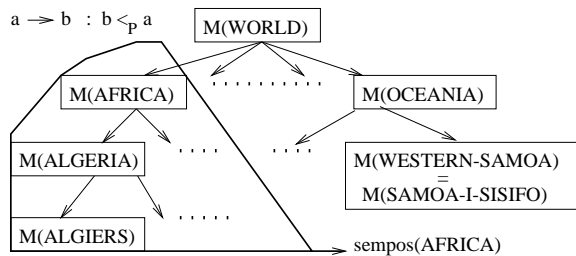


Figure 1: Example of Geographic Semantic Poset

3.5. Blind Relevance Feedback

Some words which do not appear in the query may still act as good predictors of relevance. However, such words may be difficult to find from individual query terms.

As is well known, some information about the relevance of the documents could be used to identify words that are good relevance predictors. This information can be used to reweight existing query words or add new ones.

It is also possible just to assume that the highest ranked documents in an initial search are relevant. Such Blind Relevance Feedback (BRF) adds the top T terms drawn from the top B retrieved documents, where the top T are defined by their Offer Weight³ as described in [2]. This set of terms $brf(\Omega_D)$ is added to the previous document universe, thus producing the new document universe, defined from the events :

$$e'_i = brf(\Omega_D) \cup e_i$$

where e'_i is the new event related to document d_i and created from the previous event e_i and document universe Ω_D . We may also add terms taken from the document universe of a parallel corpus which is denoted Parallel Blind Relevance Feedback (PBRF).

4. EXPERIMENTS

4.1. Data

The experiments reported here use the TREC-7 SDR test data. For this evaluation, the audio documents came from American broadcast radio and TV news programs which had been manually divided into separate news stories. The text requests were expressed in natural language text, such as “Where are communists and communist organizations active in the world today?”. The participating teams had to transcribe the audio automatically and run an IR engine on this transcription to provide a ranked list of matching, i.e. potentially relevant, documents.

Human relevance assessments were used to evaluate the ranked list and determine performance using standard measures based on Precision and Recall.

There are two main considerations when describing the data for SDR. Firstly the audio data used for transcription, and secondly the request/relevance set used during retrieval. Table 1 describes the main properties of the former, while Table 2 describes the latter.

4.2. Transcription System

The transcription of spoken documents was done using part of our HTK broadcast news transcription system [8].

The input data is presented to the system as complete episodes of broadcast news shows and these are

³a particular formula is given in section 4.3.4

Nominal Length of Audio	100 hours
Number of Documents	2866
Number of Different Shows	8
Approx. Number of Words	770,000
Average Document length	269 words

Table 1: Description of data used

Number of Requests	23
Average Length of Request	14.7 words
Number of Relevant Docs (NRD)	390
Average NRD per Request	17.0 docs

Table 2: Description of request and relevance sets used

first converted to a set of segments for further processing [9]. The segmentation uses Gaussian mixture models to divide the audio into narrow and wide-band and also to discard parts of the audio stream that contain no speech (typically pure music). The output of a phone recogniser is used to determine the final segments which are intended to be acoustically homogeneous.

Each frame of input speech to be transcribed is represented by a 39 dimensional feature vector that consists of 13 (including c_0) cepstral parameters and their first and second differentials. Cepstral mean normalisation is applied over a segment.

Our system uses the LIMSI 1993 WSJ pronunciation dictionary augmented by pronunciations from a TTS system and hand generated corrections. Cross-word context dependent decision tree state clustered mixture Gaussian HMMs are used with a 65k word vocabulary. The full HTK system [8] operates in multiple passes and uses complex language models via lattice rescoring and quinphone HMMs. This system gave a word error rate of 16.2% in the 1997 DARPA Hub4 broadcast news evaluation.

The TREC-7 HTK SDR system uses the first two passes in a modified form for reduced computational requirement. The first pass uses gender independent, bandwidth dependent cross-word triphone models with a trigram language model to produce an initial transcription. The output of the first pass is used along with a top-down covariance-based segment clustering algorithm [10] to group segments within each show to perform unsupervised test-set adaptation using maximum likelihood linear regression-based model adaptation.

A second recognition pass through the data is then performed using a bigram language model to generate word lattices using adapted gender and bandwidth specific HMMs. These bigram lattices are expanded using a 4-gram language model and the best path through these lattices gives the final output. This system runs in about 50 times real-time on a Sun Ul-

tra2 and achieves an error rate of 17.4% on the 1997 Hub4 evaluation data. It should be noted that the error rates on Hub4 data and TREC data are not strictly comparable in part due to the differences in quality of the reference transcriptions.

The HMMs used in TREC-7 were trained on 70 hours of acoustic data and the language model was trained on manually transcribed broadcast news spanning the period of 1992 to May 1997 supplied by the LDC and Primary Source Media (about 152 million words in total). The language model training texts also included the acoustic training data (about 700k words) and 22 million words of text from the Los Angeles Times and Washington Post covering the span of the evaluation period (June 1997 to April 1998 inclusive).

Using all these sources a 65k wordlist was chosen from the combined word frequency list while ensuring that a manageable number of new pronunciations had to be created. The final wordlist had an out-of-vocabulary rate of 0.3% on the TREC-7 data. The overall system gave a WER of 24.8% which corresponded to a Processed Term Error Rate [11] (which more closely represents the error rate as seen by the retriever) of 32.1%.

4.3. Retrieval Systems

4.3.1. Baseline System (BL)

Our current SDR Baseline System, BL, uses most of the strategies applied in our TREC-7 SDR evaluation system.

The list of compound words was generated for geographical names taken from a travel web server (for example: *New-York*, *New-Mexico*, *Great-Britain*). The compound name processing described in section 3.1 was then applied.

Stopping as described in section 3.2 was then applied using a list of 400 stop words. Finally, stemming as described in section 3.3 was implemented using Porter's algorithm [4], along with an extra stage to correct possible incorrect spellings in the transcriptions. These devices are *query independent* and therefore were implemented as a text pre-processing phase on the queries and documents.

The index file was then generated. It contains the number of documents in the collection N , the length of each document, $dl(d_j)$, the number of documents containing each term, $n(t_i)$, and the number of times the term occurs in the given document (term frequency), $tf(t_i, d_j)$.

The document representation which was described in the preceding sections, together with a score [2], specifies how to generate a final ranking of documents from a given document universe. For each document, d_j , a score is generated for each query by summing the combined weights, $cw(t_i, d_j)$ for each term t_i produced from the following formulae:

$$\begin{aligned}
cw(t_i, d_j) &= \overline{pos(t_i)} \times \\
&\quad \frac{(\log N - \log n(t_i)) \cdot tf(t_i, d_j) \cdot (K + 1)}{K \cdot (1 - b + b \cdot ndl(d_j)) + tf(t_i, d_j)} \\
n(t_i) &= \sum_{d_i \in \mathcal{D}} \begin{cases} 0 & tf(t_i, d_i) = 0 \\ 1 & tf(t_i, d_i) > 0 \end{cases} \\
dl(d_j) &= \sum_{w \in V} tf(w, d_j) \\
ndl(d_j) &= \frac{dl(d_j) \cdot N}{\sum_{d \in \mathcal{D}} dl(d)}
\end{aligned}$$

where V is the term vocabulary for the whole document collection \mathcal{D} ; K and b are tuning constants and $\overline{pos(t_i)}$ the part-of-speech [12] weight of term t_i . A ranked list of documents is thus produced for each query by sorting the returned match scores in descending order. The POS weights were those used in the Cambridge TREC-7 SDR evaluation system:

Proper Noun	1.2
Common Noun	1.1
Adjective & Adverbs	1.0
Verbs and the rest	0.9

4.3.2. Adding Geographic Semantic Posets (GP)

Location information is very common in requests in the broadcast news domain. Our first extension implements the expansion of geographic names occurring in the original query of the BL system into the list of their components, e.g :

US \rightarrow Arizona, ..., Wisconsin, ...
Atlanta, ..., Washington-D.C., ...

We manually built a semantic poset containing 484 names of continents, countries, states and major cities, extracted from a travel web server. The poset is represented by a semantic tree whose nodes are location names and edges are the *contains* relation. The process of using posets, described in section 3.4 is applied, creating a new index term for each *sempos(at)*, with $at \in Q''$.

4.3.3. Adding WordNet Hyponyms Posets (WP)

The previous approach can be generalised to every kind of term, provided that they only have one possible sense in the document file. We obtained a list of unambiguous nouns from WordNet 1.6 [13] and then, assuming that these words are actually unambiguous in the file and also in the query, generated the corresponding noun hyponym trees (*is-a* relation. For instance, the query term *disease* is expanded into *flu* and *malaria* but words like *air* (e.g. gas or aviation or manner) are ignored in this expansion process as they have more than one sense. In these experiments we do not consider WordNet compound words, their proper handling being much more complicated than in the geographic names domain.

4.3.4. Adding Parallel Blind Relevance Feedback (PBRF)

We assembled a parallel corpus of 18628 documents from a manually transcribed broadcast news collection covering January to May 97, thus not overlapping the TREC-7 collection recording time⁴. Parallel Blind Relevance Feedback, as described in section 3.5 was then applied on this document collection.

The five terms which obtain the highest Offer Weight were appended to the query. The Offer Weight of a term t_i is :

$$ow(t_i) = r \cdot \log \frac{(r + 0.5)(N - n - B + r + 0.5)}{(n - r + 0.5)(B - r + 0.5)}$$

where B is the number of top document which are assumed relevant, r the number of assumed relevant documents in which at least one $at \in t_i$ occurs, n the total number of documents in which at least one $at \in t_i$ occurs and N the total number of documents. For these experiments we used $B = 15$. The terms added to the query were then given a weighting equal to their Offer Weight in a similar way to the part-of-speech weighting described in section 4.3.1.

4.3.5. Adding Blind Relevance Feedback (BRF)

The BRF process was also applied to the actual TREC-7 corpus. This time, only one term was added from the top five documents that were retrieved for each PBRF expanded query.

5. RESULTS

The five systems described in section 4.3 were evaluated on the TREC-7 SDR test data. An illustration of the complete expansion process is given in Table 3.

	what diseases are frequent in Britain ?
BL	disease frequent Britain
+GP	disease frequent {Britain, UK, ..., Cambridge}
+WP	{disease, flu, ..., malaria} frequent {Britain, UK, ..., Cambridge}
+PBRF	+ cold Blair rheumatism queen
+BRF	+ endemic

Table 3: Illustration of the query expansion process

The results in terms of Average Precision, and also Precision at 5 documents retrieved, for both the manual and HTK transcriptions, are given in Table 4.

We can see that improving the IR system produced a combined relative gain of 17% in Average Precision on the automatic transcriptions (lines 1 and 5 of Table 4). Lines 6-8 of Table 4 show the results we obtain when each expansion technique in turn is omitted.

⁴This corpus is a subset of the Primary Source Media broadcast news transcriptions used to train the language model of our speech recognition system.

	Transcriptions			
	Manual		HTK	
	AvP	P-5	AvP	P-5
1 = BL	49.11	57.39	47.30	56.52
2 = 1 + GP	51.55	60.00	49.77	58.26
3 = 2 + WP	52.33	60.00	50.75	58.26
4 = 3 + PBRF	53.59	64.35	51.73	64.35
5 = 4 + BRF	55.88	60.87	55.08	64.35
6 = 5 - PBRF	53.54	60.00	52.56	59.13
7 = 5 - WP	54.40	60.00	54.20	63.48
8 = 5 - GP	54.95	59.13	53.56	60.87

Table 4: Average Precision (AvP) and Precision at a 5 documents cut-off (P-5) on the TREC-7 test collection (results in %)

It confirms that each expansion device is necessary to reach the final performance of line 5 but that very good results can be obtained by using only Blind Relevance Feedback techniques.

It is worth noting that for each level of IR performance, recognition errors do not affect the results by more than 4% relative and by more than 2% relative for the final system.

Assuming that the manual transcriptions place an upper bound on performance, these experiments suggest that the adaptation of IR to ASR (e.g. using word lattices) would be less profitable than future improvements in IR techniques.

6. CONCLUSION

In this paper we have confirmed that, for a small document file, the degradation of Average Precision due to a WER of 25% is relatively small (2%). We have shown that several devices which modify the document representation, each improve the system performance slightly and that the new methods based on semantic posets might be successfully combined with Blind Relevance Feedback, and are therefore worthy of further study.

More generally, we have shown that pure information retrieval techniques can improve Average Precision by 17% relative on the TREC-7 SDR task. We hope to repeat these experiments when larger document collections become available, in order to evaluate the scalability of our techniques.

7. REFERENCES

[1] S. E. Johnson, P. Jourlin, G. L. Moore, K. Spärck Jones, and P. C. Woodland, "Spoken document retrieval for TREC-7 at Cambridge University," in *To appear in Proc. TREC-7*, (Gaithersburg, MD), 1999.

[2] K. Spärck Jones, S. Walker, and S. E. Robertson, "A probabilistic model of information retrieval : Development and status," TR 446,

Cambridge University Computer Laboratory, September 1998.

- [3] 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.
- [4] M. F. Porter, "An algorithm for suffix stripping," *Program*, no. 14, pp. 130–137, 1980.
- [5] G. Salton and M. E. Lesk, "Computer evaluation of indexing and text processing," in *The SMART Retrieval System: Experiments in Automatic Document Processing* (G. Salton, ed.), pp. 143–180, Prentice Hall, 1971.
- [6] E. M. Voorhees, "Query expansion using lexical-semantic relations," in *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pp. 61–69, 1994.
- [7] B. Dushnik and E. W. Miller, "Partially ordered sets," *American Journal of Mathematics*, no. 63, pp. 600–610, 1941.
- [8] P. C. Woodland, T. Hain, S. E. Johnson, T. R. Niesler, A. Tuerk, E. W. D. Whittaker, and S. J. Young, "The 1997 HTK broadcast news transcription system," in *Proc DARPA Broadcast News Transcription and Understanding Workshop*, pp. 41–48, 1998.
- [9] T. Hain, S. E. Johnson, A. Tuerk, P. C. Woodland, and S. J. Young, "Segment generation and clustering in the HTK broadcast news transcription system," in *Proc. 1998 DARPA Broadcast News Transcription and Understanding Workshop*, pp. 133–137, 1998.
- [10] S. E. Johnson and P. C. Woodland, "Speaker clustering using direct maximisation of the MLLR-adapted likelihood," in *Proc. ICSLP 98*, vol. 5, pp. 1775–1779, 1998.
- [11] S. E. Johnson, P. Jourlin, G. L. Moore, K. Spärck Jones, and P. C. Woodland, "The Cambridge University spoken document retrieval system," in *Proc. of ICASSP'99*, vol. 1, pp. 49–52, 1999.
- [12] S. F. Knight, "Personal communication," 1998.
- [13] C. Fellbaum, *WordNet : An Electronic Lexical Database*. ISBN 0-262-06197-X, MIT Press, 1998.