Spoken document representations for probabilistic retrieval

Pierre Jourlin, Sue E Johnson, Karen Spärck Jones, Philip C. Woodland

ELSEVIER

Speech Communication 32 (2000) 21–36

SPEECH COMMUNICATION

www.elsevier.nl/locate/specom

Spoken document representations for probabilistic retrieval

Pierre Jourlin [a,*], Sue E. Johnson [b], Karen Spärck Jones [a], Philip C. Woodland [b]

[a] *Computer Laboratory, Cambridge University, New Museums Site, Pembroke Street, Cambridge, CB2 3QG, UK*
[b] *Engineering Department, Cambridge University, Trumpington Street, Cambridge, CB2 1PZ, UK*

Abstract

This paper presents some developments in query expansion and document representation of our spoken document retrieval system and shows how various retrieval techniques affect performance for different sets of transcriptions derived from a common speech source. Modifications of the document representation are used, which combine several techniques for query expansion, knowledge-based on one hand and statistics-based on the other. Taken together, these techniques can improve Average Precision by over 19% relative to a system similar to that which we presented at TREC-7. These new experiments have also confirmed that the degradation of Average Precision due to a word error rate (WER) of 25% is quite small (3.7% relative) and can be reduced to almost zero (0.2% relative). The overall improvement of the retrieval system can also be observed for seven different sets of transcriptions from different recognition engines with a WER ranging from 24.8% to 61.5%. We hope to repeat these experiments when larger document collections become available, in order to evaluate the scalability of these techniques.

Zusammenfassung

Dieser Bericht präsentiert einige Entwicklungen zur Fragenerweiterung and Dokumentendarstellung in unserem Spracherfassungssystem. Verschiedene Erfassungsmethoden, die auf einer Menge von Transkriptionen basieren welche von gängigen Textquellen abgeleitet werden, können auf die Leistung einen Einfluss haben. Wir zeigen, daß Modifikationen an der Dokumentendarstellung in Kombination mit verschiedenen Techniken der Fragenerweiterung, die einerseits wissensbasiert und andererseits statistikbezogen sind, Durchschnittsgenauigkeitverbesserungen von mehr als 19% relativ zu einem änlichem System im Vergleich zu dem hier präsentiertem TREC-7, ergeben. Diese neuen Experimente haben ebenfalls bestätigt, daß die Durchschnittsgenauigkeitdegradierung, bezogen auf die Wortfehlerrate (WER) von 25% ziemlich klein sind (3.7% relativ) und sie auf fast auf null reduziert werden können (0.2% relativ). Die Gesamtverbesserungen des Spracherfassungssystems wurde auch für sieben verschiedene Mengen von Transkriptionen, mit Wortfehlerrate zwischen 24.8% und 61.5%, observiert. Wir hoffen, diese Ergebnisse für grössere Dokumentensammlungen zu wiederholen, um die Skalierbarkeit dieser Techniken abschätzen zu können.

Résumé

Cet article présente quelques développements dans l'expansion de requête et la représentation des documents de notre système de recherche documentaire et montre comment les diverses techniques de recherche affectent la performance pour différents ensembles de transcriptions dérivées d'une source de parole commune. Des modifications de la

* Corresponding author. Tel.: +44-1223-334424; fax: +44-1223-334678.
E-mail address: pierre.jourlin@cl.cam.ac.uk (P. Jourlin).

0167-6393/00/$ - see front matter
PII: S0167-6393(00)00021-2

représentation des documents sont effectuées, qui combinent plusieurs techniques pour l'expansion de requête, fondées sur des connaissances d'une part et sur des statistiques d'autre part. Utilisées conjointement, ces techniques peuvent améliorer la Précision Moyenne de plus de 19%, relativement à un système semblable à celui que nous avons présenté à TREC-7. Ces nouvelles expérimentations ont également confirmé que la dégradation de la Précision Moyenne due à un Taux d'Erreur de Mot (WER) de 25% est vraiment faible (3,7% relatif) et peut être réduite à une quantité négligeable (0,2% relatif). L'amélioration globale du système de recherche documentaire peut aussi être observée pour sept ensembles différents de transcriptions provenant de différents systèmes de reconnaissance ayant un WER variant de 24,8% à 61,5%. Nous espérons reproduire ces expérimentations, lorsque de plus grandes collections de documents parlés seront disponibles, afin d'évaluer le comportement de ces techniques sur de plus gros volumes de données.

Keywords: Spoken document retrieval; Automatic speech recognition; Information retrieval

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* intended to retrieve documents to meet the user's information need. Thus a good spoken document retrieval (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 [1] and use an automatic speech recognition (ASR) system to produce word-based transcriptions for the documents.

Following earlier scattered studies, the SDR evaluation within the seventh Text REtrieval Conference (TREC-7) (Garofolo et al., 1999) provided further support for the claim that conventional information retrieval (IR) methods are applicable to automatically transcribed documents. When our retrieval system was run on seven different sets of automatically transcribed broadcast news texts, from different recognisers and with a WER varying from 24.8% to 66.2%, the corresponding range in retrieval performance as measured by Average Precision was 45–35% (Johnson et al., 1999b).

But the difference in Average Precision between the best ASR-based system and the manual reference transcriptions was only 5% relative for our retrieval engine. We therefore concluded that improving IR performance would probably be more profitable than improving ASR performance, and since the TREC-7 evaluation we have thus focussed our research on general IR techniques, such as query expansion, implemented within the probabilistic retrieval model (PRM) (Spärck Jones et al., 1998), rather than on ASR refinement.

This conclusion may appear surprising to those engaged in ASR, but is in line with the results of three cycles of SDR evaluation in the TREC Programme. Thus while there is a near-linear relation between recognition quality and retrieval performance (Garofolo et al., 1999), the gap between baseline and best retrieval methods is much larger than that between human and system transcription for any particular retrieval method, a point reinforced by cross-recogniser retrieval results. The implication is that once a reasonable level of ASR performance has been achieved, trying to improve it further is less valuable than concentrating on the retrieval techniques used. When serious work on SDR began early in the decade, this was not evident. Specifically, it was thought that it might be necessary to develop special-purpose retrieval devices to compensate for the particular characteristics of imperfect transcriptions but, at least for the type of test material used, there has appeared to be more mileage in trying to apply retrieval techniques already proven

[1] There are separate problems to overcome in accepting spoken requests.

for the text case to recogniser output than in starting afresh in devising retrieval approaches for speech data. It remains to be shown whether this holds for low-quality speech, and also what the effects of, e.g., very short queries and/or documents might be, though retrieval is a coarse task and so is resistant to some degree of error in content representation. Given the growing importance of speech material, the apparent dominance of retrieval methods over recognition refinement, as so far illustrated by TREC, is a finding as relevant to the speech as to the retrieval communities; however given the small scale of evaluations so far, there is still reason to explore potential differences, as further below.

Our tests as a whole have thus applied a range of familiar general ideas, of varying proven value for text retrieval, to the speech case, both individually and in combination: we concentrate on the latter here. 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.

1.1. Unusual symbols

at	atomic term
atf	atomic term frequency
tf	term frequency
wf	word frequency
idf	inverse document frequency
adl	average document length
ndl	normalised document length
dl	document length
st	stemming function
sempos	Semantic Poset function
brf	blind relevance feedback function

2. A brief description of the probabilistic retrieval model (PRM)

The PRM framework (Spärck Jones et al., 1998) 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 $P_{\mathrm{Q}}(R|D)$. [2] 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 can 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:

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

$$e_1 = \{(\text{information}, 1), (\text{retrieval}, 1)\}$$
$$e_2 = \{(\text{information}, 1), (\text{retrieval}, 0)\}$$
$$\Omega_{\mathrm{D}} = \{e_1, e_2\}$$
$$\Omega_{\mathrm{R}} = \{\text{yes}, \text{no}\}$$
$$P(R = \text{yes} \,|\, D = e_1) > P(R = \text{yes} \,|\, D = e_2).$$

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

[2] To estimate $P_{\mathrm{Q}}(R|D)$, additional information outside the document universe as defined through the query terms, such as the document length, may be used.

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, both in general and in news-based applications. 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. Then, assuming some method of forming the compound word vocabulary Φ, this 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 contain the sequence **new-york**, then the sentence "New York is in USA but York is in the county of North Yorkshire" produces "**new-york** is in the usa but **york** is in the county of north yorkshire". [3]

A new document universe is now defined in a similar way to 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. Following standard lines of argument about indexing specificity, these new atomic terms should act as better relevance predictors than the previous ones. For example, a document about *New York* is not likely to be relevant to a query about *York* and vice versa. Such compound words should however 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 defined as a compound word since we may have the alternative *retrieval of information* or simply *retrieval* alone. Our approach to compound words is therefore more restricted, but hopefully more accurately targeted, than those generally applied (see, e.g., Mitra et al., 1997).

3.2. Removing stop words

Non-content words (e.g. the, at, with, do, ...) are generally of no retrieval value (Fox, 1992). 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 (Porter, 1980) 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. $st(\text{trains}) = \{\text{train, trainer, trained, training, trains} \ldots\}$, where *trains* is used simply as the set label, not to show the linguistic form of the stem.).

We now define a *term* t as a set of *atomic terms*. In this particular case, $t = st(at)$. 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

[3] As opposed to "new **york** is ...".

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

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"
$st(\text{trains}) = \{\text{trainer, train, training}\}$
$st(\text{new-york}) = \{\text{new-york}\}$

$$e_1 = \{(st(\text{trains}), 1), (st(\text{new-york}), 1)\},$$
$$e_2 = \{(st(\text{trains}), 2), (st(\text{new-york}), 1)\}$$
$$P(R = \text{yes} \mid D = e_1) < P(R = \text{yes} \mid D = e_2).$$

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

In addition, we can 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 exploit this particular kind of semantic structure within the framework of automatic indexing have been described in the literature, e.g., Salton and Lesk (1971), Voorhees (1994) and Mandala et al. (1999) report TREC experiments with WordNet in particular. There is some slight evidence from TREC as a whole that if information about specific proper name relations (as in the geographic case) is available, this may be moderately useful. The evidence for the value of semantic relations generally, unless applied by humans in careful manual searching, is weak; i.e., there appears to be no material advantage for such purposes as query expansion in using explicit, rather than implicit, statistically captured relations. We have, however, considered thesaural relations here as part of our attempt to explore possible differences between spoken document and text retrieval, in relation both to applying techniques to the former that have been found beneficial for the latter and to checking whether methods that have not been required in the 'easy' text case might still be helpful, because of genre differences or poor transcription, in the more difficult speech case. However, especially as our query set is small, we have not been able to draw firm conclusions about any text/speech differences.

We find it convenient to represent this behaviour of term classes by considering a semantic partially ordered set (*poset*) (Dushnik and Miller, 1941) 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 (and only if) $M(at) \leqslant_P M(at')$. The two thesauri are kept consistent by ensuring that 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

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

The document universe is then defined from the events

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

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

3.5. Blind relevance feedback

Some words which do not appear in the query may still act as good predictors of relevance,

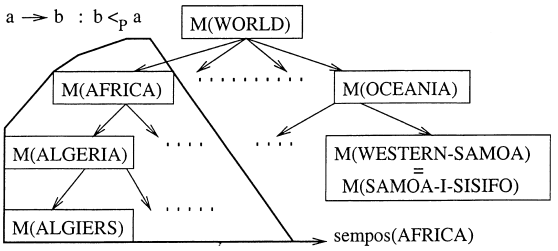

Fig. 1. Example of Geographic Semantic Poset.

though they may be difficult to find directly through individual query terms. As is well known, information about the relevance of documents can be used to identify words that are good relevance predictors, and hence to re-weight existing query words or add new ones. It is also possible, when explicit relevance information is lacking, just to assume that the highest ranked documents in an initial search are relevant. We use such blind relevance feedback (BRF) in query expansion to add the top T terms drawn from the top B retrieved documents, where the top T are defined by their Offer Weight [4] as described in (Spärck Jones et al., 1998). This new set of tuples (*term*, *term frequency*), that we call $brf(\Omega_D)$ 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.

Query expansion using assumed relevance information has become popular in TREC (Spärck Jones, 2000) and some SDR participants (e.g., Abberley et al., 1999; Singhal and Pereira, 1999; Allan et al., 1999) have shown it is profitable in the SDR domain.

We may also, by analogy with several uses of such additional resources by TREC participants (e.g., Singhal and Pereira, 1999), add terms taken from the document universe of a parallel corpus, in parallel blind relevance feedback (PBRF), in particular from a parallel *text* corpus, larger and cleaner than the automatically transcribed test file. It will be evident that while all indexing affects the document universe, some of the devices listed, notably the use of Semantic Posets, are particularly motivated by the wish to make query expansion more effective.

[4] A particular formula is given in Section 4.3.4.

4. Experiments

4.1. Data

The experiments reported here use the TREC-7 SDR test data. The audio documents are from American broadcast radio and TV news programmes which have been manually divided into separate news stories. The test requests are simple natural language text, such as "Where are communists and communist organizations active in the world today?". In TREC-7, 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. Independent human relevance assessments were used to evaluate the ranked list and to determine standard performance measures based on precision (percent of retrieved documents which are relevant) and recall (percent of relevant documents which are retrieved).

Table 1 describes the main properties of the audio document data, while Table 2 describes the request/relevance data. It should be noted that from the retrieval point of view this is an extremely small document set, although it is large by conventional ASR standards.

Table 1
Description of data used

Nominal length of audio	100 hours
Number of documents	2,866
Number of different shows	8
Approx. number of words (before any indexing operation)	770,000
Average document length	269 words

Table 2
Description of request and relevance sets used

Number of requests	23
Average length of request (before any indexing operation)	14.7 words
Number of relevant documents	390
Average number of relevant documents per request	17.0 docs

4.2. Transcription used

The experiments reported in this paper use the manually generated transcriptions provided by NIST as the reference set for retrieval performance, two baseline transcriptions, Base 1 and Base 2, generated by NIST using instantiations of the SPHINX-3 recognition system provided by CMU, and our own HTK transcriptions. These document versions were mandatory for runs in the 1997 TREC-7 full SDR evaluation (Garofolo et al., 1999). Results are also reported for the transcriptions from Dragon (Allan et al., 1999), AT&T (Singhal et al., 1999), Sheffield University (Abberley et al., 1999) and DERA (second system) (Nowell, 1999) which we used as cross-recogniser conditions in the evaluation.

In this paper, we only describe our own transcription system, the description of the others can be found in (Voorhees and Harman, 1999).

4.2.1. The HTK transcription system

The transcription of spoken documents was done using part of our HTK broadcast news transcription system (Woodland et al., 1998).

The input data is presented to the system as complete episodes of broadcast news shows and are first converted to a set of segments for further processing (Hain et al., 1998). 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.

The HTK system uses cross-word context-dependent hidden Markov models (HMMs), in which the context dependent states are clustered using decision tree techniques and the state output distribution modelled using Gaussian mixtures (Young and Odell, 1994). The pronunciations for the system's 65k word vocabulary are taken from the LIMSI 1993 WSJ dictionary augmented by automatically produced pronunciations from a text-to-speech system along with hand corrections. The full HTK system (Woodland et al., 1998) operates in multiple passes and uses complex language models applied using lattice rescoring and quinphone HMMs. This system gave a word error rate (WER) of 16.2% in the 1997 DARPA Hub4 broadcast news evaluation.

The TREC-7 HTK SDR system uses the first two passes from the full system in a modified form to reduce computational requirements. The first pass uses gender-independent, bandwidth-dependent cross-word triphone models with a trigram language model to produce an initial transcription. This first-pass transcription is used to select the most likely gender of the speaker for each segment by alignment with gender specific models. The segments for each show (within gender and within bandwidth) are then grouped using output top-down covariance-based clustering algorithm (Johnson and Woodland, 1998). The first pass transcriptions are used to generate adaptation transformations for each segment cluster using maximum likelihood linear regression (MLLR) (Leggetter and Woodland, 1995; Gales and Woodland, 1996).

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 300-MHz Sun Ultra2 and achieves an error rate of 17.4% on the 1997 Hub4 evaluation data. It should, however, be noted that the error rates on Hub4 data and TREC data are not strictly

comparable: the TREC references are of a lower quality and the data is more difficult.

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 word-list was selected from the combined word frequency list, whilst ensuring that the number of new pronunciations which had to be created remained manageable. The final word-list 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 story-averaged Processed Term Error Rate (PTER) (Johnson et al., 1999a) (which more closely represents the error rate as seen by the retriever) of 34.6%.

4.2.2. Other transcriptions

Traditionally, 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 model the input to the retriever more closely, a term error rate (TER) has been introduced (Johnson et al., 1999a) which does not depend on word order and counts substitution errors (where one word is misrecognised as a different one) as two errors, since a correct word is missing and a spurious word is added.

This can also be calculated after preprocessing, to take into account the effects of stopping, stemming etc., and it has been shown that this PTER can offer a better predictor of retrieval performance than WER (Johnson et al., 1999b). The recogniser transcriptions used for the tests described in this paper, are therefore compared for WER, TER (original words) and PTER (after processing), as shown in Table 3, where TER and PTER are averaged over stories.

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 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, along with an extra stage to correct possible incorrect spellings in the manual transcriptions. These devices are *query-independent* and therefore were implemented as a text pre-processing phase on the queries and documents. The resulting index file was then generated. It notes 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), $tf(t_i, d_j)$.

We have also used one query-dependent device, which has been found of modest value with our test data. With our comparatively short requests there is no reason to consider within-query term frequency. But we have applied part-of-speech weighting to the query terms, using syntactic

Table 3
Error rates for the transcriptions

	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 4
Part-Of-Speech (POS) weights values

Proper noun	1.2
Common noun	1.1
Adjective and adverbs	1.0
Verbs and the rest	0.9

categories identified by applying a local tagger (Knight, 1998) to the source requests and weights defined by Table 4. These particular values were found optimal on the TREC-6 document set (Johnson et al., 1999b).

The document representations we have described can be exploited, according to the model of Spärck Jones et al. (1998), to give a document-query matching score and hence a search output ranking of retrieved documents. In our case the score takes account, in a simple way, of the query term part-of-speech weights. Thus we first use the model to define the combined weight $cw(t_i, d_j)$ for a query term t_i:

$$cw(t_i, d_j) = \frac{(\log N - \log n(t_i)) * tf(t_i, d_j) * (K + 1)}{K * (1 - b + b * ndl(d_j)) + tf(t_i, d_j)}, \tag{1}$$

$$n(t_i) = \sum_{d_i \in D} \begin{cases} 0 & tf(t_i, d_i) = 0, \\ 1 & tf(t_i, d_i) > 0, \end{cases} \tag{2}$$

$$dl(d_j) = \sum_{w \in V} tf(w, d_j), \tag{3}$$

$$ndl(d_j) = \frac{dl(d_j) * N}{\sum_{d \in D} dl(d)}, \tag{4}$$

where V is the term vocabulary for the whole document collection D and K and b are tuning constants. We then define the part-of-speech modified combined weight $pos\_cw(t_i, d_j)$ for t_i as $pos\_cw(t_i, d_j) = pos(t_i) * cw(t_i, d_j)$, where $pos(t_i)$ is the part-of-speech weight according to Table 4, and sum the values of pos_cw for all matching terms to obtain the final document score.

4.3.2. Adding Geographic Semantic Posets (GP)

Location information is very common in requests in the broadcast news domain. Our first extension of our TREC-7 system implements the expansion of geographic names occurring in the original query of the BL system into the list of their components, e.g.,

```
US → Alabama,...,Ohio,...,Wyoming
     Atlanta,...,New-York,...,
     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)

This 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 (Fellbaum, 1998) 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, as their proper handling is much more complicated than in the geographic names domain.

4.3.4. Adding Parallel Blind Relevance Feedback (PBRF)

Five different parallel corpora have been used for testing this method. They are subsets of the Primary Source Media (PSM) broadcast news transcriptions used to train the language model of our speech recognition system and were created from manually transcribed broadcast news collections covering the years 1993 to 1997. They do not overlap with the TREC-7 collection recording time (see Table 5).

In the feedback stage, when using the parallel corpus, the five terms with the highest Offer Weight were appended to the query. The Offer

Table 5
Parallel blind relevance corpora

Collection	Recording time	Number of stories
1993	January–December	31,515
1994	January–December	37,703
1995	January–December	41,364
1996	January–April	14,484
1997	January–May	18,629
Trec-7 SDR	June 1997–January 1998	2,866

Weight of a term t_i, following Spärck Jones et al. (1998), is

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

where B is the number of top documents 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 way expansion terms were exploited was motivated by the idea that the value of information embodied in Offer Weights based on Blind Feedback, and especially on Parallel Blind Feedback, is limited; so it should be viewed only as a modifier of a query term's basic weight, and a weight based on putative relevance data should not be substituted for the basic weight in the model style of Spärck Jones et al. (1998). But at the same time, given that the document data do not specify part of speech, this element of the weighting should be discarded for any original query term 'boosted' by appearing in the expansion set (new expansion terms could only have some 'neutral' part-of-speech weight of, say, 1, which could safely be ignored). All expansion terms, whether original query terms or new ones, were therefore weighted with $ow(t_i) * cw(t_i)$. At the same time, original terms not boosted into the expansion set on Offer Weight could be seen as so lacking in relevance information as to justify retaining their original weighting only. This rather complex set of assumptions needs further investigation. Caution in relying on the Blind Feedback data was similarly a reason for using only a small expansion set of five terms.

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 in order to reflect the smaller nature of the collection. This was perhaps over modest, but by choosing the best candidate, with a good chance of being helpful even on weak information, such limited expansion might be useful.

5. Results

The five systems described in Section 4.3 were evaluated on the TREC-7 SDR test data. Specifically, the comparison is across BL and BL progressively augmented using GP, WP, PBRF and BRF as shown in Table 6. There is a subsidiary comparison between the different corpora for PBRF.

The two evaluation measures which are used in this paper are Average Precision (AvP) and Precision at document cut-off 15 (P@15). In the case of AvP, Precision is calculated after each relevant document is retrieved (if no relevant document is retrieved, Precision is 0.0). Precision values are then averaged per query and query values in turn averaged over all queries. P@15 is the precision after 15 documents (whether relevant or non-relevant) have been retrieved, averaged over all queries (if 15 documents were not retrieved for a query, then all missing documents are assumed to

Table 6
Illustration of the query expansion process

	What diseases are frequent in Britain?
BL	{disease} {frequent} {britain}
+GP	{disease} {frequent} {britain, uk, united-kingdom, ..., cambridge}
+WP	{disease, flu, ..., malaria} {frequent} {britain, uk, ..., cambridge}
+PBRF	+ {cold} {blair} {rheumatism} {queen}
+BRF	+ {cow}

be non-relevant). These are standard measures: see Appendix A, Voorhees and Harman (1999).

There are some problems about the choice of a significance test for retrieval experiments, and our query sample is also rather small. However, for each performance comparison, we used the Wilcoxon Matched-Pair Signed-Ranks Test, which takes into consideration the ranks of the absolute differences of Average Precision between two system for all queries. When the Wilcoxon significance level (WSL) is lower than 5%, the difference is considered as significant, and the exact value of WSL is given.

We first consider performance for the expansion options without feedback as shown in Table 7, giving AvP for all transcriptions.

Table 7
Average Precision (AvP) on the TREC-7 test collection (results in %) for different transcriptions using the baseline system (BL), geographic posets (GP) and WordNet posets (WP)

Transcription	PTER	Average Precision		
		BL	BL + GP	BL + GP + WP
Manual	–	49.11	51.55	52.33
HTK	34.6	47.30	49.77	50.75
AT&T	39.7	44.84	47.47	48.39
Dragon	41.6	44.27	46.08	46.59
Base 1	48.5	42.95	45.09	46.53
Sheff	50.4	44.27	46.17	46.84
Base 2	68.9	33.95	35.71	36.26
DERA	93.0	38.70	39.74	40.47

The use of Semantic Posets, especially using both GP and WP, leads to a universal and almost constant improvement in AvP for all transcriptions. There is, for example, a 7.3% relative improvement (WSL = 2.9%) with our own HTK recogniser. In addition, Table 7 shows that PTER is a reliable predictor of retrieval performance when its value is low (up to 39.7%), but becomes unreliable for higher values (from 41.6% to 93.0%) as higher PTER values only sometimes lead to better AvP.

Table 8 shows that we can further improve performance with feedback using PBRF. Thus using corpus data for any individual year increases AvP for all the sets of transcriptions by between 0.57% and 7.67% absolute. For example, using the 1995 corpus, AvP for the manual transcriptions increases from 52.33% using BL + GP + WP alone (see Table 7), to 56.26% with PBRF added (WSL = 1.35%), from 50.75% to 54.01% with HTK (WSL = 4.15%), and from 36.26% to 41.77% with Base 2 (WSL = 1.9%).

These results suggest that the size of the parallel collection is a critical factor: using the years 1996 and 1997, with sizes of 14,484 and 18,629 documents, respectively, gives less improvement than using older collections containing between 31,515 and 41,364 documents. Moreover, the 1996 and 1997 corpora have comparable sizes and the AvP results show no significant difference (a maximum increase of 0.99% absolute for 1996 with AT&T

Table 8
Average Precision (AvP) on the TREC-7 test collection (results in %) for different transcriptions using the BL + GP + WP system augmented with PBRF based on different corpora

Trans.	BL	BL + GP	BL + GP + WP	BL + GP + WP + PBRF						
				Recording year of parallel corpus						
				1993	1994	1995	1996	1997	96 + 97	94 to 97
				Number of documents of parallel corpus						
				31,515	37,703	41,364	14,484	18,629	33,113	112,180
Manual	49.11	51.55	52.33	54.56	55.24	56.26	52.90	53.59	53.42	55.46
HTK	47.30	49.77	50.75	52.47	54.78	54.01	51.84	51.73	51.78	53.47
AT&T	44.84	47.47	48.39	51.83	52.87	53.83	51.63	50.64	50.15	53.50
Dragon	44.27	46.08	46.59	51.46	51.12	52.35	49.35	48.99	49.53	53.32
Base 1	42.95	45.09	46.53	49.98	51.44	52.22	48.65	48.03	47.76	51.54
Sheff	44.27	46.17	46.84	51.22	52.67	54.51	49.29	49.26	49.91	52.89
Base 2	33.95	35.71	36.26	39.56	40.97	41.77	39.26	40.13	38.79	42.15
DERA	38.70	39.74	40.47	44.56	43.54	43.25	43.67	44.22	43.31	47.33

and of 0.87% for 1997 with Base 2). In addition, the 1993 parallel corpus and the 1996 + 1997 corpus have a similar size, but although they are separated by a period of at least two years there is no significant difference of performance. Comparing the performance for years 1996, 1997 and their combination, we can also notice that a relatively small increase of size does not lead to any improvement. In addition, even a relatively large increase of parallel corpus size (last column of Table 8) does not lead to a significant improvement. Thus the overall implication is that if corpus size is more important than recency in this experiment, a reasonable amount of data seems sufficient for PBRF to reach its maximum efficiency. In order to obtain a more subtle analysis of the influence of size and recency of the parallel collection on IR performance, one would need to vary the number of terms added and the number of documents assumed relevant in the PBRF process. However, we have not taken experiments further since performance must be affected not only by, e.g., the number of documents assumed relevant, but by properties of the search collection which we do not have the test data to vary.

Table 9, on the other hand, shows that adding BRF as well does not have a significant effect on performance. For example, by adding BRF to the BL + GP + WP + PBRF'94 system with the HTK transcriptions, AvP is increased from 54.78% to 56.33% (not significant), but when using the Base 2 transcriptions, performance actually slightly decreases from 40.97% to 40.41%. While it might be thought the minimal expansion used with BRF explains these results, other experiments, not reported here, have shown that a sensible variation of the number of terms to be expanded by both PBRF and BRF does not produce much variation of performance. It thus seems that when PBRF is used, the addition of BRF, especially if this involves less good quality documents, can hardly improve a SDR system, though this needs checking on a larger test file.

The techniques we have examined can be assigned to two classes, namely knowledge-based (GP + WP) and statistics-based (PBRF + BRF). Fig. 2 shows the relative contributions of both these classes to the overall improvement of our SDR system. It is evident that the improvements are not crudely additive but that the two types of approach are partly complementary. It is possible that the statistics-based feedback techniques require a high precision at top rank levels. The knowledge-based approaches promote this precision and, through the better feedback information that follows, enhance the benefits of statistics-based techniques.

However, precisely how the two types of expansion interact, and what form their respective contributions take, is not clear. Table 10 shows a direct comparison between the two, for our own transcriptions, giving precision at document cut-off 15 as well as the AvP performance included in Fig. 2. The measures are very different, but though Feedback expansion appears to deliver more improvement over the Baseline than posets for AvP as opposed to P@15, there is in fact no statistically significant difference in either case. There are in

Table 9
Average Precision (AvP) on the TREC-7 test collection (results in %) for different transcriptions using the BL + GP + WP system augmented with PBRF based on different corpora and with BRF on the retrieval document collection

Trans.	BL	BL + GP + WP + PBRF'94	BL + GP + WP + PBRF + BRF				
			1993	1994	1995	1996	1997
Manual	49.11	55.24	54.85	56.47	56.14	54.81	55.88
HTK	47.30	54.78	53.69	56.33	53.79	54.42	55.08
AT&T	44.84	52.87	52.21	54.08	52.00	53.25	53.16
Dragon	44.27	51.12	52.09	52.39	51.64	49.35	50.86
Base 1	42.95	51.44	49.98	49.80	52.11	50.71	51.96
Sheff	44.27	52.67	50.95	52.72	52.93	52.20	51.97
Base 2	33.95	40.97	38.42	40.41	40.61	40.18	39.73
DERA	38.70	43.54	44.80	44.21	43.23	43.54	44.15

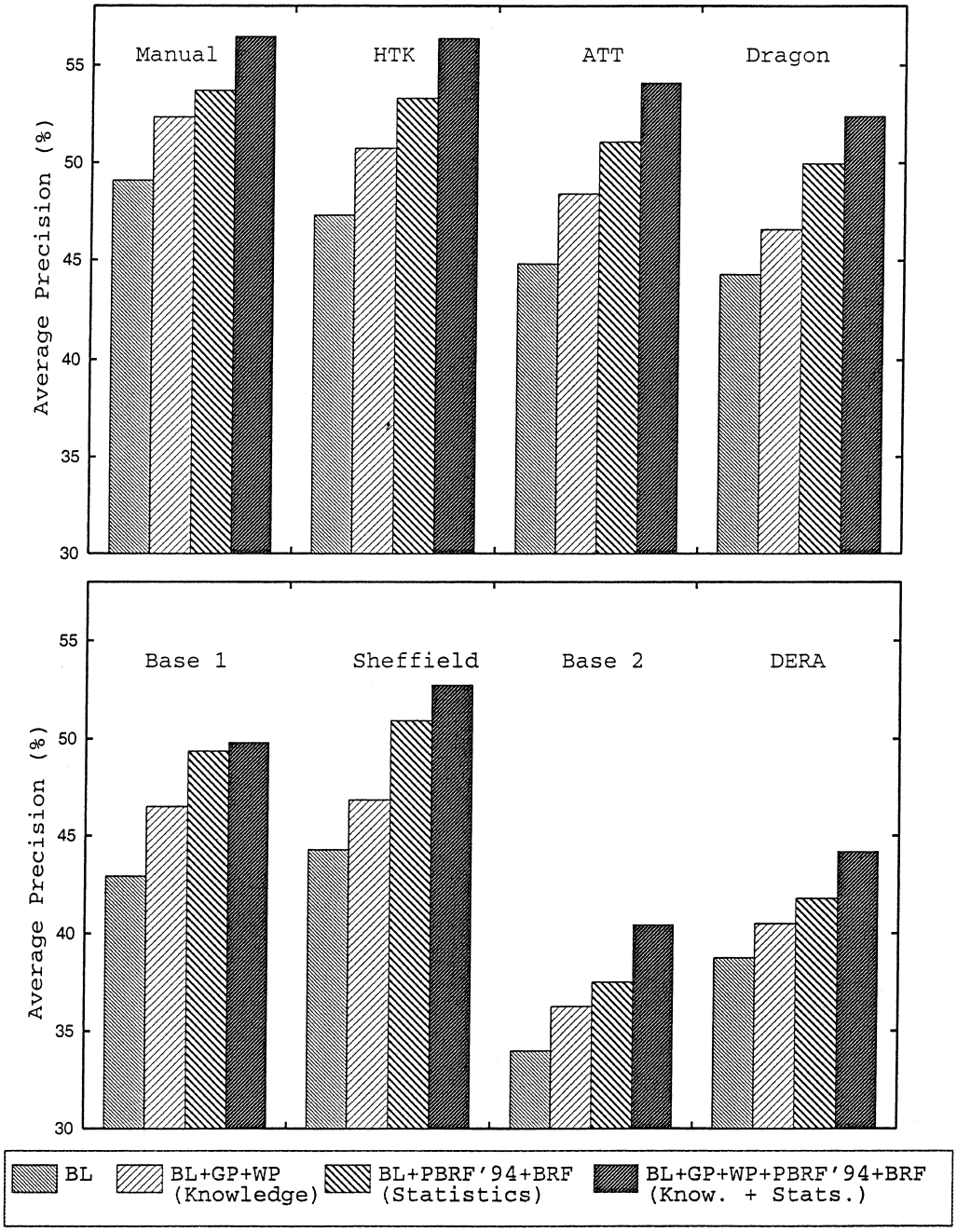

Fig. 2. Average Precision for different transcriptions and combinations of knowledge-based (WP + GP) and statistics-based (PBRF '94 + BRF) expansion techniques.

Table 10
Average Precision and Precision at document cut-off 15 on the TREC-7 test collection (results in %), for different transcriptions using the Baseline System (BL), the knowledge-based expanded system (BL + GP + WP) and the statistic-based expanded system (BL + PBRF '94 + BRF)

Trans.	AvP			P@15		
	BL	BL + GP + WP	BL + PBRF '94 + BRF	BL	BL + GP + WP	BL + PBRF '94 + BRF
Manual	49.11	52.33	53.68	40.29	44.06	42.03
HTK	47.30	50.75	53.28	39.13	42.32	43.48
AT&T	44.84	48.39	51.07	38.55	40.87	40.58
Dragon	44.27	46.59	49.97	37.97	41.45	41.74
Base 1	42.95	46.53	49.37	37.68	40.00	40.00
Sheff	44.27	46.84	50.93	37.39	39.71	41.74
Base 2	33.95	36.26	37.53	32.75	34.20	35.36
DERA	38.70	40.47	41.75	33.91	37.39	37.68

Table 11
Average Precision (AvP) on the TREC-7 test collection (results in %) for different transcriptions using the baseline system (BL) augmented with PBRF based on different corpora and with BRF

Trans.	1993	1994		1995	1996	1997
		1st Iter.	2nd Iter.			
Manual	52.82	53.68	53.61	53.00	53.34	53.46
HTK	51.74	53.28	50.29	51.55	52.36	52,53
AT&T	49.10	51.07	51.19	50.97	49.80	50.77
Dragon	49.91	49.97	49.80	50.54	50.20	48.21
Base 1	48.52	49.37	46.71	48.64	48.91	49.29
Sheff	48.94	50.93	48.88	51.47	50.20	50.28
Base 2	35.02	37.53	39.12	39.12	37.26	36.88
DERA	42.91	41.75	43.32	41.51	41.51	41.66

any case limits to the value of Feedback, since further iterations, as shown in Table 11, are not useful.

6. Conclusion and further directions

In this paper we have confirmed, in a larger range of experiments than previously reported, that the degradation of retrieval performance due to imperfect transcription can be relatively small. Thus for our baseline retrieval system, and our HTK recogniser WER of 25%, retrieval performance compared with that for the reference manual transcriptions shows a loss of only 3.7% relative in AvP (WSL = 19.7%).

Further, when recogniser accuracy is good and WER is low, it is possible to improve performance using different types of query expansion, based, on one hand, on thesaural methods, in Semantic Posets, and on the other on statistical methods, in Blind Relevance Feedback (BRF). When these types of expansion are combined, performance loss compared with the reference case is only 0.2% relative.

However with such a high WER as DERA's 61.5%, it appears to be much less easy to improve retrieval performance by using these devices. Thus while an AvP gain of 14% relative (WSL = 0.001%) can be obtained for these transcriptions, the substantial difference between automatic and manual transcriptions remains almost constant. The tests with the transcriptions from other recognition systems we have reported follow this pattern, suggesting that if recogniser output is reasonable (and taking manual transcription performance as an upper bound), it is more profitable to concentrate on applying good general retrieval devices than on specific adaptation to the speech data (e.g., by using expected term frequencies derived from ASR word lattices). Thus our tests show that with the HTK transcriptions the combined set of

expansion devices raises AvP performance from the baseline by 19% relative (WSL = 0.004%).

The available general techniques include document expansion, which Singhal and Pereira (1999) have shown also has the merit of offsetting higher WER, and which we have now begun to study (Johnson et al., 2000). The methods we have tested, moreover, represent only specific ways of applying the underlying general ideas, and need more investigation. Finally, like others working on spoken document retrieval, we have so far been confined to small test collections and have lacked a proper development collection: we intend to repeat our tests to see how our methods scale when larger collections become available.

Acknowledgements

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

References

Abberley, D., Kirby, D., Renals, S., Robinson, T., 1999. Retrieval of broadcast news documents with the THISL system. In: Voorhees, E.M., Harman, D.K. (Eds.), The Seventh Text REtrieval Conference (TREC-7). NIST Special Publication 500-242. Department of Commerce, National Institute of Standards and Technology, Gaithersburg, MD, pp. 181–190.

Allan, J., Callan, J., Sanderson, M., Xu, J., Wegmann, S., 1999. InQuery and TREC-7. In: Voorhees, E.M., Harman, D.K. (Eds.), The Seventh Text REtrieval Conference (TREC-7). NIST Special Publication 500-242. Department of Commerce, National Institute of Standards and Technology, Gaithersburg, MD, pp. 201–226.

Dushnik, B., Miller, E.W., 1941. Partially ordered sets. American Journal of Mathematics 63, 600–610.

Fellbaum, C., 1998. WordNet: An Electronic Lexical Database, ISBN 0-262-06197-X. MIT Press, Cambridge, MA.

Fox, C., 1992. Lexical analysis and stoplists. In: Frakes, W.B., Baeza-Yates, R. (Eds.), Information Retrieval: Data Structures and Algorithms. Prentice-Hall, Englewood Cliffs, Chapter 7, pp. 102–130.

Gales, M.J.F., Woodland, P.C., 1996. Mean and variance adaptation within the MLLR framework. Computer Speech and Language 10, 249–264.

Garofolo, J.S., Voorhees, E.M., Auzanne, C.G.P., Stanford, V.M., 1999. Spoken document retrieval: 1998 evaluation and investigation of new metrics. In: Proceedings of the ESCA Workshop: Accessing Information in Spoken Audio. Cambridge University, Cambridge, pp. 1–7.

Hain, T., Johnson, S.E., Tuerk, A., Woodland, P.C., Young, S.J., 1998. Segment generation and clustering in the HTK broadcast news transcription system. In: Proceedings of 1998 DARPA Broadcast News Transcription and Understanding Workshop. pp. 133–137.

Johnson, S.E., Woodland, P.C., 1998. Speaker clustering using direct maximisation of the MLLR-adapted likelihood. In: Proceedings of the Fifth International Conference on Spoken Language Processing. Vol. 5. pp. 1775–1779.

Johnson, S.E., Jourlin, P., Moore, G.L., Spärck Jones, K., Woodland, P.C., 1999a. The Cambridge University spoken document retrieval system. In: Proceedings of the 1999 IEEE International Conference on Acoustics, Speech and Signal Processing. Vol. 1. pp. 49–52.

Johnson, S.E., Jourlin, P., Moore, G.L., Spärck Jones, K., Woodland, P.C., 1999b. Spoken document retrieval for TREC-7 at Cambridge University. In: Voorhees, E.M., Harman, D.K. (Eds.), The Seventh Text REtrieval Conference (TREC-7). NIST Special Publication 500-242, Department of Commerce, National Institute of Standards and Technology, Gaithersburg, MD, pp. 191–200.

Johnson, S.E., Jourlin, P., Spärck Jones, K., Woodland, P.C., 2000. Spoken document retrieval for TREC-8 at Cambridge University. In: Voorhees, E.M., Harman, D.K. (Eds.), The Eighth Text REtrieval Conference (TREC-8). NIST Special Publication, Department of Commerce, National Institute of Standards and Technology, Gaithersburg, MD.

Knight, S.F. (n.d.), 1998. Personal communication.

Leggetter, C.J., Woodland, P.C., 1995. Flexible speaker adaptation using maximum likelihood linear regression. In: Proceedings of ARPA 1995 Spoken Language Technology Workshop. Morgan Kaufmann, Los Altos, CA, pp. 110–115.

Mandala, R., Tokunaga, T., Tanaka, H., Okumura, A., Satoh, K., 1999. Ad hoc retrieval experiments using wordnet and automatically constructed thesauri. In: Voorhees, E.M., Harman, D.K. (Eds.), The Seventh Text REtrieval Conference (TREC-7). NIST Special Publication 500-242, Department of Commerce, National Institute of Standards and Technology, Gaithersburg, MD, pp. 475–480.

Mitra, M., Buckley, C., Singhal, A., Cardie, C., 1997. An analysis of statistical and syntactic phrases. In: Proceedings of Recherche d'Information Assistée par Ordinateur (RIAO '97, Montreal). Centre de Hautes Études Internationales d'Informatique Documentaire, Paris, pp. 200–214.

Nowell, P., 1999. Experiments in spoken document retrieval at DERA-SRU. In: Voorhees, E.M., Harman, D.K. (Eds.), The Seventh Text REtrieval Conference (TREC-7). NIST Special Publication 500-242, Department of Commerce, National Institute of Standards and Technology, Gaithersburg, MD, pp. 353–362.

Porter, M.F., 1980. An algorithm for suffix stripping. Program 14, 130–137.

Salton, G., Lesk, M.E., 1971. Computer evaluation of indexing and text processing. In: Salton, G. (Ed.), The SMART

Retrieval System: Experiments in Automatic Document Processing. Prentice-Hall, Englewood Cliffs, pp. 143–180.

Singhal, A., Pereira, F., 1999. Document expansion for speech retrieval. In: Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval. pp. 34–41.

Singhal, A., Choi, J., Hindle, D., Lewis, D.D., Pereira, F., 1999. AT&T at TREC-7. In: Voorhees, E.M., Harman, D.K. (Eds.), The Seventh Text REtrieval Conference (TREC-7). NIST Special Publication 500-242, Department of Commerce, National Institute of Standards and Technology, Gaithersburg, MD, pp. 239–252.

Spärck Jones, K., 2000. Further reflections on TREC. Information Processing and Management 36, 37–85.

Spärck Jones, K., Walker, S., Robertson, S.E., 1998. A probabilistic model of information retrieval: Development and status. TR 446, Computer Laboratory, University of Cambridge.

Voorhees, E.M., 1994. 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.

Voorhees, E.M., Harman, D.K. (Eds.), 1999. The Seventh Text REtrieval Conference (TREC-7). NIST Special Publication 500-242, Department of Commerce, National Institute of Standards and Technology, Gaithersburg, MD.

Woodland, P.C., Hain, T., Johnson, S.E., Niesler, T.R., Tuerk, A., Whittaker, E.W.D., Young, S.J., 1998. The 1997 HTK broadcast news transcription system. In: Proceedings of DARPA Broadcast News Transcription and Understanding Workshop. pp. 41–48.

Young, S.J., Odell, J.J., Woodland, P.C., 1994. Tree-based state tying for high accuracy acoustic modelling. In: Proceedings of 1994 ARPA Human Language Technology Workshop. Morgan Kaufmann, Los Altos, CA, pp. 307–312.