Experiments in Spoken Document Retrieval at DERA-SRU

P. Nowell
Speech Research Unit
DERA Malvern
St Andrews Road
Malvern
Worcs, WR14 3PS

Introduction

A small amount of internal funding allowed DERA-SRU to participate in the TREC-7 SDR evaluations for the first time this year. Since we had almost no experience of entering this or related NIST evaluations (e.g. ARPA HUB-4 LVCSR) there was a rather steep learning curve along with intense development of the experimental infrastructure. The intention was to generate a base for future participation and to build upon this using experience gained from related work on topic spotting. To this end, a straightforward (i.e. non-optimised) speech recogniser was used to generate transcripts and retrieval was performed using the *okapi* [6,9] search engine. Previous work on topic spotting [7] suggested that term expansion using a semantic network (in this case *wordnet* [2,3]) might be useful. This hypothesis appeared to be supported by preliminary work on TREC-6 SDR data which yielded text (i.e. R1) results that were comparable with the best achieved elsewhere.

Speech Recognition

Two sets of speech recognition transcripts were generated. The first set (S1) was generated quickly using available acoustic and language models with aim of developing and validating the necessary infrastructure. Subsequently a second set of transcripts (S2) was generated using the same acoustic and language models but with the recognition process beginning to be tailored to the task.

The first set of speech recognition transcripts (S1) was generated using a continuous large vocabulary speech recogniser constructed using pre-existing (i.e. not Hub4 task specific) acoustic and language models. The language model consisted of a 50000 word vocabulary and 500 word class clustered bigram model which had been trained using the North American News Transcripts (NANT) corpus. A single set of 2548 full-bandwidth (i.e. 8kHz) 8 mixture component triphone models were used along with 45 single mixture component fast-match monophone models. The acoustic models had been trained using the SI284 Wall Street Journal corpus.

The speech recogniser used a 25 channel mel-scale filterbank from which 12 cosine terms (C1 ... C12) and an energy value were computed. The inclusion of delta and delta-delta gave a 39 element feature vector. No online adaptive noise masking or channel normalisation algorithms were used. A summary of the official NIST results for S1 are given in table 1.

	Corr	Sub	Del	Ins	Err	S.Err
Sum/Avg	39.3	47.4	13.3	5.6	66.4	99.8

Table 1; NIST results for S1 speech recognition transcripts

In light of the S1 results, it was decided to generate a second set of recognition transcripts. A public-domain copy of the CMU speech segmenter (CMUSeg_0.4) was obtained from NIST. The scripts and code were modified slightly to handle files containing speech from a single focus group and occasional arithmetic underflows. Otherwise, the scripts were used as given with the default parameter settings (i.e. those used by CMU for the 1996 Hub-4 evaluations).

The speech was then segmented prior to recognition into focus groups F0 (full-bandwidth) and F2 (telephone bandwidth). Each segment was then recognised separately using either full-bandwidth or telephone bandwidth models (again trained using SI284 Wall Street Journal) and the recognition transcripts concatenated. Online adaptive noise masking and channel normalisation were also used. N-best recognition results were generated (using a depth cut-off of 20) and these were then rescored using a trigram language model also trained using the NANT corpus. The trigram rescorer generated a further set of N-best lattices of which only the first choice were used in subsequent retrieval experiments. The official NIST results for S2 are summarised in Table 2.

	Corr	Sub	Del	Ins	Err	S.Err
Sum/Avg	47.3	44.8	7.9	8.8	61.5	99.8

Table 2; NIST results for S2 speech recognition transcripts

The average word error rate is still high at 61% due mainly, we believe, to the use of non-task-specific acoustic and language models along with relatively tight pruning thresholds. Given time these results could be improved significantly by developing optimised models and relaxing pruning constraints to more normal levels. However we do not believe that optimising recognition performance should be a major aim for these particular evaluations.

Both recognition runs were carried out on a bank of 200Mhz Pentium Pros with either 64Mb or 128Mb of local memory. In each case the test data was divided up and assigned to one of seventeen processing units. Unfortunately timing information was not kept for the first recognition runs although it is estimated that recognition would have taken around 10 times real-time on a single CPU. The second recognition was timed and in this case the recogniser ran at approximately 22 times real-time on each CPU.

Information Retrieval

During the course of early SRU work on topic spotting it was proposed that a semantic network be constructed and used to generate keywords from a topic descriptor or vice-versa [5]. Unfortunately, due to funding constraints, it was not possible to follow this line of research further at that time. It was therefore decided

that this proposal would be investigated further under the realm of information retrieval. Term expansion using *wordnet* has also been briefly explored in previous TREC evaluations [4,5,8].

The text retrieval test set contains 23 queries, one of which is shown in figure 1. The function of the retrieval engine is to take such queries and generated a ranked list of up to 1000 matching section. Note however the mismatch between the query, which requests a list of cities, and the evaluation which requires a ranked list of episode sections.

```
<num> Number: 55

<desc> Description:
What cities other than Washington D.C. has the First Lady
visited on official business (i.e., accompanying the President
or addressing audiences/attending events)?
```

Figure 1; Example of a spoken document retrieval (SDR) query

The query text is syntactically tagged and keywords are selected on the basis of their part-of-speech (POS) tags. This largely avoids the need for an explicit stop-list as well as helping to reduce the amount of over-generation during term-expansion.

The syntactic tagger, known as LTPOS [12], contains three major components: a tokeniser, a morphological classifier and a morphological disambiguator. The tokeniser segments the input string into words and sentences which are then classified according to a range of morpho-syntatic features (number, case, gender, etc.). Each feature set is unambiguously mapped to one or more part-of-speech (POS) tags. Unknown words are catered for by a rule set which attempts to guess the POS tag(s). Finally, words that have been assigned more than one POS tag (e.g. 'books' which can be a noun or a verb) are probabilistically disambiguated using a pre-trained Hidden Markov Model. Accuracy on known words is reported to be 96-98% and on unknown words 88-92%.

Noun and verb groups are also identified by means of a syntactic chunker or partial parser. The parser uses the POS information provided by the tagger and mildly context-sensitive grammars to detect syntactic boundaries. The output consists of simple noun groups (e.g. [[Washington D.C]]) and verb groups (e.g. ((have occurred))).

```
What_WP [[ cities_NNS  ]]other_JJ than_IN [[ Washington_NNP
D.C._NNP ]]( has_VBZ  ))[[ the_DT First_NNP Lady_NNP  ]](
visited_VBD ))on_IN  [[ official_JJ business_NN
]](_(i.e._FW,_,accompanying_VBG [[ the_DT President_NNP  ]]
or_CC addressing_VBG[[ audiences_NNS/_NN ]]attending_VBG
[[ events_NNS ]])_)?_.
```

Figure 2; Example of typical output from the syntactic tagger / chunker

The tags are used to extract a set of keywords and keyphrases for the retrieval engine. This is achieved by simply extracting words with tags that are likely to be discriminative (In these experiments JJ, NN, VB[DGNP], RB, PRP, MD and WRB were

used). A small ad-hoc stop list, containing just under 20 words, was also used to remove mostly common verbs such as ‘are’, ‘be’, ‘do’, ‘get’, ‘have’, etc.

The output from the tagger is also processed to extract compound nouns and adjectival phrases. Where phrases contain three or more words the words are successively removed from the left hand side to yield progressively shorter sub-phrases (e.g. ‘military air crash’ would also yield ‘air crash’). Such a simple process however does occasionally generate erroneous sub-phrases. Figure 3 shows the basic keywords and phrases extracted from the query shown in figure 1.

```
<keywords>
cities Washington D.C. First Lady visited official business
accompanying President addressing audiences attending events

<phrases> Compound Nouns:
Washington D.C.
First Lady

<phrases> Adjectival Phrases:
official business
```

Figure 3; Example of keywords and phrases extracted from a tagged query

Term expansion is performed using *wordnet*, a semantic network originally developed for testing psycholinguistic theories of human lexical memory [2,3]. This program contains a network of nouns, verbs, adjectives and adverbs which is organised into synonym sets representing different underlying lexical concepts. Entering a single word produces a list of related words and phrases at various levels of abstraction. *Wordnet* has been used in the TREC-6 but no details appear to have been published. A related concept of semantic forests [10] has also been used to infer topic categories from keywords.

A major problem in any term-expansion scheme is one of over generalisation. This problem has been reduced to some extent due to knowledge of lexical POS tags. Function words are ignored and ambiguous keywords (such as ‘issues’ which can be a noun or a verb) are generally expanded in the correct manner. Even so, there are still a large number of options by which nouns, verbs and adjectives may be expanded. A brief examination on the previous year’s data showed that most options resulted in overgeneration and/or the production of words that were unlikely to have any beneficial effect on retrieval. Only three options were used for keyword expansion, these being the senses of nouns, senses of verbs and hypenyms of nouns.

Furthermore, only the most frequent sense of any word was used in the hypenym expansion and terms were only taken from the top-most (i.e. least abstract) expansion. Given these constraints and the tagged query in figure 3, the following expansion terms were generated.

<sensn> Senses	<hypen> Hypenyms	<sensv> Senses
city	municipality	have
metropolis	national capital	have got
urban center	federal district	hold
lady	rank	visit
business	woman	see
concern	adult female	attach to
business concern	enterprise	attend
business organization	corporate executive	accompany
audience	business executive	come with
event	gathering	go with
	assemblage	address
		speak to
		turn to
		go to

Table 3; Wordnet expansion of a tagged query

Text retrieval was performed using Okapi, a probabilistic retrieval engine developed by City University and the Polytechnic of Central London. Okapi has been used extensively at previous TREC evaluations by City University as well as other participants [6,9]

All texts were first indexed using default parameter settings i.e. strong stemming and an empty GSL file. A GSL file can be used to specify terms that are dealt in a special way by the indexing processing such as stop terms or words that should be indexed as a single item. Searches were performed using the standard BM25 weighting function reported in TREC-3 [6] with the default parameter settings as used by City in TREC-6 [9].

An off-line retrieval engine was developed around the Basic Search System (BSS) library. This engine is similar to the test-engine provided with the *okapi-pack* distribution but has been modified to enable searches for co-adjacent words (e.g. 'First Lady'). The manner in which multi-word search items are handled is however different from that used by the interactive version of Okapi. In the former case the words must be strictly co-adjacent whereas in the latter the words only have to occur in the same sentence or paragraph. Otherwise the search engine is completely standard with the default weighting functions being used for both single word and multi-word items.

Three retrieval experiments were run, the first using keywords only (KW), the second keywords and phrases (KWP) and the last using keywords, keyphrases and *wordnet* expansions (KWPE). As it was only possible to submit one set of retrieval results for official scoring the latter (KWPE) were submitted.

Official TREC-7 SDR Results

A summary of the official retrieval results for KWPE (keywords, phrases and wordnet expansion) is presented in figure 4. As described previously, two sets of recognition transcripts were generated.

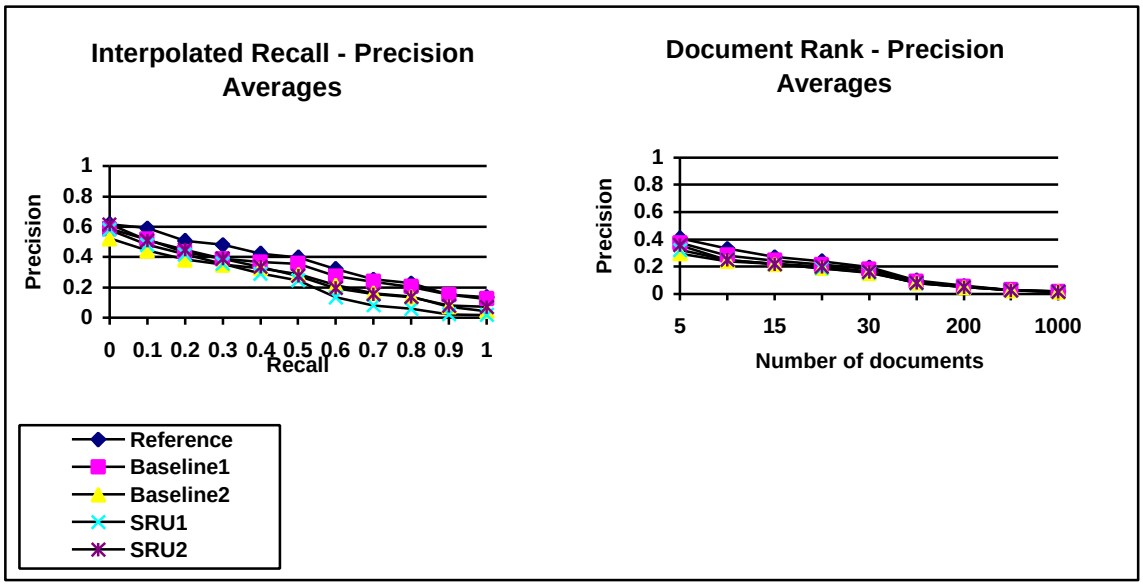

Figure 4; Official TREC-7 SDR summary results

The results show that there does not appear to be a great difference in retrieval performance even though the word error rate ranges from 0% to 66%. It is also interesting that results using SRU2 appear to be slightly better than with the second baseline (B2) although the word error rate is substantially higher.

The following graph compares results on the reference transcripts and therefore shows differences due to the retrieval strategies used by the various sites. Results using recogniser transcripts are likely to be similar but with lower absolute precisions.

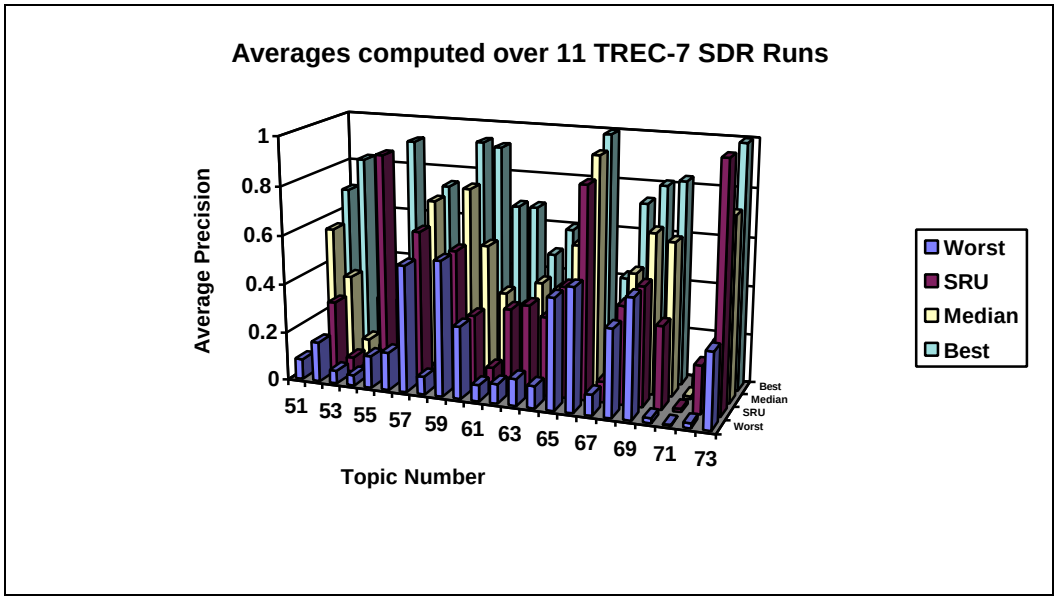

Figure 5; Official TREC-7 SDR results for each topic

Of the 23 topics three have the highest precision, two are above the median, twelve are below the median and six have the lowest precision. Topic 55 is particularly interesting since in this our results have the highest precision and are also significantly higher than the median. One would expect that this is due to *wordnet* expansion yielding one or more particularly useful keywords or phrases. Further investigation is required to determine whether this is indeed the case.

Examination of the queries revealed that only two words were unknown to the speech recogniser, these being 'paparazzi' (topic 60) and 'Trie' - a Chinese name (topic 64). Paparazzi occurs frequently in the 'perfect' (LTT) transcripts whereas Trie does not appear at all. Unsurprisingly our retrieval results are poor on topic 60 and term expansion is unable to help since 'paparazzi' is also unknown to *wordnet*.

Unofficial TREC-7 SDR Results

Figure 6 shows the performance of different retrieval strategies according to the condition and word error rate (WER). Labels along the x-axis refer to the source of the transcripts and associated word error rate (R1=Reference, CU=Cambridge University, DG=Dragon, B1,B2=Baselines 1 and 2, DS=DERASRU S2 and S1). The key labels refer to the source of the retrieval results, these being University of Massachusetts, Sheffield, Cambridge University, University of Maryland and DERASRU. The last key 'Simple' shows the results that we would have obtained by simply using keywords i.e. omitting keyphrases and term expansion (results using keyphrases are virtually identical).

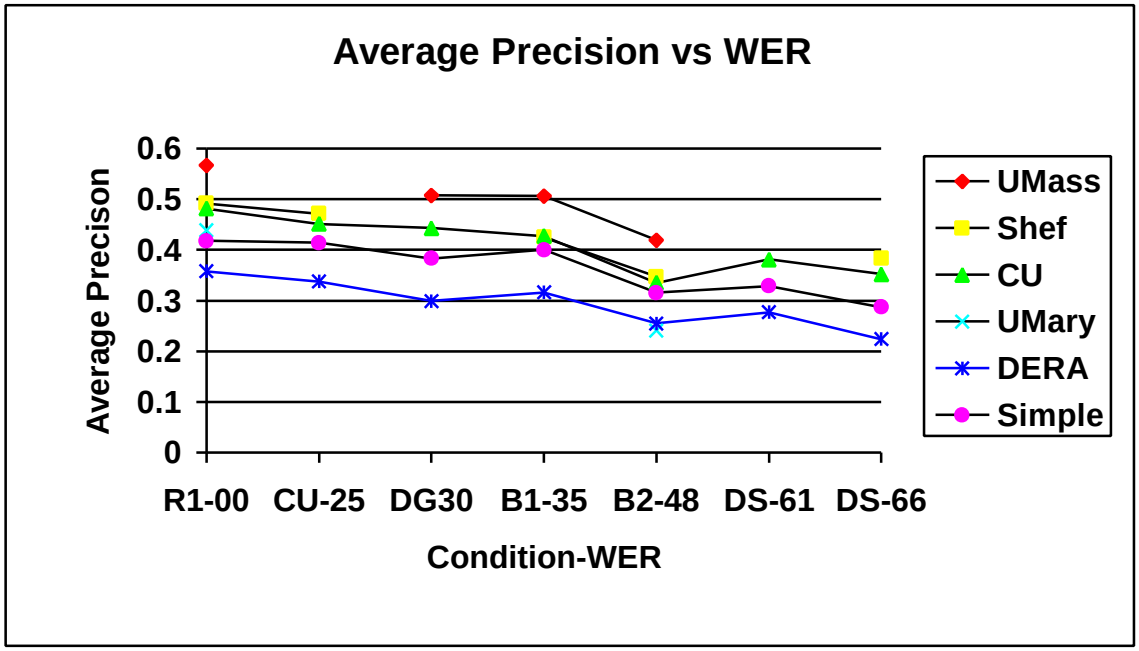

Figure 6; Unofficial comparative TREC-7 SDR results

The graph, which only represents those retrieval results known at the time of writing this report (Oct. 1998), shows that UMass appears to be better and DERA unfortunately worse than average. It is interesting and also disappointing that the

simple keyword-only approach works consistently better than the more complex approach involving term expansion. Furthermore, retrieval performance using DS-61 is consistently higher than with B2-48 suggesting that not all errors are created equal.

Taking figures 5 and 6 together it seems that term expansion using *wordnet* can occasionally be of great benefit but more often than not leads to a degradation in performance. Further investigation is required to uncover the reasons for the performance shortfall either using simple keywords or keywords, phrases and term expansion.

Summary

Although this is the first time that DERA has participated in such evaluations it is nevertheless disappointing that the current performance is less than that of more experienced participants. Given time, the performance of the speech recogniser could be easily improved through the use of task-specific acoustic and language models. There is also scope for improvement of the information retrieval component as shown by the keyword only results. It is however somewhat surprising that term-expansion should lead to such a large and consistent decrease in performance. However, now that the basic infrastructure is in place the performance of both components should improve quite rapidly.

Of more interest for the future, are the application of techniques that address the specific problems of spoken document retrieval. When DERA first entered the evaluation the intention was to try and make use of phoneme based techniques developed for topic spotting. It is believed that these will be of benefit where the query contains out-of-vocabulary words and / or the recogniser incorrectly recognises significant words. Time constraints meant that it was not possible to carry out these investigations this time around. It is hoped that these and other techniques will be investigated as part of future evaluations although funding sources have yet to be identified.

References

[1] S.E. Robertson and K. Sparck Jones, Relevance weighting of search terms, Journal of the American Society for Information Science 27, May-June 1976, p129-146.
[2] G.A. Miller, Wordnet: A dictionary browser, Proc. of the first conference of the UW centre for the New Oxford dictionary, Univ. of Waterloo, Canada, 1985.
[3] G.A. Miller et al. Introduction to wordnet: an on-line lexical database, Princeton University, Aug. 1993.
[4] E. Voorhees and Y.W. Hou, Vector Expansion in a Large Collection, NIST special publication 500-207, 1993.
[5] E. Voorhees, On Expanding Query Vectors with Lexically Related Words, NIST special publication 500-215, 1994.
[6] S.E. Robertson et al, Okapi at TREC-3, NIST special publication 500-236, 1995.
[7] P. Nowell, Topic Spotting Progress Report, DRA Memorandum DRA/CIS(SE1)/374/04/PR1/1.0, March 1996.
[8] A. Smeaton et al, TREC-4 Experiments at Dublin City University: Thresholding Posting Lists, Query Expansion with Wordnet and POS Tagging of Spanish, NIST special publication 500-236, 1996.

[9] S. Walker et al, Okapi at TREC-6: Automatic ad hoc, VLC, routing, filtering and QSDR, Proceedings of the Sixth Text Retrieval Conference (TREC-6), NIST special publication 500-240, 1998.
[10] P. Schone et al, Text Retrieval via Semantic Forests, Proceedings of the Sixth Text Retrieval Conference (TREC-6), NIST special publication 500-240, 1998.
[11] E.M. Vorhees and D. Harman, Overview of the Sixth Text Retrieval Conference (TREC-6), NIST special publication 500-240, 1998.
[12] University of Edinburgh, Language Technology Group, 2Buccleuch Place, Edinburgh, EH8 9LW, (http://www.ltg.ed.ac.uk)

Addendum

Following receipt of the official SDR submission results it has been possible to devote some time to investigating areas where the retrieval approach could be improved. This has involved analysis of performance on the test data as well as testing of alternative stop-lists and parameter settings. There is therefore an element of training on the test data. However, the intention is that any modifications should be consistent with those that might had been made given sufficient time and experience of previous evaluations. Incremental results using keywords only on hand-transcripts (R1) are shown in table 4.

Previously it was described how keywords were selected based upon POS tags and a small stop-list containing less than 20 words. Replacing this small list by the standard 'van Rijsbergen' stop-list gives a small improvement when applied to the queries only and a larger improvement when also applied prior to database indexing.

Examination of the queries and retrieval output revealed a problem with the handling of abbreviations. One of the queries contains the single term 'U. S.' which is represented in all transcripts as two separate words, i.e. ' U. S. '. Simple attempts at concatenating such words leads to problems since the *okapi* pre-processor substitutes 'us' which is in the stop-list and thereforeunindexed. Removing 'us' from the stop-list would of course create other problems. Rewriting all abbreviations in transcripts and queries to the form 'UxxSxx, DxxCxx etc.' overcomes these problems and gives a small increase in performance.

It has been shown previously (e.g. [11]) that using templates of the form 'What data | information is availableon ....' to prune non-topic-descriptive words from queries is advantageous. Table 4 shows that query pruning also gives a relatively large increase in performance on this years data.

Alternative parameters settings for the BM25 weighting function [9] have also been tested but these were not found to give a clear advantage.

Average Precision (R1)	Modification
0.4179	Official / baseline results
0.4216	Stoplist applied to queries only
0.4334	Stop-list applied to queries and database
0.4349	Proper treatment of abbreviations
0.4516	Query pruning using templates

Table 4; Effects of modifications on average precision

The above changes have also been incorporated into the full system employing keywords, keyphrases and *wordnet* expansion which was submitted to NIST. In addition weighting factors [4,5,8] have been added to each term with values of 1.0 for keywords and keyphrases and 0.5 for terms arising from*wordnet* expansion. The effect is de-weight expanded terms which may be less relevant those obtained from the original query.

Figure 7 shows comparative results on the standard R1, B1, B2, S1 and S2 transcripts. No cross-recogniser results have been generated at this time (October 1998).

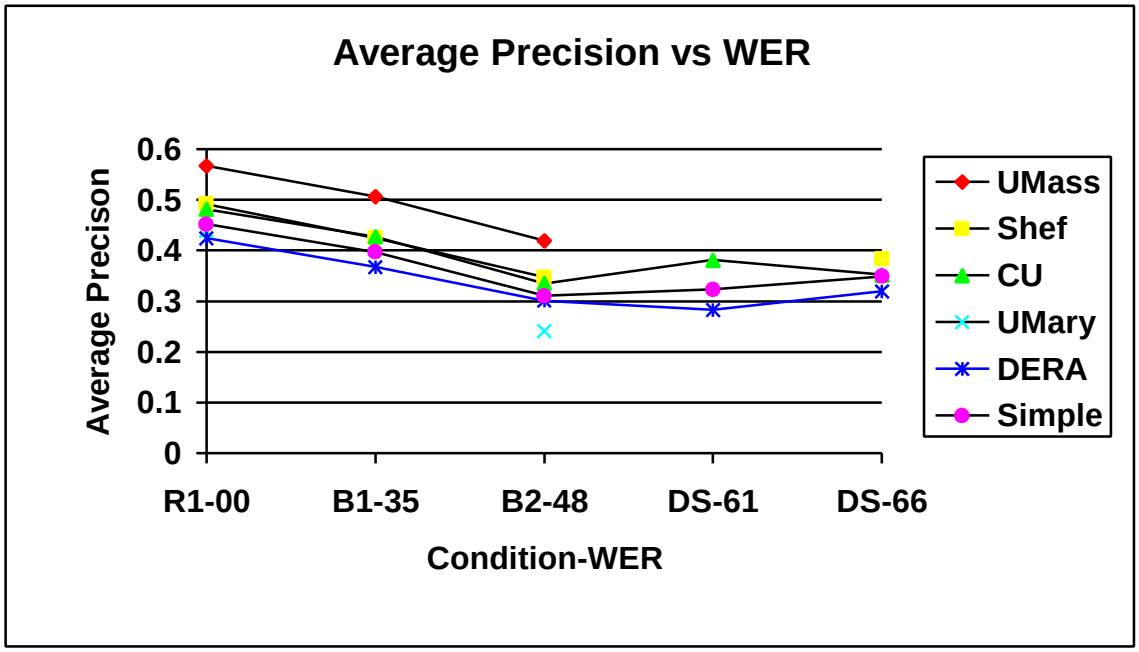

Figure 7; Second iteration unofficial TREC-7 SDR results

Keyword only results (Simple) are improved on R1-00 and DS-66 and are closeto the median performance level. There is however still a small performanceshortfall which shows that there is more work to be done. Although term-expansion still leads to a fall in performance the average precision has been greatly increased. Both sets of results of results could undoubtedly be improved further given time and resources.