Effects of Out of Vocabulary Words in Spoken Document Retrieval*

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

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

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

Abstract

The effects of out-of-vocabulary (OOV) items in spoken document retrieval (SDR) are investigated. Several sets of transcriptions were created for the TREC-8 SDR task using a speech recognition system varying the vocabulary sizes and OOV rates, and the relative retrieval performance measured. The effects of OOV terms on a simple baseline IR system and on more sophisticated retrieval systems are described. The use of a parallel corpus for query and document expansion is found to be especially beneficial, and with this data set, good retrieval performance can be achieved even for fairly high OOV rates.

1 Introduction

Recently there has been much interest in IR from audio sources such as television and radio broadcasts. The standard approach to this problem is to use a large vocabulary continuous speech recognition (LVCSR) system to automatically generate word-level transcriptions and then apply standard text IR techniques to the transcripts.

One of the major issues with the large vocabulary transcription approach to SDR is the effect on retrieval of words in the audio stream outside the particular recognition vocabulary used, i.e. out-of-vocabulary (OOV) words. This is clearly an important problem since however large the recognition vocabulary is made, there will always be new words (e.g. new names) that will occur in the input audio stream. The aim of this paper is to investigate the scale and nature of these effects on SDR performance for a particular recogniser.

The paper first discusses the effects of OOV on speech recognition and SDR performance. An extensive set of experiments using the TREC-8 SDR audio corpus and query sets have been performed to evaluate retrieval performance using sets of automatic transcriptions with varying OOV rates. Performance is given for a basic retrieval engine and for variants enhanced using query and document expansion from parallel newspaper text corpora since, independent of other potential benefits, expansion can compensate for transcription errors caused by OOV.

*This work is in part supported by EPSRC grant GR/L49611

To appear in SIGIR 2000, July 2000, Athens, Greece

2 OOV Effects in LVCSR and IR

The speech recogniser interprets each audio segment as a sequence of items from the recognition vocabulary. Even assuming there are no errors due to other causes, the recogniser will always make at least one word error whenever an OOV word is spoken. Therefore OOV words are *missing* from automatically transcribed spoken documents and are *replaced* by alternatives that are probable given the recognition acoustic model of phone realisations as acoustic features, the allowable phone sequences that constitute vocabulary items and the recognition language model of word sequence probabilities. The size and nature of the recogniser vocabulary is clearly the crucial factor. In order to minimise the expected OOV rate on new data, the vocabulary is chosen by taking the most frequent word forms from a large text corpus.

For IR, it is especially important if query words are OOV (QOV items). QOV words may have in fact occurred in the original spoken audio and not been recognised, which leads to a word *miss* in searching. (and potentially also a word *false drop* with its replacement on other occasions). The retrieval damage from misses depends on such factors as query length and word frequencies across, or in, documents, and the frequency across queries. Overall, QOV misses will affect both Precision and Recall by reducing the number of matching words.

Stemming further complicates matters. The speech recognition vocabulary is defined using *full word forms*. Therefore if one form is missing, another may be included and if a word form is misrecognised as a variant, due to acoustic and semantic similarity, such OOV transcription errors will not have any impact on retrieval performance.

Further issues arise with query expansion using terms from documents. In principle there can be miss/false drop effects on expansion sets. However in practice the effects may be somewhat mitigated because expansion prefers 'well-supported' terms, i.e. ones for which there is cross-document evidence, and these are less likely to be OOV ones (if the OOV rate is low).

The potential consequences of OOV for retrieval are quite complex and can be summarised as

(a) OOV query terms which directly affect matches either through lower scores for otherwise matching documents or through failure to match at all;

(b) Missing term relations, which affect expansion sets through failure of 'demand' from the query or 'supply' from the documents;

(c) Spurious in-vocabulary (IV) document terms caused by the recogniser incorrectly substituting for OOV words, leading to false matches with IV query terms.

3 Speech Transcriptions

The TREC-8 audio contains 500 hours of US broadcast news data that was recorded between February and June 1998. We processed this data to automatically detect and remove commercials [2]. Transcription used a simplified version of the HTK broadcast news system which corresponds to the "first-pass" recognition system described in [2]. The system automatically segments the audio into acoustically homogeneous chunks for transcription. The speech recogniser uses cross-word context-dependent hidden Markov models and a 4-gram language model trained on broadcast news data.

To investigate OOV effects, five different sized recognition vocabularies were constructed. Starting from a 55k word vocabulary the size was successively halved so that the smallest vocabulary tested contained only 3k words, at each stage retaining the most frequently occurring words. For each vocabulary size, the entire 500 hours of TREC-8 data transcribed. The system runs in 2.5-3x real time and hence computing the 5 complete sets of TREC-8 transcriptions took 10 CPU-months.

3.1 Closed Caption Transcriptions

Manually generated closed-caption transcriptions (`cc`) were provided by NIST for the full 500 hours of test data.[1] After text-normalisation, these transcriptions had a word error rate of 8.8% and hence offer a reasonable approximation to an accurately transcribed document collection for use as a reference is assessing SDR performance.

3.2 Transcription Quality

A high quality manual transcription of a 10 hour subset of the TREC-8 SDR corpus [1] was used to determine the error rates of the recogniser. The word error rate (WER); out of vocabulary word rate (OOVW); query out of vocabulary word rate (QOVW) for the 684 original query words; and size of wordlist are given in Table 1.

ID	# Words	WER	OOVW	QOVW
3k	3,413	43.9	14.1	14.6
7k	6,819	35.4	8.4	8.5
14k	13,638	29.8	4.3	3.2
27k	27,279	27.3	2.0	1.3
55k	54,746	26.5	1.0	0.2
cc	N/A	8.8	N/A	N/A

Table 1: Transcription quality vs vocabulary size

4 IR Strategies

The experiments reported in this paper try to capture some of the effects of OOV words on retrieval performance without analyzing in detail the different problems caused by OOV words. The TREC-8 SDR test collection[1] used for the experiments consists of 21,754 spoken documents and 49 written queries.

A parallel collection of newspaper text documents, consisting of 62,926 articles spanning the same epoch as the test audio collection, was also used for some of the experiments with query and document expansion. This parallel collection was larger and thought to be more reliable than the automatically transcribed test collection.

4.1 Baseline System

The documents were first stopped and stemmed in the standard way. To remove further complications, no phrase information was used during indexing. The number of terms in the (stopped/stemmed) recognition vocabulary; the story-averaged term error rate (STER) [2] and the OOV term rate on the 10 hour subset (OOVT); and the 323 terms of the 49 queries (QOVT) after pre-processing (stopping and stemming) are given in Table 2. The average number of different terms per query was 6.4 after pre-processing.

ID	#Terms	STER	OOVT	QOVT
3k	2,142	68.8	22.3	21.4
7k	3,992	53.4	12.8	12.7
14k	7,777	41.4	6.5	5.9
27k	15,442	36.7	3.0	1.9
55k	30,811	35.1	1.6	0.0
cc	N/A	10.5	N/A	N/A

Table 2: 10hr subset term OOV and error rates

The baseline retrieval system, (`BASE`), uses Okapi-style combined weights (CW), exploiting term frequencies within and across the documents[4]. The Mean Average Precision (MAP) results for the different transcriptions are given in Table 3.

The results show that for relatively low WER and OOV rates, performance drops slightly with increasing error rate, but the fall off is much greater for the smaller vocabulary systems and higher OOV rate. The %OOV rate, WER and STER all show similar trends when just the vocabulary size is changed. Although STER would be the most appropriate measure to relate transcription quality to retrieval performance across different speech recognition engines which have varying word error rates for the same vocabulary size, for a single recogniser the OOV rate provides a similar function.

The following experiments explore ways of counteracting OOV effects in retrieval. We assume that misses are the major problem encountered and therefore consider potential ways of overcoming missing query terms and missing term relations.

4.2 Query Expansion

Query expansion could be used to attack the problem of missing query terms directly and missing term relations indirectly.

A strategy proven to work well in SDR [2] is blind relevance feedback (BRF) on the test collection. Although the expected benefit from this device may weaken as the quality of the test collection degrades, some missing term relations due to OOV effects might be recaptured using this technique. Here we consider explicitly *adding* new terms to the existing query Q, to form Q', not simply replacing one term set by another. Candidate expansion terms are defined by their offer weights OW, and the terms in Q' by their combined iterative weights CIW [4].

Table 3 gives the performance for BRF when adding 3 new terms drawn from the top 10 documents. It is difficult to disentangle the general effects of enlarging queries from any specifically anti-OOV effects; and BRF

is a rather weak beneficial strategy. However, the results show an improvement for all transcriptions.

Since BRF is already a weak technique and the OOV words in the test collection have not been correctly recognised it is natural to consider the use of a parallel collection for both query and document expansion. First we consider query expansion. The parallel collection is larger and more reliable than the test collection and should provide better expansion information, both for terms and weights. This helps address the problem of missing query terms directly and missing term relations indirectly.

There are several ways of exploiting the parallel collection, but we present just a simple union method here. We assume that the parallel and test collections are similar and work with the union of the test and parallel collections, U. Therefore BRF is applied directly to U, denoted UBRF, with term weights computed using U. However since (we assume) the user only wants documents from the test collection, the expanded query was then applied only on the test collection.

ID	BASE	BRF	UBRF
3k	22.2	24.4	33.3
7k	33.8	37.5	44.3
14k	41.4	47.6	51.8
27k	43.0	49.7	53.8
55k	43.5	50.3	54.6
cc	47.9	54.4	56.1

Table 3: Effects of Query Expansion on MAP

The results from adding 6 new terms to each query from the 20 top documents with UBRF is given in Table 3 which shows that this is a very effective strategy.

4.3 Document Expansion (DE)

Anti-OOV effects with query expansion can come from facilitating more joint term matches, or from pulling in terms semantically related to missing OOV terms. However, the parallel text collection could be used to retrieve OOV terms for the test documents directly via a process of *document expansion* [3]. If OOV terms are reintroduced this should help matching with the original queries (or other useful terms may be imported).

In our implementation of DE, each document in the test collection is expanded by forming a *pseudo-query* of the 100 terms with the lowest collection frequency from that document. BRF is run on the parallel collection adding the best 200 new terms taken from the top 10 documents to the original document with a term frequency equal to 1. The baseline system is then run on the expanded test collection documents using the original query with CW from the document-expanded index file.

The results in Table 4 (column DE) show that the retrieval performance improves for all transcriptions with somewhat larger improvements from the smaller, higher OOV-rate vocabularies.

A further experiment was conducted to see if the benefits from UBRF and document expansion could be combined. The query file from the UBRF experiment in section 4.2 with CIW[1] was run on the document-expanded index file.

[1] With term frequencies computed from the document-expanded collection but relevance weights computed using U.

ID	BASE	DE	UBRF	DE+UBRF
3k	22.2	27.9	33.3	37.0
7k	33.8	38.6	44.3	46.6
14k	41.4	46.7	51.8	53.4
27k	43.0	48.4	53.8	55.8
55k	43.5	48.7	54.6	56.6
cc	47.9	51.1	56.1	56.3

Table 4: Effects of Document Expansion on MAP

The results in Table 4 (DE+UBRF) show that performance improves for all automatic transcriptions by using the UBRF query. This confirms that query expansion and document expansion are compensating for errors in different ways. It can also be seen that using this final scheme the performance from the automatic 55k transcriptions matches that from the closed caption transcriptions.

4.4 Summary of Results

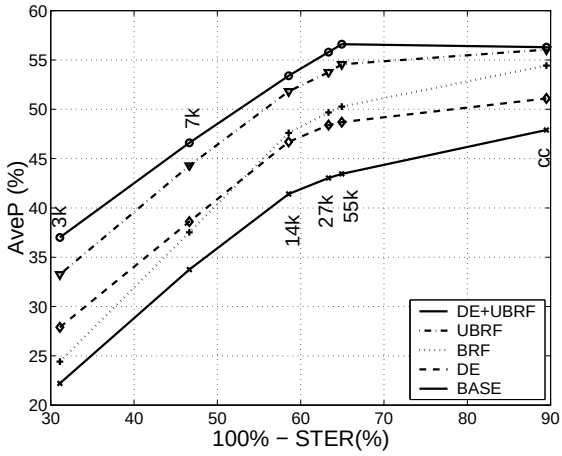

Figure 1: Summary of Results

The results using the various retrieval setups for the different transcription sets are shown in Figure 1. It can clearly be seen that more advanced IR techniques compensate for recognition errors caused by OOV words and this allows retrieval performance to reach the level of a more basic system with a much lower OOV rate.

References

[1] J S Garofolo, C G P Auzanne & E M Voorhees *1999 TREC-8 Spoken Document Retrieval Track: Overview, Results and Analyses* To appear Proc. TREC-8, 2000

[2] S E Johnson, P Jourlin, K Spärck Jones & P C Woodland. *Spoken Document Retrieval for TREC-8 at Cambridge University.* To appear Proc. TREC-8, 2000

[3] A Singhal & F Pereira *Document Expansion for Speech Retrieval* Proc. SIGIR '99, pp. 34-41, 1999

[4] S E Robertson & K Spärck Jones *Simple, Proven Approaches to Text Retrieval* Technical Report TR356, Cambridge University Computer Laboratory, May. 1997