Effects of Out of Vocabulary Words in Spoken Document Retrieval

Phil Woodland†, Sue Johnson†, Pierre Jourlin‡ & Karen Spärck Jones‡

† Cambridge University Engineering Department ‡ Cambridge University Computer Laboratory

{ pcw , sej28 }@eng.cam.ac.uk { pj207 , ksj }@cl.cam.ac.uk http://svr-www.eng.cam.ac.uk/Research/Projects/mdr/

SUMMARY

- Investigate the effects of out-of-vocabulary (OOV) items in spoken document retrieval (SDR).
- Multiple automatically generated transcription sets for TREC8-SDR with different OOV rates but same recogniser.
- The performance with a basic IR engine and with query and document expansion is given.
- OOV effects can be partially compensated for by more advanced retrieval strategies.

INTRODUCTION

The standard approach to SDR uses a large vocabulary continuous speech recognition (LVCSR) system to generate word-level transcriptions and then applies standard text IR techniques to the transcripts. One of the major issues is the effect on retrieval of OOV words. The aim of this work is to investigate the scale and nature of these effects on SDR performance for a particular recogniser, and to see how the effects vary as query and document expansion are added to the retrieval system.

OOV EFFECTS IN LVCSR AND IR

In LVCSR:

- Recogniser interprets each audio segment as a sequence of items from the recognition vocabulary
- OOV words are *missing* from automatically transcribed spoken documents and are *replaced* by the most probable alternatives.
- Note that recogniser uses full word forms: not root forms.

In IR:

- Query OOVs (QOV) can lead to *word miss* in searching
- Stemming implies that only stem recognition errors are important
- Can lead to missing term relations which affect query expansion sets

SPEECH TRANSCRIPTIONS

- 500 hours of TREC8 SDR audio
- Single pass non-adapted HTK LVCSR system (context-dependent HMMs + 4-gram language model)
- Recognition vocabularies approx 55k, 27k, 14k, 7k, 3k (most frequent words)
- 2-3xRT: 10 CPU months for transcription generation
- Also closed caption (cc) transcriptions
- Error rates assessed on 10 hour carefully transcribed portion

TRANSCRIPTION QUALITY

ID	# Words	WER	OOVW	QOVW	#Terms	STER	OOVT	QOVT
3k	3,413	43.9	14.1	14.6	2,142	68.8	22.3	21.4
7k	6,819	35.4	8.4	8.5	3,992	53.4	12.8	12.7
14k	13,638	29.8	4.3	3.2	7,777	41.4	6.5	5.9
27k	27,279	27.3	2.0	1.3	15,442	36.7	3.0	1.9
55k	54,746	26.5	1.0	0.2	30,811	35.1	1.6	0.0
cc	—	8.8	—	—	—	10.5	—	—

Table 1 : *Error and OOV rates for different vocabulary sizes (10 hour subset)*

OOVW/OOVT *Out of Vocabulary Rate (Word/Term)*
QOVW/QOVT *Query Out of Vocabulary Rate (Word/Term)*
WER/STER *Word Error Rate/Story Averaged Term Error Rate*

- %OOV, WER and STER all show similar trends
  - STER would be preferred when comparing across recognisers

IR STRATEGIES

- 49 TREC8-SDR test queries & 21,754 documents
- For expansion used 62,926 newspaper documents from same epoch
- Evaluated in terms of Mean Average Precision MAP

BASELINE SYSTEM

- Okapi-style combined weights: no query expansion BASE
- All automatic transcriptions give a MAP lower than cc
  - Increasingly sharp degradation in performance as OOV increases (see Table 2).

QUERY EXPANSION

- BRF *Blind Relevance Feedback*
  - Adds new terms to query with maximum Offer Weights from initial search on test collection
  - Added 3 terms from top 10 documents
  - Better results with BRF for low-OOV cases
- UBRF *Union Blind Relevance Feedback*
  - Do BRF on *union* of test collection and parallel collection
  - Added 6 terms from top 20 documents
  - Very substantial improvements for all transcriptions : especially at high OOV rates.

DOCUMENT EXPANSION

- Try to utilise the parallel collection specifically to overcome OOV defects in documents
- Use DE *Document Expansion* (similar to Singhal, SIGIR'99)
- Use test documents as pseudo-query and add 200 new terms from top 10 parallel documents to each test document.
- Improvements for all transcriptions (from BASE): bigger improvements for higher OOV cases
- Combining DE+UBF gives further gains (small for (cc)) .

ID	BASE	BRF	UBRF	DE	DE+UBRF
3k	22.2	24.4	33.3	27.9	37.0
7k	33.8	37.5	44.3	38.6	46.6
14k	41.4	47.6	51.8	46.7	53.4
27k	43.0	49.7	53.8	48.4	55.8
55k	43.5	50.3	54.6	48.7	56.6
cc	47.9	54.4	56.1	51.1	56.3

Table 2 : *Effects of Expansion on MAP*

SUMMARY OF RESULTS

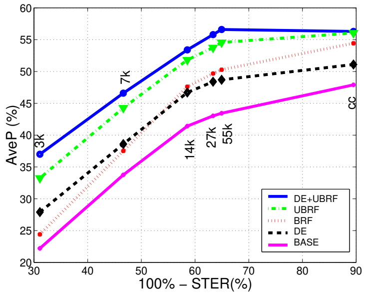

Figure 1 : *Summary of results for all transcriptions*

CONCLUSIONS

- Use of a parallel collection for UBRF and DE compensates for OOV words
- Can get results as good as those for cc with BASE system from a 7k vocab system
- For TREC8-SDR good performance can still be achieved with a moderate-size vocabulary