Spoken Document Retrieval at Cambridge University

Sue Johnson†, Pierre Jourlin‡

Karen Spärck Jones‡ & Phil Woodland†

November 14th, 2000

† Cambridge University Engineering Department

‡ Cambridge University Computer Laboratory

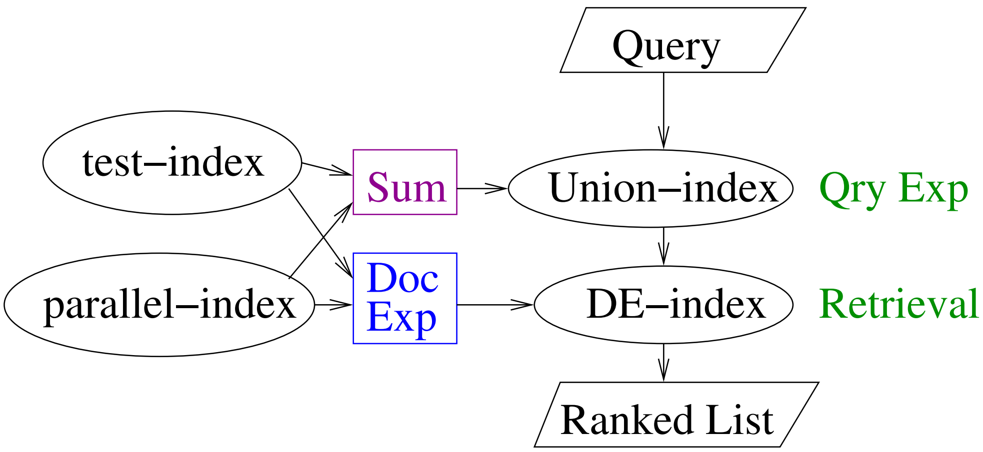
Outline

- Summary of Story Known (SK) system
- Summary of Story Unknown (SU) system
- SU results
- SU/SK comparison
- NEW Document expansion for SU
- NEW Non-lexical information
- Conclusions

[NB: Terse 'TREC-8' queries developed in-house]

Summary of SK System

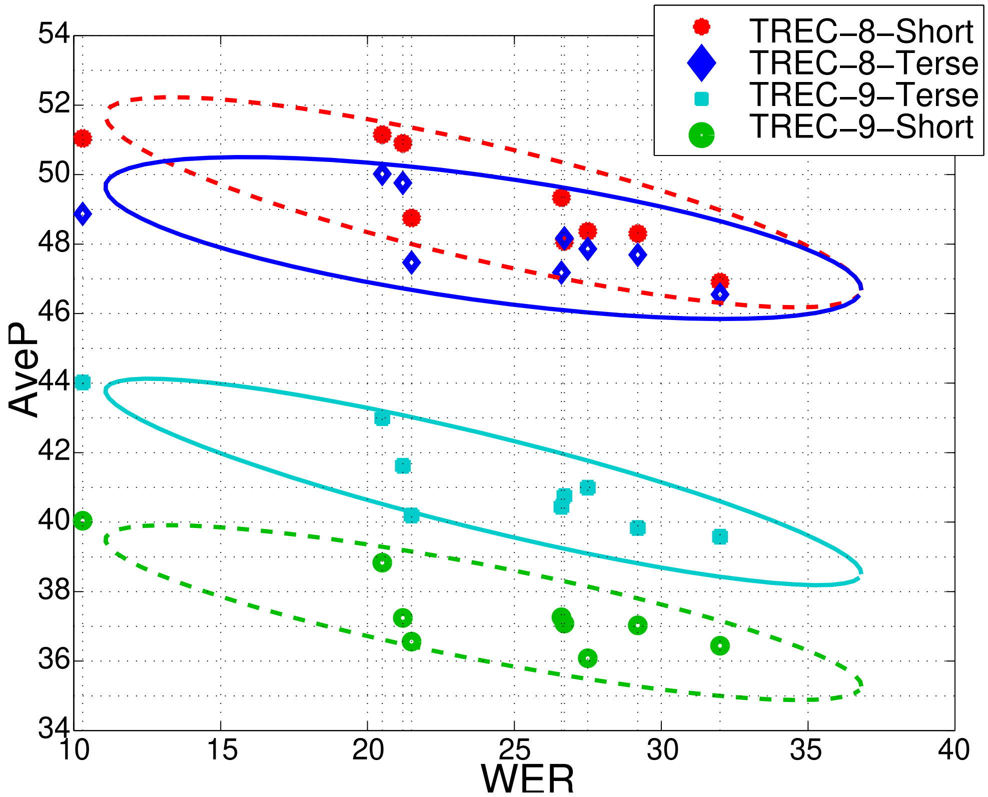
- Used HTK transcripts from TREC-8 SDR eval.
- Text-based sources normalised to ASR form. NEW
- Used stopping, stemming, Okapi-based weighting.
- Used document expansion from parallel corpus.
- Used 1-step query expansion on union corpus. NEW



Summary of SU System

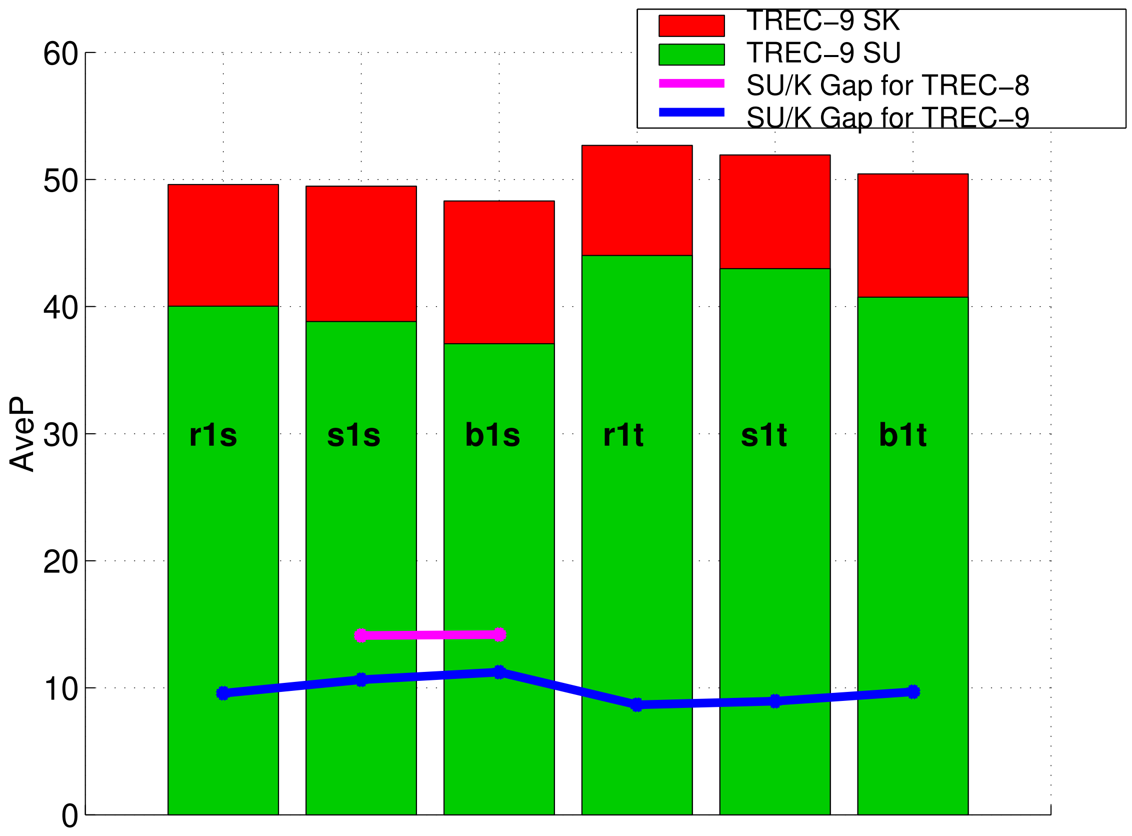
- Non-lexical information generated from audio. NEW
- 57% non-story (65h) detected @ 0.23% story loss.
- This "non-story" info used to filter transcriptions.
- Windowing 30s@15s to partition audio-stream.
- Retrieval similar to SK but no document expansion (and no document length normalisation).
- Post-retrieval merging (240s) to remove duplicates. (Assumes the probability of 2 distinct relevant stories existing in the same episode is small)
- 51% AveP on TREC-8 queries. (c.f. 41% last year)

Summary of SU Results



- Performance falls only slightly with increasing WER
- Fall-off is unchanged over query sets
- TREC-9 query sets were generally harder than TREC-8
- Optimal parameters for short/terse queries different.

SU/SK Comparison



- SK/SU gap reduced from TREC-8 but still exists
- Gap partly due to
  - Story (not passage) based scoring
  - Longer, coherent, single-topic documents for SK
  - Document expansion used in SK
    (also helps further compensate for recognition errors)

Document Expansion for SU NEW

Algorithm

1. Form pseudo-query of each term in window (>10 terms)
2. Find 100 terms with top OW on parallel corpus (r=15)
3. Increment term frequency by 1

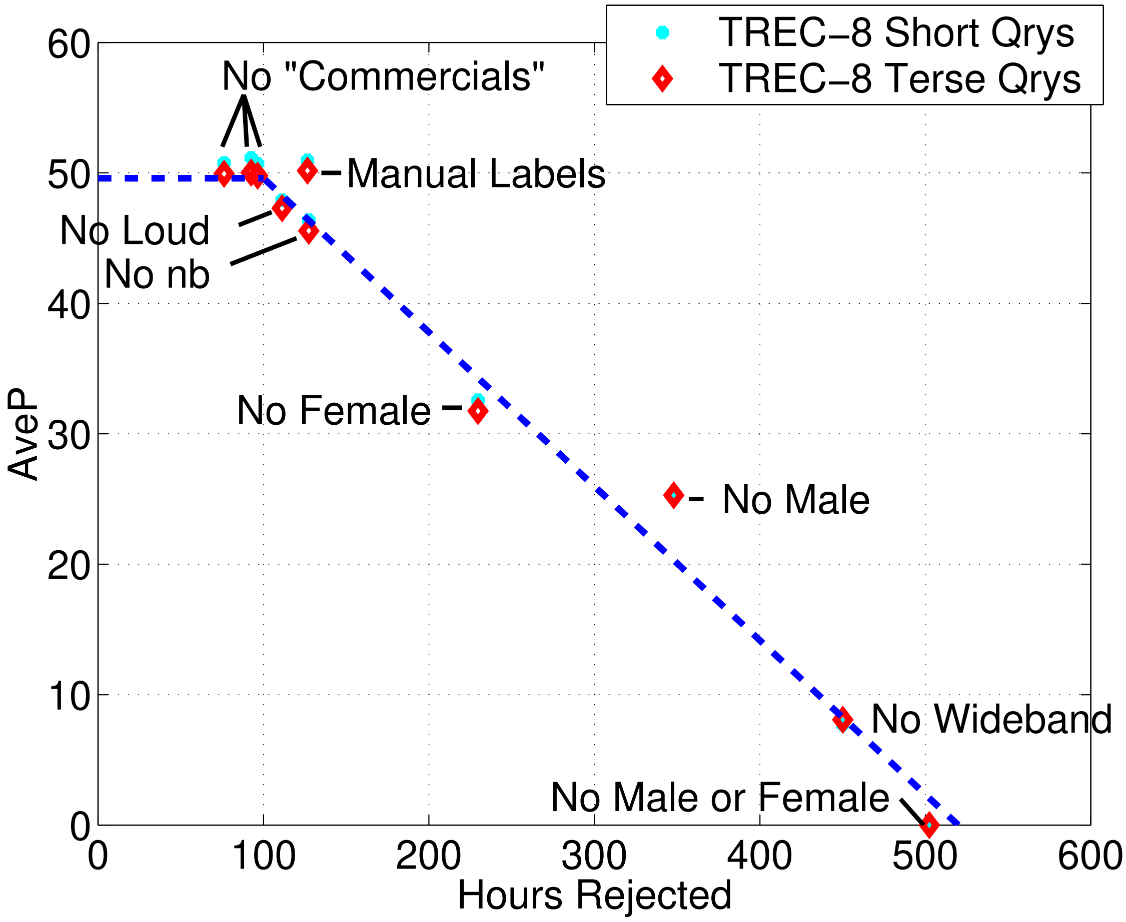
Results (AveP for TREC-8 queries)

QryExp (t@r)	Short Q	+DE	Terse Q	+DE
-	30.89	+7.79	32.51	+4.76
8 @ 20	50.84	+0.88	47.28	+0.37
17 @ 18	51.53	-2.34	49.37	-2.37
23 @ 27	51.06	-2.03	50.56	-6.08

- Works well for low to moderate query expansion
- Is detrimental for high level of query expansion (esp. terse queries)

Non-lexical Information NEW

- Non-lexical info can be useful in SDR (e.g. interactive)
- {gender,bandwidth,segment,(high)-energy nospeech,repeat,commercial} found acoustically
- Per unit time, female-only best, male-only worst



Conclusions

- Standard SDR techniques work well

- Query expansion from parallel corpus is very useful

- Robustness against recognition errors confirmed

- Non-lexical info. can be generated automatically for
  - Retrieval e.g. remove commercials
  - Broadcast structure e.g. finding story boundaries
  - Real systems e.g. to help browsing/understanding

- Document expansion is possible for SU systems, but improvements decrease for high level of query expansion

- Community is ready for new challenges such as video

Detailed questions? → email : sej28@eng.cam.ac.uk