Audio Indexing and Retrieval of Complete Broadcast News Shows

Sue Johnson, Andreas Tuerk, Phil Woodland

Pierre Jourlin, Karen Spärck Jones

April 13th 2000

Cambridge University Engineering Department
Cambridge University Computer Laboratory

Outline

- The SU-SDR task
- Our 1999 SU-SDR System
  - Eliminating Commercials
  - Speech Recognition
- Experiments
- Results
  - Eliminating Commercials
  - Enforcing Broadcast Structure
  - Changing the Retriever
  - Optimising Post-Retrieval Processing
- Future Work
- CU MDR Demo System

Description of the Task

- Motivation
  - Increased amount of audio and video data available
  - Retrieve information with no manual intervention
  - Need to automatically transcribe, retrieve and select relevant parts (stories)

- Related areas
  - Story Known Spoken Document Retrieval (SK-SDR)
  - Passage-based information retrieval
  - Topic Detection and Tracking (TDT)

- Approaches
  1. Segment into stories and run SK-SDR
     (e.g. use TDT, TextTiling, Audio/Video cues)
  2. Window transcriptions, run passage retrieval and recombine close ones to remove duplicates

- Evaluation Framework
  - Formal: 1999 TREC-8 SU-SDR evaluation
  - Informal: User response from demo system

Our 1999 Complete SU-SDR System

GENERATE TRANSCRIPTIONS	Eliminate commercials Eliminate music and silence
GENERATE SLIDING WINDOWS	Generate 30s windows @ 15s Enforce boundaries at gaps >5s Truncate shorter windows Remove short/few-word windows
RUN RETRIEVER	weight query terms (POS), stop, stem Enhance Indexing (SPI , PCFW) Retrieve with feedback (PBRF, BRF)
POST-PROCESSING	Enforce boundaries at gaps > 5s Take only first for windows within 75s Output top 1000 per query

Uses
- large vocabulary speech recogniser
- enhanced OKAPI-based retriever.

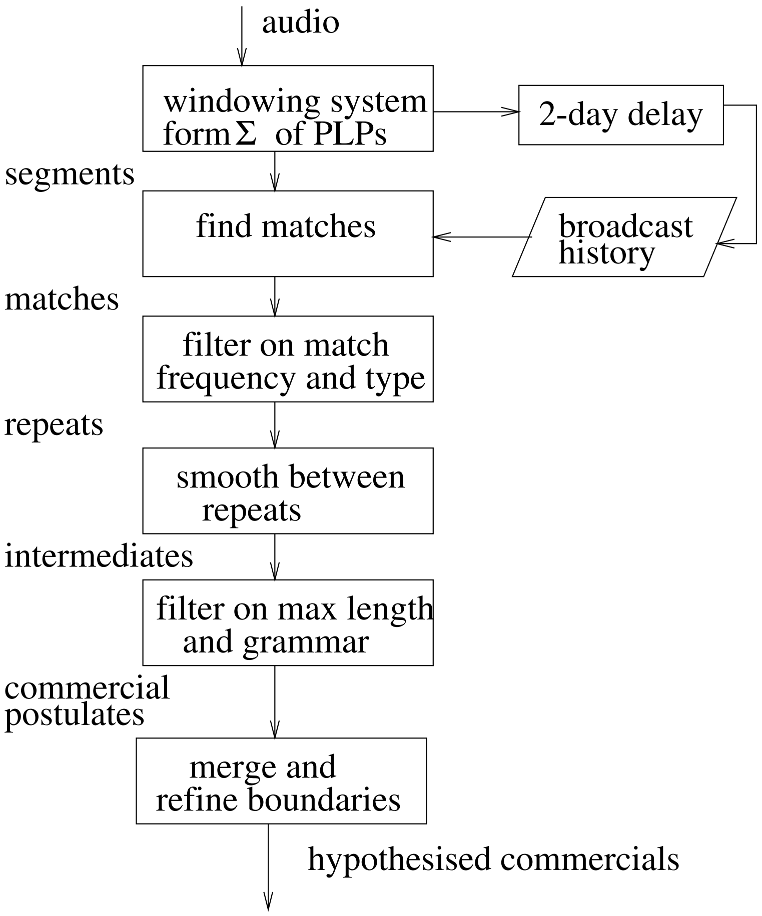
Combines windowing/recombination system with knowledge about broadcast structure learnt from:

- commercial detection
- audio segmentation and music rejection

Eliminating Commercials

Try to remove commercials because they:

- Contain little content information
- Are unlikely to be relevant, hence reduce precision
- Increase processing time required
- May help separate stories



Results from Eliminating Commercials

Conservative system used for evaluation (C-E)

- Removes 8.4% data (42.3hr audio $\equiv$ 25 CPU days)
- 97.8% manually labelled non-story
- 36.3% non-story removed with loss of only 0.23% story
- Most story loss is due to very short stories in CNN

Contrast run removes more data (C-2)

- Removes 12.6% data (63.4hr audio $\equiv$ 40 CPU days)
- 95.0% manually labelled non-story
- 52.9% non-story removed, with loss of 0.81% story
- Almost all story loss occurs in CNN broadcasts

Speech Recognition

2-pass system based on CUHTK-Entropic 1998 Hub4 10xRT system

- Segmentation
  - speaker/condition changes found
  - music and silence removed (34 hrs)
  - gender and bandwidth determined

- First Pass (`HTK-pass1`)
  - Gender independent, bandwidth specific triphones
  - 60,000 word vocab, 4-gram LM
  - WER = 26.6% on 10hr subset

- Second Pass (`HTK-pass2`)
  - MLLR-adapted models
  - Gender and bandwidth dependent triphones
  - 108,000 word vocab, 4-gram LM
  - WER = 20.5% on 10hr subset

Experiments

- Results reported for 1999 TREC-8 SU SDR task:
  500 hours of audio, (5 months of news), 50 queries
  1,818 relevant stories (RS), 1,085,882 non-relevant stories (NRS), 314,700 non-stories (NS)
- Show:time stamp mapped to manually defined story
- Non-stories (e.g. commercials) scored as irrelevant
- Duplicates scored as irrelevant
  This encourages over-merging ☹
- Performance measured using Average Precision (AveP) and R-precision (R-P) across all queries
- Also use `%retrieved` for RS (Recall), NRS and NS, before and after recombination

Results - Commercial Elimination

Before Window Recombination:

Comm.	RS	NRS	NS	#Dup
-	94.7	39.3	25.6	734,897
C-E	94.2	39.1	18.6	697,143

After Window Recombination:

Comm.	RS	NRS	NS	AveP	R-P
-	77.5	3.52	2.50	41.00	40.96
C-E	77.6	3.76	1.62	41.50	41.63

- `%retrieved` Relevant Story (Recall) unaffected
- `%retrieved` Non-Story drops dramatically
- `C-E` increased AveP (1.2% rel.) and R-P (1.6% rel.)

Further Experiments show:

- AveP on LIMSI's transcriptions increased by 1.4% rel.
- AveP and R-P increased more by *pre-filtering audio* than *post-filtering windows* after retrieval
- using `C-2` increased R-P further, but not AveP

Enforcing Broadcast Structure

Hard Breaks (HB) enforced in both window generation and recombination where gap > 5s occurs in transcription.

1. Enforcing structure in Window Recombination

HB	RS	#Dup	AveP	R-P
N	78.3	2422	41.44	42.01
Y	78.4	2707	41.47	41.98

- Less merging occurs (more duplicates remain)
- AveP, R-P and Recall unaffected
- Using a smaller gap to define the hard breaks is worse

2. Enforcing gaps in Window Generation

HB	RS	NRS	NS	AveP	R-P
N	78.3	3.43	2.39	41.71	40.07
Y	78.4	3.74	1.68	41.47	41.98

- NS `%retrieved` reduced dramatically
- Recall and R-P increased slightly
- AveP falls slightly

Changing the Retriever

1. Enhanced indexing: Semantic Poset Indexing (SPI)

SPI	Recall	AveP	R-P
N	79.2	43.42	43.35
Y	78.3	41.71	40.07

- Many more windows are returned by retrieval
- Recall, AveP and R-P all decrease

2. Improved Search: Blind Relevance Feedback (BRF)

assume top `r` docs are relevant and add `t` terms to query

r	t	Recall	AveP	R-P
-	0	78.4	41.52	42.93
10	5	79.2	43.42	43.35
10	10	79.4	44.28	44.23

- BRF increased AveP (4.6% rel.) and R-P (1.0% rel.)
- changing `r` worsened performance
- Adding more terms improved performance
- Recall and R-P increased further by adding more terms

Changing the Retriever II

3. Improved Search: Parallel Blind Relevance Feedback

r	t	Recall	AveP	R-P
-	0	77.2	39.01	40.40
20	7	79.4	44.28	44.23
25	7	79.8	44.02	45.61
20	10	80.6	44.23	44.68

- PBRF increased AveP (13.5% rel.) and R-P (9.5% rel.)
- Values of `r` and `t` generalise well
- Different `(r,t)` values are best for each criterion

4. Tuning for document type: b and K parameters

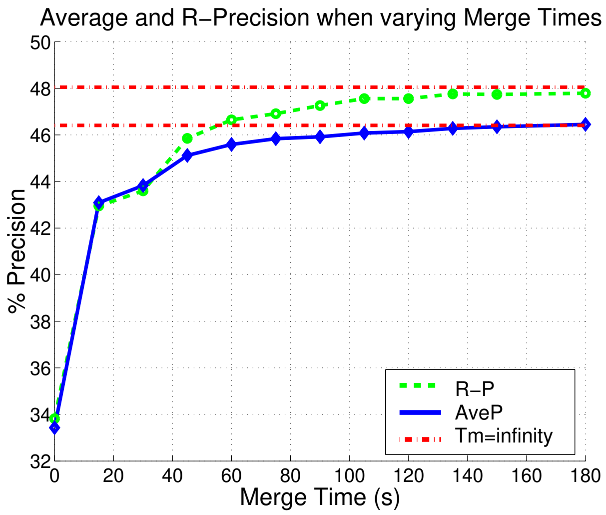
b=0 $\Rightarrow$ long documents contain multiple topics

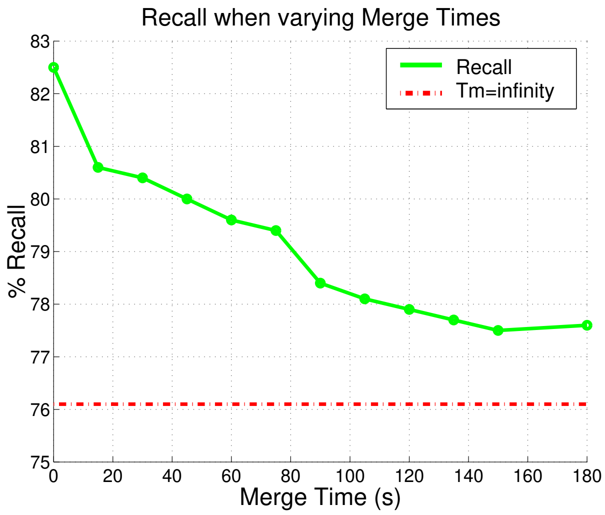
b	K	AveP	R-P
0.5	1.0	44.28	44.23
0.0	1.0	45.50	47.04
0.0	1.25	45.84	46.91

- setting b=0 increased AveP (2.8%) and R-P (6.4% rel.)
- increasing `K` gave further slight increase in AveP
- Recall unaffected by changing `b` and `K`

Optimising the Recombination

1. Altering the Merge Time





- # Duplicates falls by a factor of 3 when recombining overlapping windows and a factor of 13 when taking top window per show

- AveP and R-P increase monotonically to an asymptote

- Taking top window per show gives best precision and speeds up recombination

- Recall continues to fall as merging increases

- "Best" system depends on the desired Precision/Recall trade-off

Optimising the Recombination II

2. Reduce # Windows in Recombination

(Speed up and reduce %retrieved for NRS and NS)

Before Window Recombination

	RS	NRS	NS	#Dup	# windows
0.1	96.5	44.20	30.41	871,428	1,448,864
10	82.1	3.22	1.66	38,499	80,196

- Recall drops by 15% relative
- %retrieved drops by x13 for NRS and x18 for NS
- # duplicates drops by factor of 22

After Window Recombination

	Recall	AveP	R-P
0.1	79.4	45.84	46.91
10	77.0	45.78	46.91

- Number of windows in recombination reduced by a factor of 20 with negligible loss in AveP and R-P

Future Work

Within the TREC-8 framework

- ? Broadcaster-specific recombination
- ? Windowing parallel corpus
- ? Document expansion with SU-SDR
- ?? Combination of different SU-SDR systems
  e.g. also including story-segmentation methods

New for the TREC-9 SU-SDR framework

- ✓ Text-based audio description for retrospective filtering using audio-derived information
- ? Combination of text and audio-based cues
  (e.g. for story segmentation)

Summary

⋆ Retrieving information from broadcast news is still possible when only the audio data is available

- Information about broadcast structure can be found from acoustic segmentation and commercial detection

  ◦ High accuracy transcriptions can be produced

  ◦ Automatically removing commercials increased performance

  ◦ Using broadcast structure didn't affect performance

  ◦ BRF and PBRF increased performance

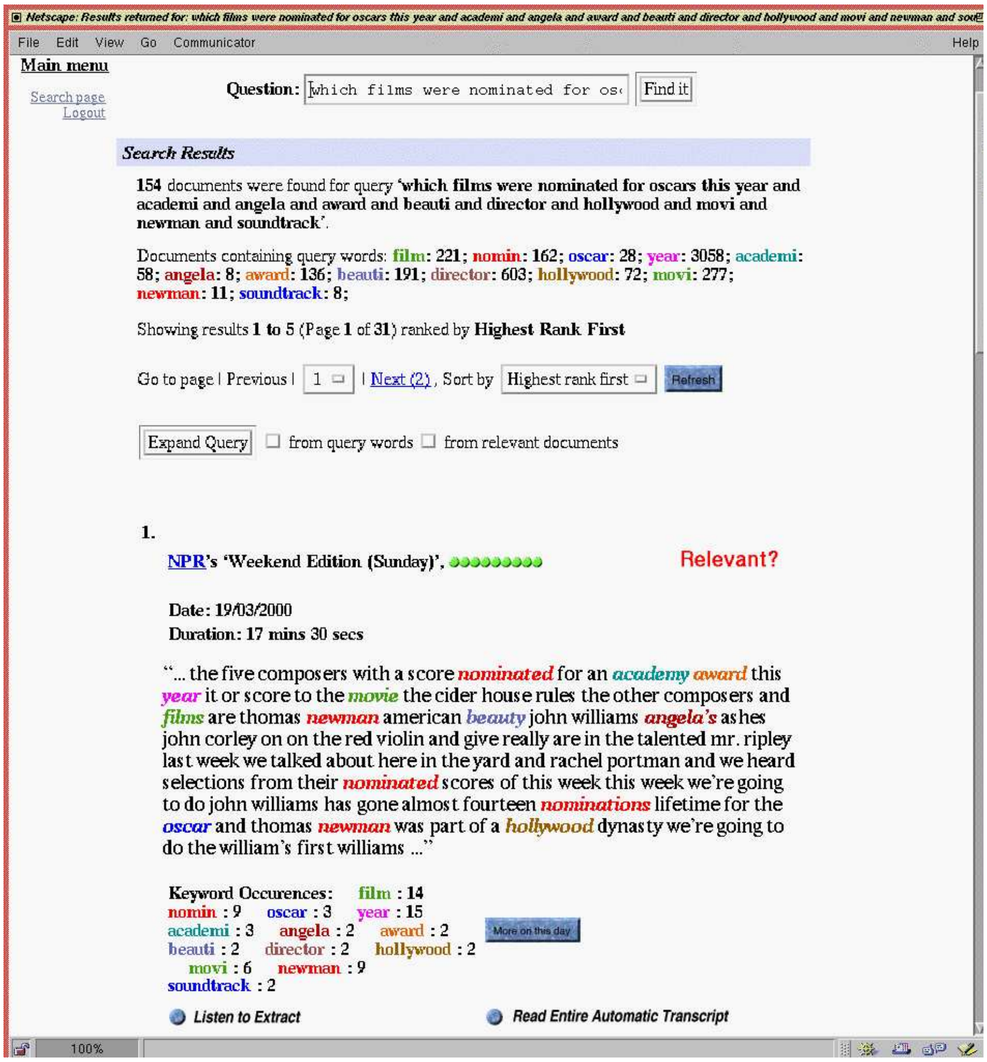
  ◦ Increasing the amount of recombination increased Precision at a cost to Recall

  ◦ Recombination can be significantly speeded up at no cost to performance

The CU MDR Demo System

- Web-based using multiple online RealAudio sources
- High quality US and British English transcriptions
- ○ Known and unknown story boundary sources combined
- ○ Interactive query expansion using WordNet
- ○ Interactive query expansion using relevance feedback
- ○ Filtering from static queries also included
- ⋆ Relative indicator of relevance given
- ⋆ Results can be reordered e.g. by relevance, date
- ⋆ Matching search terms highlighted
- ⋆ Extract and whole story can be displayed and played
- ⋆ Parts of transcription can be highlighted and played

The CU MDR Demo System



For more info

http://svr-www.eng.cam.ac.uk/Research/Projects/mdr

to find out more about the TREC-9 SDR evaluation
email john.garofolo@nist.gov