Highly Imperfect Automatic Transcript from low-quality audio source

(Excerpt from in-class student presentation: accurate words highlighted in bold)

... members are now he did John fees the premise of the Solicitor roof not it moment of that is David cone was a Allah it began an and I this by to aver and began Internet and robotic problem right to go is a new Harman's and functions that we've developed for the design will and will the preliminary research and design the river begin similar solutions the week, with and then vote in the doing next few weeks for work and the impact on a long haul that we've given was that children who wheelchair-bound once any the use on swings of modern this lens without being transferred to U.S. standards went subject of why this suit bill the slaying that eight of

Poor Index:
• too many inaccurate terms
• no ranking of importance

TERM | VALID | RANK
--- | --- | ---
mem | ? | ?
are | ? | ?
now | ? | ?
we | ? | ?
did | ? | ?
John | ? | ?
fee | ? | ?
the | ? | ?
promise | ? | ?
if | ? | ?

- Highly imperfect transcripts contain more invalid than valid terms
- In the absence of more accurate material (produced at higher cost), these inaccurate transcripts must be mined for useful information

Apply proper filter
• Validity and ranking of terms and phrases in these transcripts are necessary for proper indexing

Determine rank of words

Problem

SELECTION AND RANKING OF TEXT FROM HIGHLY IMPERFECT TRANSCRIPTS FOR RETRIEVAL OF VIDEO CONTENT

Alexander Haubold
Department of Computer Science, Columbia University

Highly Imperfect Automatic Transcript from low-quality audio source

Approach

External Corpus as Filter

- e.g. Text from Presentation Slides applied to Presentation Videos:
  - Qualified phrases (including stop-words) (iteratively test consecutive words in WordNet)
    - e.g. “Statue of Liberty”
  - Phrases not in WordNet (identically valid terms, extract unidentified terms in-between)
    - e.g. “the Circle Line to the Statue of Liberty”
  - All single words (excluding stop-words)
    - e.g. “statue”, “liberty”, ...
  - Individual words (including stop-words)
    - e.g. “amortize”, “analysis”, ...
  - Phrases (each index line is one phrase)
    - e.g. “amortize analysis”, ...

Filter Imperfect Transcript

- Filter Imperfect Transcript

- Term/Phrase not in WordNet:
  - If unidentified, then possible named entity. Constant C chosen arbitrarily (e.g. C = 10) to emphasize meaning of these phrases.

Rank Terms / Phrases

A. Term/Phrase not in WordNet:
  - WordNet
  - Rank = number of terms * C

B. Term/Phrase in WordNet:
  - 1. Count synsets: more specific terms/phrases have fewer synsets.
    - numSynsets
  - 2. Distance to root sense: more distant term/phrase has more specific meaning.
    - distRoot
  - 3. Count noun synsets: nouns are more descriptive.
    - nounEmphasis
    - Rank = numWords*distRoot + nounEmphasis

User Study

Multimedia Browser VAST MM (Video Audio Structure Text)

- Browser used ...
  - by students to self- and peer-evaluate presentation performance
  - by instructors to archive and review student performance
- Current database:
  - >150 hours of presentation material, >180 video tapes
  - 5 year duration of course, >1500 students

• Main browser features:
  - Keyframe and Speaker index
  - Filtered, ranked transcript text in yellow boxes
  - Text Search through filtered text and raw transcripts
  - Streaming Video

Evaluation of two different User Interfaces for keywords / keyphrases

- Words/phrases are not ranked
  - Text boxes are horizontally temporally aligned to video
  - Text boxes fill UI space in a greedy fashion

User Study Setup:
- 3 semesters, 442 students
- Targeted task:
  - “Give a high-level idea about presentation content using only keyphrases!”
- 3 approaches to UI and evaluation:
  1. Without visual ranking, multiple choice
  2. With visual ranking, multiple choice
  3. With visual ranking, articulate response

- Measures of completion:
  - Average duration of task
  - Accuracy of response (manual evaluation)

User Study Results:
- With ranking and visualization:
  - Accuracy of responses increases
  - Duration of summarization task decreases
    - (When articulating response, additional time is required for typing)

Application

Multimedia Browser VAST MM (Video Audio Structure Text)

- Browser used ...
  - by students to self- and peer-evaluate presentation performance
  - by instructors to archive and review student performance
- Current database:
  - >150 hours of presentation material, >180 video tapes
  - 5 year duration of course, >1500 students

• Main browser features:
  - Keyframe and Speaker index
  - Filtered, ranked transcript text in yellow boxes
  - Text Search through filtered text and raw transcripts
  - Streaming Video

- Browser used ...
  - by students to self- and peer-evaluate presentation performance
  - by instructors to archive and review student performance
- Current database:
  - >150 hours of presentation material, >180 video tapes
  - 5 year duration of course, >1500 students

• Main browser features:
  - Keyframe and Speaker index
  - Filtered, ranked transcript text in yellow boxes
  - Text Search through filtered text and raw transcripts
  - Streaming Video

- Browser used ...
  - by students to self- and peer-evaluate presentation performance
  - by instructors to archive and review student performance
- Current database:
  - >150 hours of presentation material, >180 video tapes
  - 5 year duration of course, >1500 students

• Main browser features:
  - Keyframe and Speaker index
  - Filtered, ranked transcript text in yellow boxes
  - Text Search through filtered text and raw transcripts
  - Streaming Video