Interactive Approaches to Video Lecture Assessment

August 13, 2012

Korbinian Riedhammer

Speech Group
Pattern Recognition Lab
Motivation
Computer Science 188, 001|Spring 2012|UC Berkeley

Pieter Abbeel

Description

COMPSCI 188 LEC 001 Sp12

<table>
<thead>
<tr>
<th>Name</th>
<th>Time</th>
<th>Released</th>
<th>Description</th>
<th>Popularity</th>
<th>Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Lecture 8</td>
<td>--</td>
<td>2/10/12</td>
<td>i</td>
<td>8</td>
<td>FREE</td>
</tr>
<tr>
<td>2 Lecture 7</td>
<td>--</td>
<td>2/8/12</td>
<td>i</td>
<td>8</td>
<td>FREE</td>
</tr>
<tr>
<td>3 Lecture 6</td>
<td>--</td>
<td>2/2/12</td>
<td>i</td>
<td>8</td>
<td>FREE</td>
</tr>
<tr>
<td>4 Lecture 5</td>
<td>--</td>
<td>2/1/12</td>
<td>i</td>
<td>8</td>
<td>FREE</td>
</tr>
<tr>
<td>5 Lecture 2: audio only available</td>
<td>--</td>
<td>1/30/12</td>
<td>i</td>
<td>8</td>
<td>FREE</td>
</tr>
<tr>
<td>6 Lecture 4</td>
<td>--</td>
<td>1/26/12</td>
<td>i</td>
<td>8</td>
<td>FREE</td>
</tr>
<tr>
<td>7 Lecture 3</td>
<td>--</td>
<td>1/24/12</td>
<td>i</td>
<td>8</td>
<td>FREE</td>
</tr>
<tr>
<td>8 Lecture 1</td>
<td>--</td>
<td>1/18/12</td>
<td>i</td>
<td>8</td>
<td>FREE</td>
</tr>
</tbody>
</table>

Total: 8 Items

Customer Ratings

We have not received enough ratings to display an average for this collection.

Click to rate ⭐️⭐️⭐️⭐️⭐️

Customer Reviews

Be the first to write a review →
1: Introduction and Goals; Data Types, Operators, and Variables

Topics covered: goals of the course; what is computation; introduction to data types, operators, and variables

Instructor: Prof. Eric Grimson, Prof. John Guttag
1-7 Voice-based Information Retrieval

Length:
0:01:25.5

Time Span of This Chapter:

Time Span of This Slide:

Play Summary (0:00:8.6)  Play Whole

Key Terms:
- Information retrieval
- Query
- Language model
- Adaptation
- Topic
- Special interest group
6 - Musteranalyse/Pattern Analysis

Zurück zur Clipansicht
Outline

Data Acquisition → Automatic Speech Recognition → Key Phrase Extraction & Ranking → Visualization

Textual Summary
Data Acquisition

- Data Acquisition
- Automatic Speech Recognition
- Key Phrase Extraction & Ranking
- Visualization
**LME Lectures – A Corpus of Academic Spoken English**

- Two lecture series read in 2009
  - Pattern Analysis (PA)
  - Interventional Medical Image Processing (IMIP)
- 18 recordings per series
- About 40 hours of audio/video data
  - Audio: 48 kHz, 16 bit (AIFF), resampled to 16 kHz
  - Video: HD, reduced resolutions available due to bandwidth
- Clip-on cordless speaker microphone, room microphones
- Constant recording setting
  - RRZE E-Studio
  - Single speaker
  - Same recording equipment
Transcription

- Semi-automatic segmentation into speech turns
  - Based on speech pauses and silences
  - 23,857 turns
  - Average duration of 4.4 seconds
  - Total of about 29 hours of speech
- Manual transcription
  - New tool for the rapid transcription of speech
  - Time effort: about 5 times real time
- Transcription results
  - On average 14 words per speech turn
  - 300,500 words transcribed
  - Vocabulary size: 5,383 (excluding foreign words and word fragments)
Annotations

- Individual lecture *PA06*
- Based on edited manual transcript
- 5 human subjects
- 20 phrases
  - Ranking
  - Salience: from 1 (*very relevant*) to 6 (*useless*)

- Further annotations
  - Lecturer’s key terms for series *PA*
  - Presentation slides in PDF format
Automatic Speech Recognition

Data Acquisition → Automatic Speech Recognition → Key Phrase Extraction & Ranking → Visualization
The Kaldi Speech Recognition Toolkit

- State of the art, open source
- 4-layer system modeled by weighted finite state transducers (WFST)
  - Statistical n-gram language model
  - Lexicon with pronunciation alternatives
  - Context dependent phonemes
  - Hidden Markov models
- Acoustic frontend
  - Mel-frequency cepstral coefficients (MFCC), 1st and 2nd order derivatives
  - Phoneme dependent linear transformations
- Acoustic modeling: subspace Gaussian mixture models
The **LME Lectures** Speech Recognition System

- 600 Gaussian components, 5,500 HMM states
- Vocabulary size: 5,383
- Language model
  - 5,370,040 bi- and tri-grams
  - Trained on 500+ million words (including spontaneous lecture speech)

<table>
<thead>
<tr>
<th>Name</th>
<th>Duration</th>
<th># Turns</th>
<th># Words</th>
<th>% WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>train</td>
<td>25h 31m</td>
<td>20,214</td>
<td>250,536</td>
<td>-</td>
</tr>
<tr>
<td>development</td>
<td>2h 07m</td>
<td>1,802</td>
<td>21,909</td>
<td>9.78</td>
</tr>
<tr>
<td>test</td>
<td>2h 12m</td>
<td>1,750</td>
<td>23,497</td>
<td>11.03</td>
</tr>
</tbody>
</table>

**WER**: word error rate
Key Phrase Extraction & Ranking

Data Acquisition → Automatic Speech Recognition → Key Phrase Extraction & Ranking → Visualization
Candidate Selection

- A *verb alone* may be vague
  - “discuss” – what?
- An *isolated noun* may be ambiguous
  - “question” – difficult or easy?
- Information about the topic is often in the *noun phrase*
  - “He asked a *difficult question* about the *modified processing of words*”.

- Apply part-of-speech tagging
- Extract noun phrases based on regular expression
Example

“It computes the principal axes that’s the one d axis that shows the highest spread of the points”

adjective* (noun,number)* (article+ adjective* (noun,number)+)*
Example

“it computes the principal axes that ‘s the one d axis that shows the highest spread of the points”

axes principal axes one d axis one d d axis spread highest spread points spread of the points highest spread of the points
Unsupervised Ranking

- Frequent phrases may be salient
- With a similar occurrence count, longer phrases may be more salient
- Motivated by
  - *Didactics*: less confusion by literal repetition
  - *Psycholinguistics*: lexical entrainment

\[
weight(phrase_i) = \begin{cases} 
  f_i, & n_i = 1 \\
  f_i \times (n_i + 1), & n_i > 1 
\end{cases}
\]

- Data and domain independent: simple and reliable
- Other investigated strategies include prior world or domain knowledge
Comparison of Rankings

- Compare a target ranking against a reference (human) ranking
- Standard measure: *Normalized Distributed Cumulative Gain*
- Award credit for placing valuable phrases at high ranks
- Compare lists of a certain length, e.g., top 10 phrases
- Phrases annotated with salience from 1 (*very useful*) to 6 (*useless*)

\[
gain(phrase_i) = 2^{(6 - grade_i)/5} - 1
\]

\[
NDCG(N) = C \sum_{i=1}^{N} \frac{gain(phrase_i)}{\text{ld}(1 + i)}
\]
Multiple Annotators – Objective Results

- NDCG for pair-wise comparison only
- 5 human annotators
- **Human** score
  - average NDCG value of all human-human pairings
  - 20 individual pairings
- **Automatic** score
  - average NDCG value for all human-machine pairings
  - 5 individual pairings
- Scores based on manual (*TRL*) and automatic (*ASR*) transcripts
Evaluation of Human and Automatic Rankings

- Fairly high human average agreement
- Similar quality of human and automatic ranking
- Only small differences due to ASR errors
Visualization

Data Acquisition  →  Automatic Speech Recognition  →  Key Phrase Extraction & Ranking  →  Visualization
Motivation

- Key phrases give a topical overview of the lecture
- Phrase occurrences can serve as a visual index or navigation aid
- Simple example: clickable occurrence bar
StreamGraphs

- Popular in the visualization community
- Stacked splines
- Left to right: Playback time (as with occurrence bar)
- Stream wideness: Current phrase dominance
- Dominance: Number of occurrences within certain time frame
Advantages

● Comfortably display 3 to 6 phrases simultaneously
● Stream width can suggest topical relations of phrases
  ● Similar widths at the same time indicate co-occurrence – possibly *related*
  ● Different widths indicate rare or no co-occurrence – possibly *unrelated*

User Interactions

● Click into the stream – jump to closest occurrence
● Change the phrases on display – learn about topics and relations
● Interactions can be logged to collect data for customized rankings
Implementation Details

server

Automatic Speech Recognition

Key Phrase Extraction & Ranking

Visualization

client

Data Acquisition

Key Phrase Extraction & Ranking

Visualization

server

client

data

feedback

server

client

Automatic Speech Recognition

Key Phrase Extraction & Ranking

Visualization

server

client

Automatic Speech Recognition

Key Phrase Extraction & Ranking

Visualization

server

client

Automatic Speech Recognition

Key Phrase Extraction & Ranking

Visualization
User Study
Task Based Evaluation

- Typical scenario: preparation for an exam
- Task should be independent of prior knowledge and comprehension
- Locate those segments of the video that cover certain topics
- Two groups of CS graduate students – test and control
  - Familiar with topic, speaker and lecture
  - 5 subjects per group
- Each participant is provided with
  - 3 lecture topics with short description
  - Control group: Video only
  - Test group: the presented interface
- Post-use questionnaire for test group to gather feedback
Results

<table>
<thead>
<tr>
<th>Group</th>
<th>Accuracy</th>
<th>Average time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>68 %</td>
<td>30</td>
</tr>
<tr>
<td>Test</td>
<td>69 %</td>
<td>21</td>
</tr>
</tbody>
</table>

*average time in minutes*

- Both groups have a similar accuracy
- Video duration: 42 minutes
- The *test* group was on average about 29% faster
- Most users found the…
  - interface to be helpful and easy to use
  - key phrase visualization to give a good overview
Summary

- **LMELectures**, a new corpus of academic spoken English
- Automatic speech recognition system for the *LMELectures* with a word error rate of 11%
- Unsupervised key phrase extraction and ranking that highly correlates to human rankings
- Novel video lecture browser that helps students to quickly assess the contents
Outlook

- More transcriptions for better acoustic and language models
- Integration of prior knowledge about speaker, room and topic
- Supervised methods for user-tailored rankings
- Larger user study on more lectures