VideoCLEF 2008: ASR Classification based on Wikipedia Categories
Motivation

System design and architecture
  - Training set creation
  - Test set creation
  - Classification

Evaluation
  - System parameters
  - Experimental results
Motivation
sachsMedia Project

- Annotation and Retrieval of Audiovisual Media
  - Video analysis
  - Audio analysis
  - Metadata handling and retrieval
- Graphical User Interfaces
- Digital Content Distribution
  - Digital video broadcasting
  - Next generation networks
  - IP-based services
Our partners (SME):
- TV broadcasters
- Digital distribution providers (Playout + physical distribution)

Benefits of digital archives?
- Access/share archives in production (re-use)
- Make archives available to consumers

Classification of Video and providing RSS channels is a use case
- Category mapping
- Term extraction with JWPL (UKP lab)
- Word processing (tokenization, stopword removal and stemming)
- Store training term dictionaries (TRTD) per category
System Architecture
Test Set Creation

- Parsing ASR transcripts
- Filter textual content
- Filter metadata content (optional task 2)
- Word processing (tokenization, stopword removal and stemming)
System Architecture
Classification

- Classifier training
  - Create instances from TRTDs (text to numerical representation)
  - Store classifiers
- ASR classification on term-by-term basis
  - Create instances from TSTDs
  - Load classifiers
  - Classification with Weka toolkit (4-NN, Naive Bayes)
  - Normalization + RSS Feed Generation
System Architecture
Normalization

- Training term dictionaries
  - Maximum terms threshold (cat. balancing)
  - Duplicate removal threshold (cat. discrimination)
- Test term dictionaries
  - Duplicate removal threshold (cat. discrimination)
- Classification
  - Intra-category normalization (doc. length)
  - Inter-category normalization (mean + std. dev.)
System Architecture Translation

- Parse RSS Feeds for textual attributes
- Sentence splitter
- Translation with Google’s AJAX Language API
- Write translated RSS Feeds
Training
- D: depth of wikipedia category extraction
- FS: frequency-based term selection
- TMAX: maximum number of training terms
- WT: training term duplicate removal rate

Classification/Testing
- VT: test term duplicate removal rate
- C: used classifiers (k-NN + Naive Bayes)
## Evaluation

### Experimental Results

<table>
<thead>
<tr>
<th>Run ID</th>
<th>D</th>
<th>FS</th>
<th>TMAX</th>
<th>WT</th>
<th>VT</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>cut_c1r1</td>
<td>3</td>
<td>top</td>
<td>3000</td>
<td>2</td>
<td>0.5</td>
<td>0.15</td>
<td>0.14</td>
</tr>
<tr>
<td>cut_c1r2</td>
<td>4</td>
<td>top</td>
<td>5000</td>
<td>5</td>
<td>0.5</td>
<td>0.10</td>
<td>0.12</td>
</tr>
<tr>
<td>cut_c2r1</td>
<td>3</td>
<td>top</td>
<td>3000</td>
<td>2</td>
<td>0.5</td>
<td>0.13</td>
<td>0.12</td>
</tr>
<tr>
<td>cut_c2r2</td>
<td>4</td>
<td>top</td>
<td>5000</td>
<td>5</td>
<td>0.5</td>
<td>0.12</td>
<td>0.14</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Assessor 1</th>
<th>Assessor 2</th>
<th>Assessor 3</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>fluency</td>
<td>2.88</td>
<td>2.65</td>
<td>2.93</td>
<td>2.82</td>
</tr>
<tr>
<td>adequacy</td>
<td>3.53</td>
<td>3.15</td>
<td>3.80</td>
<td>3.49</td>
</tr>
</tbody>
</table>
Classification
- Room for improvement
  - Completely blind approach
  - No language specific parameter settings

Translation
- Translating the RSS-Feeds is not the main problem
Summary

Experiment Conclusions (2)

- Improvements?
  - Normalization during training stage
  - Tuning/omitting parameters
  - Detecting Dutch and English terms in ASR outputs (remove noise)
  - Learning class distributions from development data
- Comparison to simple retrieval approach
- Creating RSS-TV channels