Daniel Richter, Jens Kürsten
Dept. Computer Science and Media
Chemnitz University of Technology

VideoCLEF 2008: ASR Classification based on Wikipedia Categories

CHEMNITZ UNIVERSITY
OF TECHNOLOGY

Bundesministerium
für Bildung
und Forschung

INNOPROFILE
UNTERNEHMEN REGION
Die BMBF-Innovationsinitiative Neue Länder

CHEMNITZ UNIVERSITY OF TECHNOLOGY

Outline

- Motivation
- System design and architecture
  - Training set creation
  - Test set creation
  - Classification
- Evaluation
  - System parameters
  - Experimental results

CHEMNITZ UNIVERSITY OF TECHNOLOGY

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

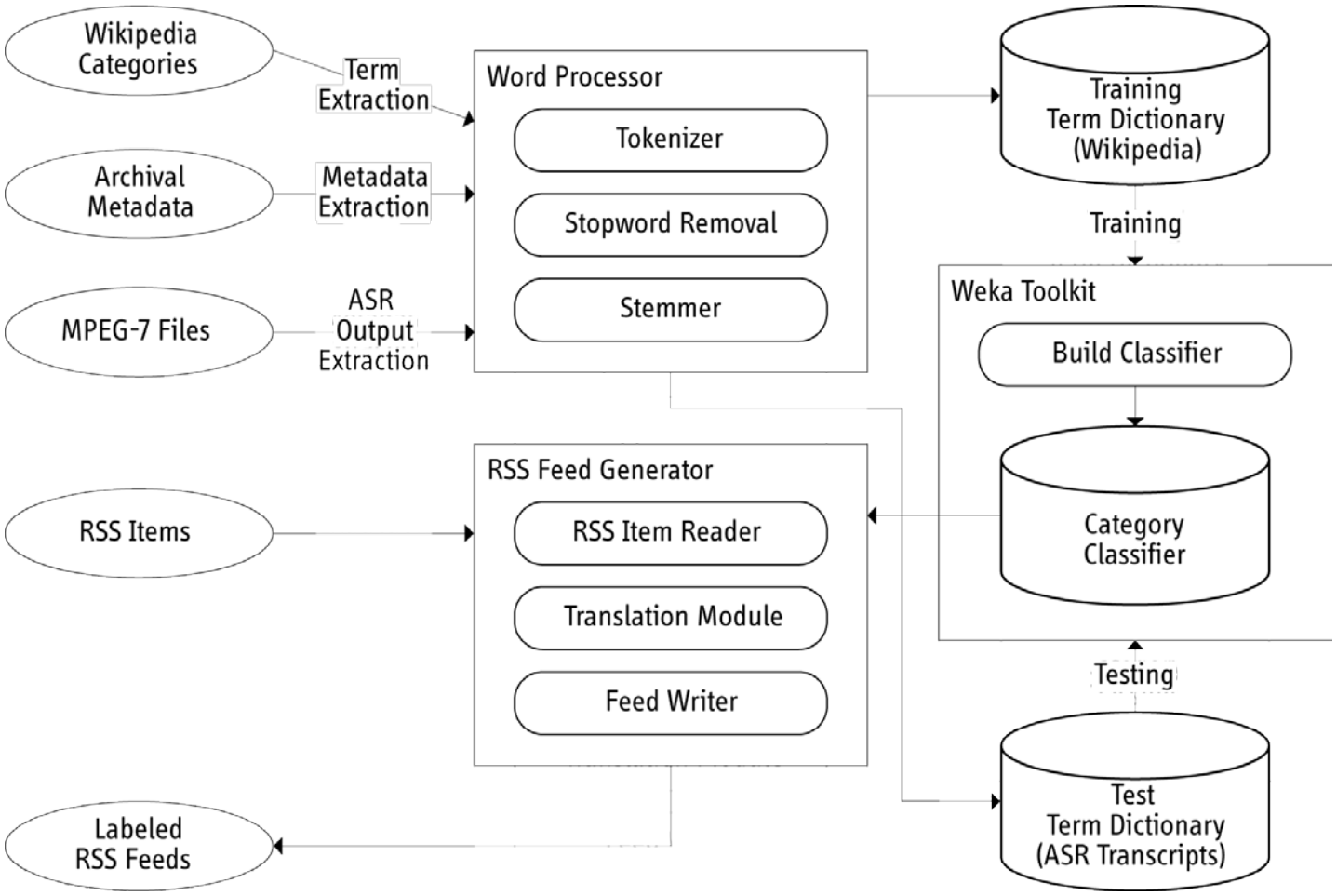
CHEMNITZ UNIVERSITY OF TECHNOLOGY

Motivation
Relation to VideoCLEF

- 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

CHEMNITZ UNIVERSITY OF TECHNOLOGY

System Architecture



CHEMNITZ UNIVERSITY OF TECHNOLOGY

System Architecture Term Extraction (Wikipedia)

- Category mapping
- Term extraction with JWPL (UKP lab)
- Word processing (tokenization, stopword removal and stemming)
- Store training term dictionaries (TRTD) per category

CHEMNITZ UNIVERSITY
OF TECHNOLOGY

System Architecture
Test Set Creation

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

CHEMNITZ UNIVERSITY OF TECHNOLOGY

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

CHEMNITZ UNIVERSITY OF TECHNOLOGY

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.)

CHEMNITZ UNIVERSITY OF TECHNOLOGY

System Architecture Translation

- Parse RSS Feeds for textual attributes
- Sentence splitter
- Translation with Google's AJAX Language API
- Write translated RSS Feeds

CHEMNITZ UNIVERSITY OF TECHNOLOGY

Evaluation System Parameters

- 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)

CHEMNITZ UNIVERSITY OF TECHNOLOGY

Evaluation Experimental Results

Run ID	D	FS	TMAX	WT	VT	Precision	Recall
cut_c1r1	3	top	3000	2	0.5	0.15	0.14
cut_c1r2	4	top	5000	5	0.5	0.10	0.12
cut_c2r1	3	top	3000	2	0.5	0.13	0.12
cut_c2r2	4	top	5000	5	0.5	0.12	0.14

Criterion	Assessor 1	Assessor 2	Assessor 3	Average
fluency	2.88	2.65	2.93	2.82
adequecy	3.53	3.15	3.80	3.49

CHEMNITZ UNIVERSITY OF TECHNOLOGY

Summary Experiment Conclusions (1)

- Classification
  - Room for improvement
    - Completely blind approach
    - No language specific parameter settings
- Translation
  - Translating the RSS-Feeds is not the main problem

CHEMNITZ UNIVERSITY OF TECHNOLOGY

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