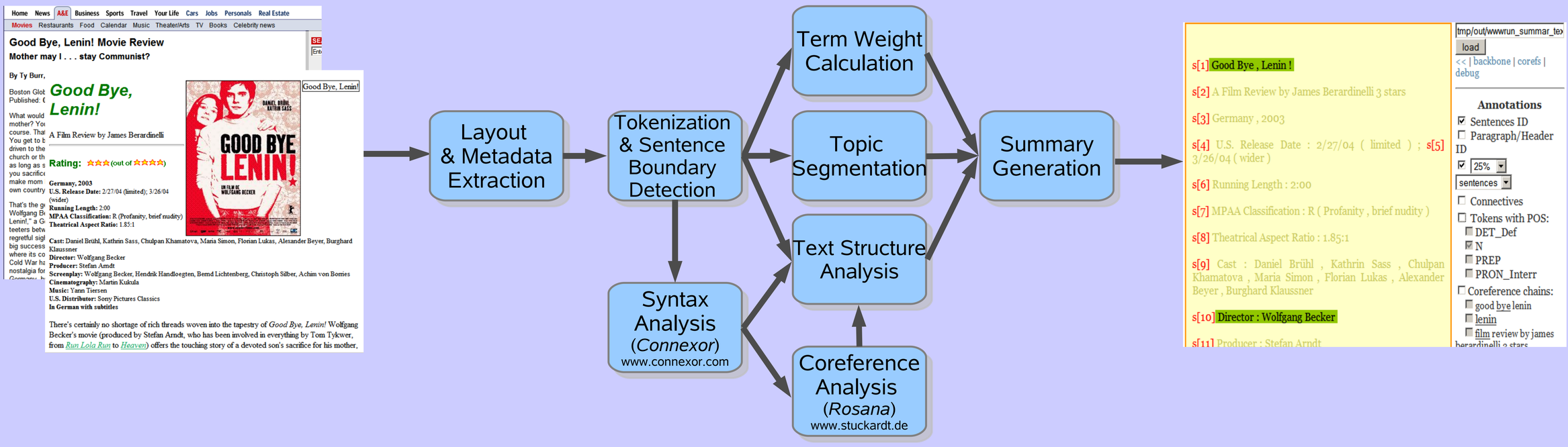
SUMMaR: Combining Linguistics and Statistics for Text Summarization

Applied Computational Linguistics Lab
Department of Linguistics
University of Potsdam, Germany

Manfred Stede
Heike Bieler
Stefanie Dipper
Arthit Suriyawongkul

{stede, bieler,dipper,art}@
ling.uni-potsdam.de



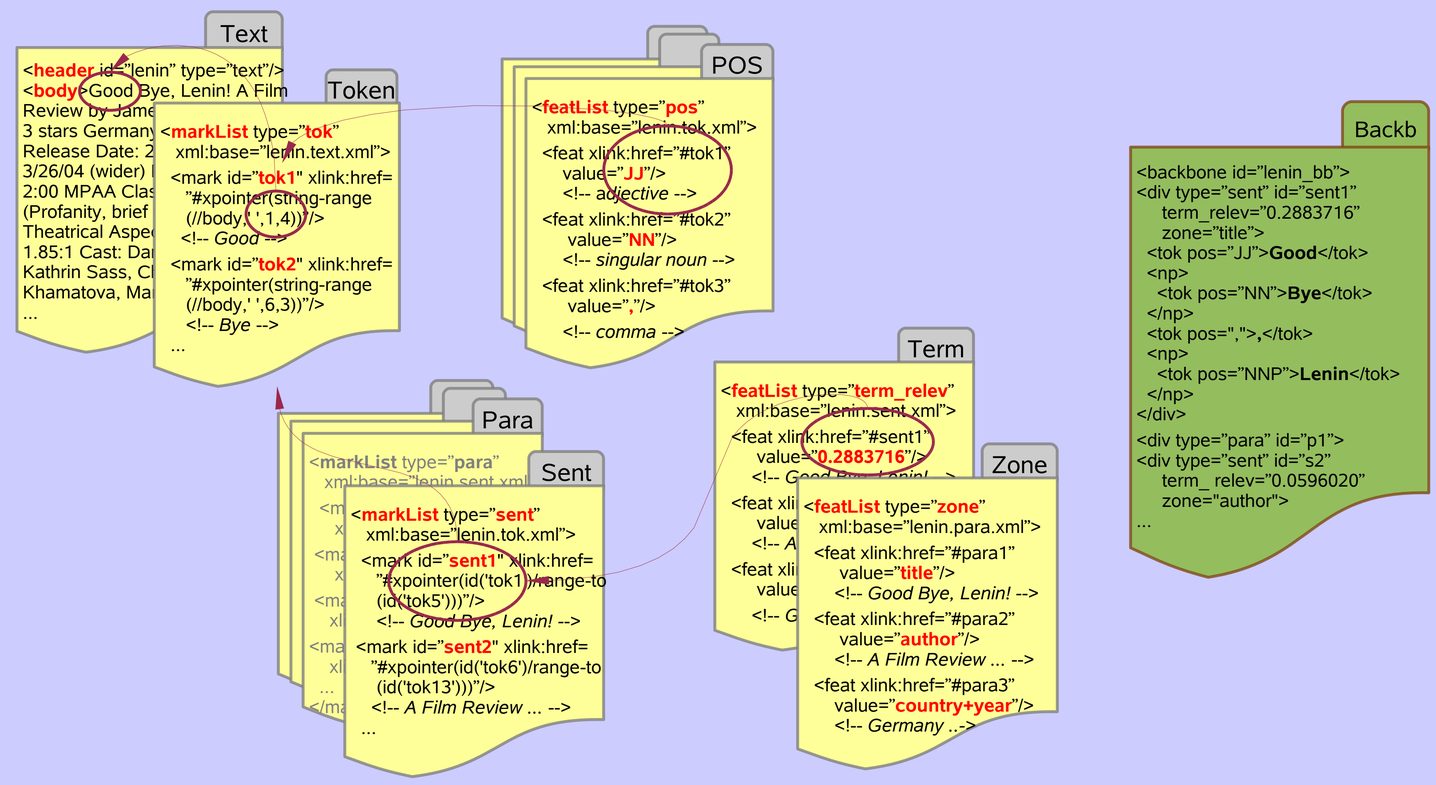
Data Structures and Manipulation

Standoff XML representations

- Source text and individual annotations (token, sentence and paragraph boundaries, POS, term relevances, content zones, ...) are kept apart
- <mark> tags for unit selection, <feat> tags for annotations
- Linking via XPointer expressions
- Straightforward representation of:
  - overlapping annotations
  - annotations from different sources
  - additional annotations "on request"

"Backbone": input to sentence extractor

- Integrated inline version of most relevant information
- Generation of backbone
  - input: freely-selected list of annotations
  - if necessary: uses milestones



Background

Automatic Text Summarization

Basic steps

- Identify the most important sentences in a document
- Extract these sentences
- Apply re-generation techniques to ensure the summary's coherence
- Preserve document layout

Key question: What is *important*?

Answer 1, for robustness: Use statistics, following Luhn 58, Edmundson 69

- Term relevance scoring: the most frequent terms are taken to be characteristic for the document and thus indicate importance of sentences
- Document frequencies: compute term relevance relative to domain vocabulary (tf*idf, Sparck-Jones 72)
- In addition, use features such as sentence position (beginning, end of paragraph), heading words, and generic cue words („importantly“, „we emphasize“, ...)

Answer 2, for quality: It depends on the type of text, really -> See next box!

Multi-document summarization: Given >1 documents on the same topic, identify portions of identical / different / conflicting information in the documents, and produce a single summary

Sample application scenario: Multiple reviews of the same film

Document Structure

Logical structure: identification of headers and paragraph breaks

- For XML and HTML: map the existing structure
- For plain text documents: heuristic rules considering average length of lines

Content structure: semantic labelling

- Label sets are associated with text-sort ontology (e.g.: article – opinion article – review – film review)
- Labels for film reviews include TITLE, RATING, DESCRIBE-CONTENT, COMMENT-ON-ACTORS, ... (total: 40 labels, organized in hierarchy)
- Procedure for identifying content zones
  - Identify „simple“ zones (such as AUTHOR) with strict rules encoding sufficient conditions
  - Assign probabilities to more difficult zones with heuristic rules not considering zone context
  - Employ 3-gram model of zone labels to assign probabilities to remaining zones based on context information; introduce a temporary <DESCRIPTION-OR-COMMENT> label for running text
  - Employ a statistical classifier for making the distinction between DESCRIPTION and COMMENT
    Evaluation: 158 paragraphs: 88.71% / 86.46% accuracy

Benefit (for our example): Ensure that the summary includes both a short description of the film and a synopsis of author's opinion

SUMMaR Procedure

- Pre-Processing:
  - Logical document structure
  - Tokenization
  - Sentence splitting
- Syntactic analysis followed by pronoun resolution
- Content zone identification
- Local rhetorical parsing within paragraphs (following RST)
- Sentence weight calculation based on
  - Term weights
  - Text-sort-specific content zone considerations
  - RST-nuclearity of segments
- Sentence extraction
- Partial re-generation
  - Avoid dangling pronouns: replace with antecedent NP if available
  - Avoid ellipses, dangling comparatives and connectives

PinK
Bundesministerium für Bildung und Forschung
UNTERNEHMEN REGION
Die BMBF-Innovationsinitiative Neue Länder