Lecture:Textmining

Sprachverarbeitung (NLP) und Text Mining

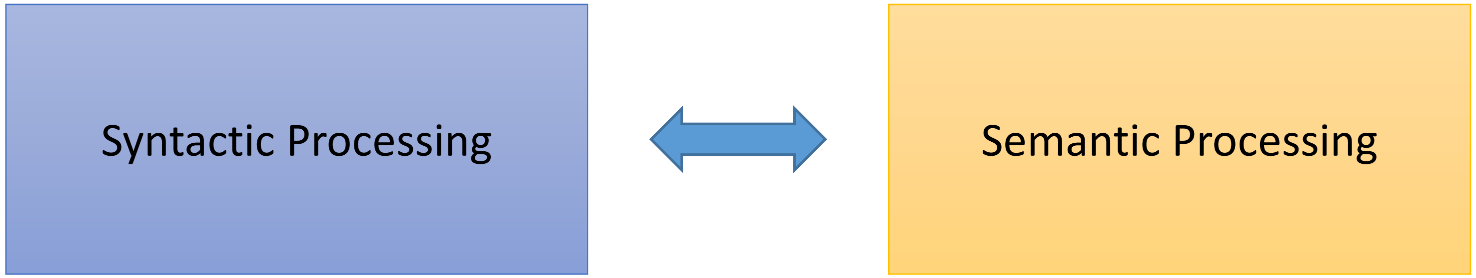
- Lectures:
  - Prof. Dr. Andreas Hotho
  - Time: Wednesday 10:15-11:45
  - Location (starting 27.10): On-site, Übungsraum II (AH 003) (Computer Science Building, Basement)
- Exercises:
  - Janna Omeliyanenko
  - Time: Friday 12:15-13:45, Friday 14:00-15:30 (starting 05.11) (on-site: Seminarraum III (A 005) (Computer Science Building, Ground floor))
  - You will get one sheet of exercises (and after a week the solutions to it)
  - There will be one appointment per week, where you can ask questions
  - There will be separate exercises that you can complete and earn a bonus for the exam
- Exam:
  - Date will be announced as soon as it is scheduled
- Literature:
  - D. Jurafsky & J. Martin: Speech and Language Processing, Pearson, 2009, 3rd edition.
  - Scientific Papers (linked in WueCampus)
- Language: To be determined today

Content

- What are you going to learn?

  ➔In essence how to deal with text!

- History for text mining has brought up a lot of interesting tasks and solutions, but in essence we can group them into two:

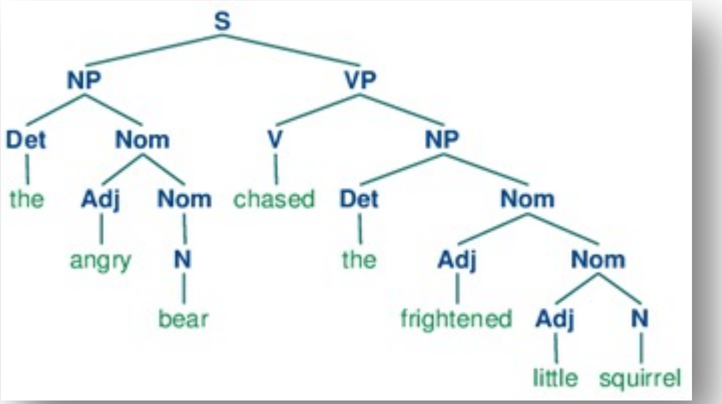


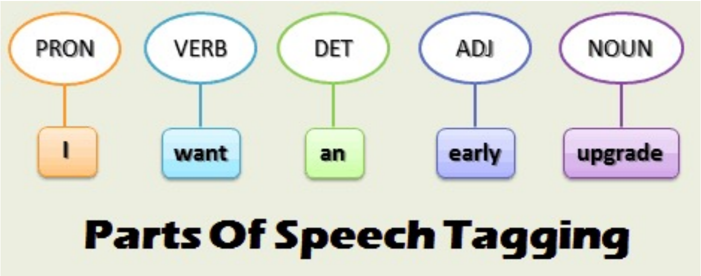
Syntactic Processing

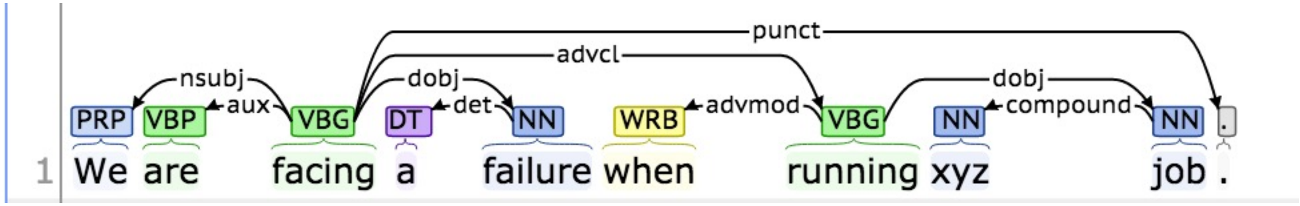
Syntactic Processing

- We are trying to model the „grammar“ of natural language

1. Tokenization, Sentence Splitting, Word Normalization
2. Part of speech Tagging
3. Syntactical Parsing







Tokenization

Syntactic Processing

- Input: Plain text
- Task: Split the text into tokens
- Output: Token annotations
- Example:

Aber nun zu eine Geschichte! Wer erzählt uns eine Geschichte?

⇓

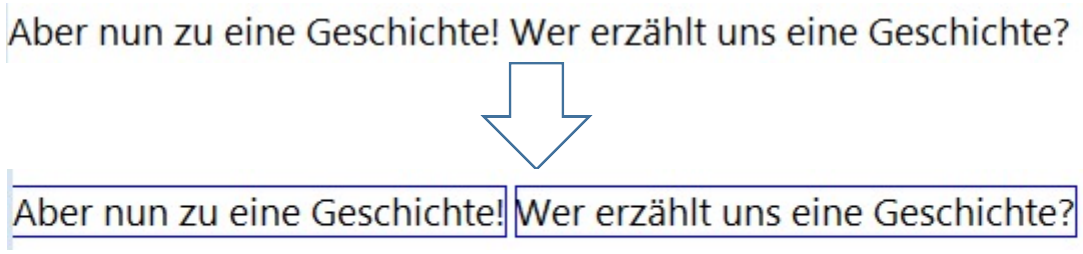
| Aber | nun | zu | eine | Geschichte | ! | Wer | erzählt | uns | eine | Geschichte | ? |

- Problems: Abbreviations, domain specific tokens (paragraphs, phone numbers, etc...)
- Techniques: Regular expressions, grammars, machine learning
- Tools: „Stanford Segmenter“, „OpenNLP-Tokenizer“
- Problem class: Sequence classification

Sentence Splitting

Syntactic Processing

- Input: Plain text, sometimes token annotations
- Task: Split the text into sentences
- Output: Sentence annotations
- Example:

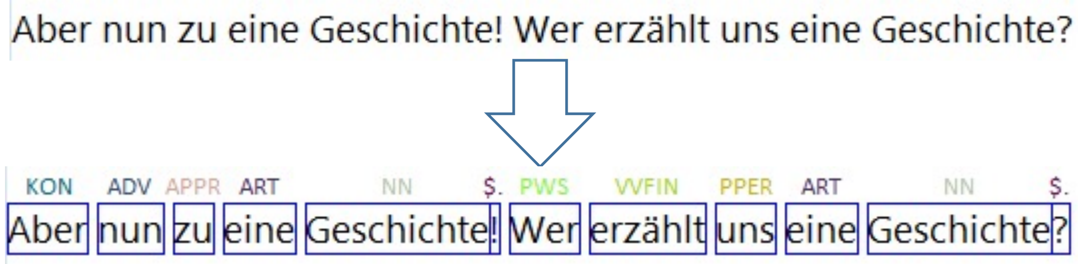


- Problems: Abbreviations, ambiguities e.g. with a semicolon ";".
- Techniques: Regular expressions, machine learning
- Tools: „Stanford Segmenter“, „OpenNLP-SentenceRecognition “
- Problem class: Sequence classification

Word Type Recognition

Syntactic Processing

- Input: Plain text, tokens, mostly sentences
- Task: Determine the word types of the individual tokens
- Output: Part-of-Speech-Tags („POSTags“) for tokens
- Example:

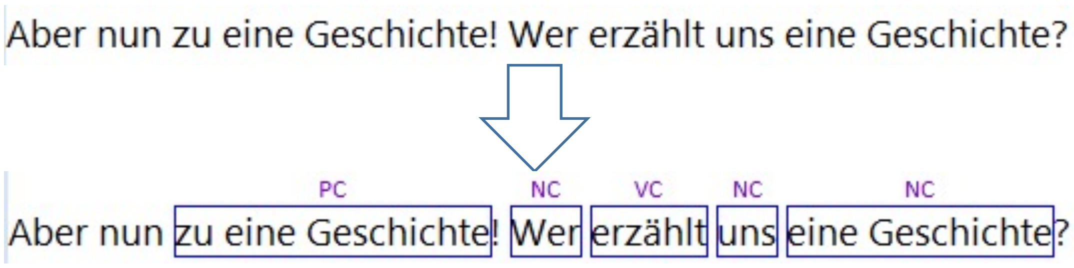


- Problems: ambiguities (e.g. “cut"), unknown words
- Techniques: Rule-based algorithms, machine learning
- Tools: „RFTagger“, „TreeTagger “, „Brill-Tagger“, „CleverTagger“ …
- Problem class: Sequence classification

Phrase Recognition

Syntactic Processing

- Input: Plain text, tokens + POSTags, mostly sentences
- Task: Determine the minimal phrases (chunks)
- Output: Minimal phrases (chunks) and their type
- Example:



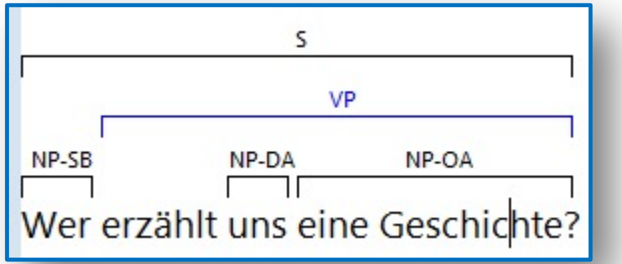
- Problems: Not always possible in German!
- Techniques: Rule-based algorithms, machine learning
- Tools: „TreeTagger “, „OpenNLPChunker“ …
- Problem class: Sequence classification

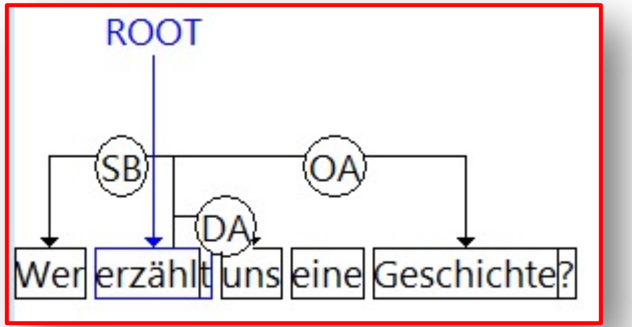
Parsing

Syntactic Processing

- Input: Plain text, tokens, sentences, sometimes POSTags
- Task: Determine the syntactic parse tree of a sentence
- Output: Usually constituent phrases or dependency edges
- Example:

Wer erzählt uns eine Geschichte?

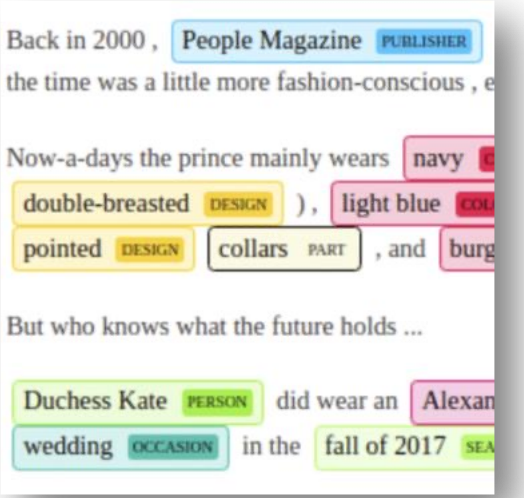




- Problems: Ambiguities, dialects, domains
- Techniques: Grammar-based algorithms, machine learning
- Tools: Mate Parser, Berkeley Parser, Stanford Parser, Parsey McParseface, ParZu,…
- Problem class: Structural Hierarchical Classification, Sequence Classification

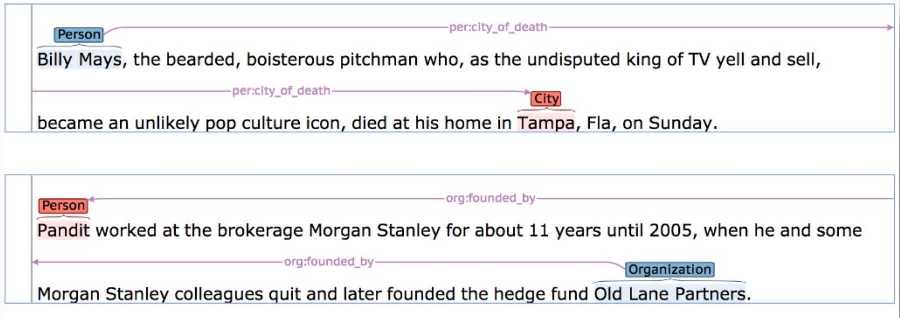
Semantic Processing

- We are trying to model the „meaning“ of the text
  1. Named Entity Recognition
  2. Relation Classification
  3. Coreference Resolution
  4. …



Example: Coreference

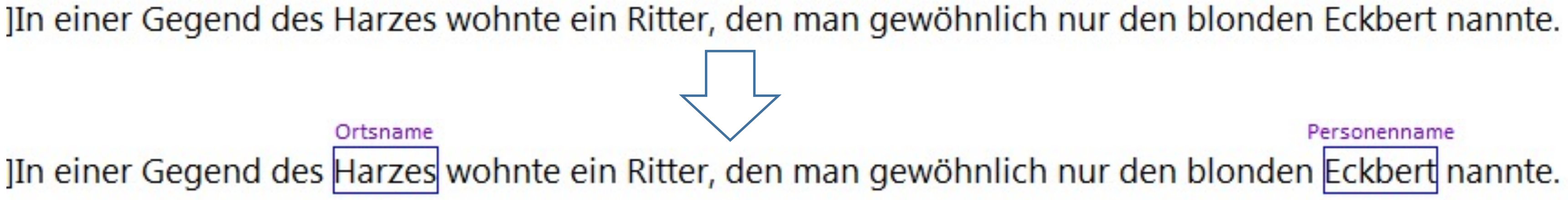
$[\text{Barack Obama}]^1_1$ nominated $[\text{Hillary Clinton}]^2_2$ as $[[\text{his}]^1_3$ secretary of state$]^3_4$ on $[\text{Monday}]^4_5$. $[\text{He}]^1_6$



(Named) Entity Recognition

Semantic Processing

- Input: The plain text, tokens, sentences, chunks, mostly POSTags
- Task: Recognize entities in a sentence/document (e.g. person name, place name)
- Output: Phrases that represent entities
- Example:

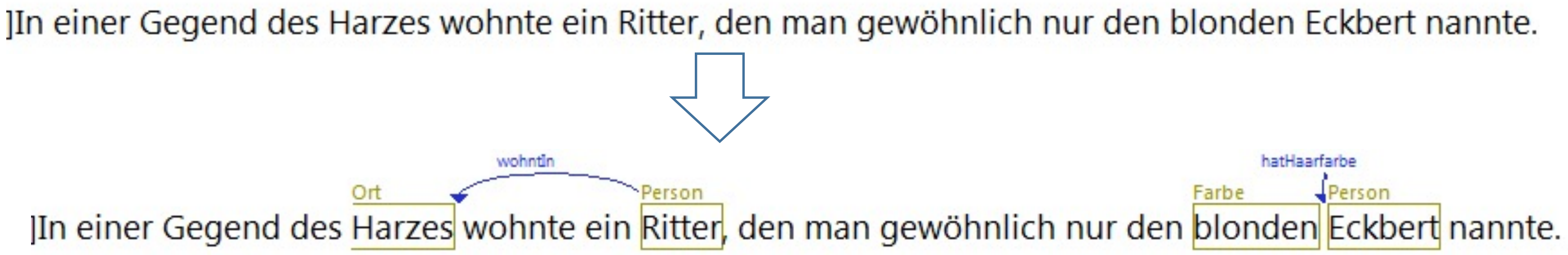


- Problems: Ambiguities, dialects, domains
- Techniques: Grammar-based algorithms, machine learning
- Tools: Mate Parser, Berkeley Parser, Stanford Parser, Parsey McParseface, ParZu,…
- Problem class: Structural Hierarchical Classification, Sequence Classification

Relation Detection

Semantic Processing

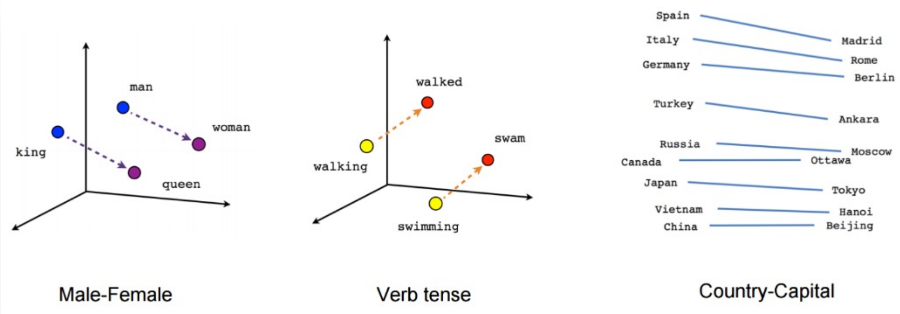
- Input: The plain text, tokens, sentences, chunks, mostly POSTags, entities
- Task: Recognize relations between entities
- Output: Relations between phrases
- Example:



- Problems: Ambiguities, data scarcity
- Techniques: Rule-based algorithms, machine learning
- Tools: ??? (object of research)
- Problem class: (non-structural) Hierarchical Classification

Modelling the text

- Instead of trying to operate on specific aspects, we could model the text
  1. Vector Semantics („Embeddings“)
  2. Language Models
  3. Topic Modelling



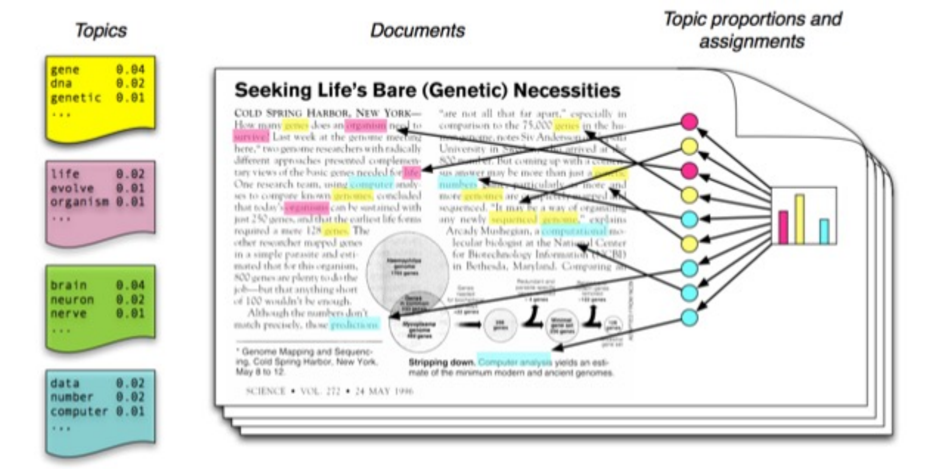
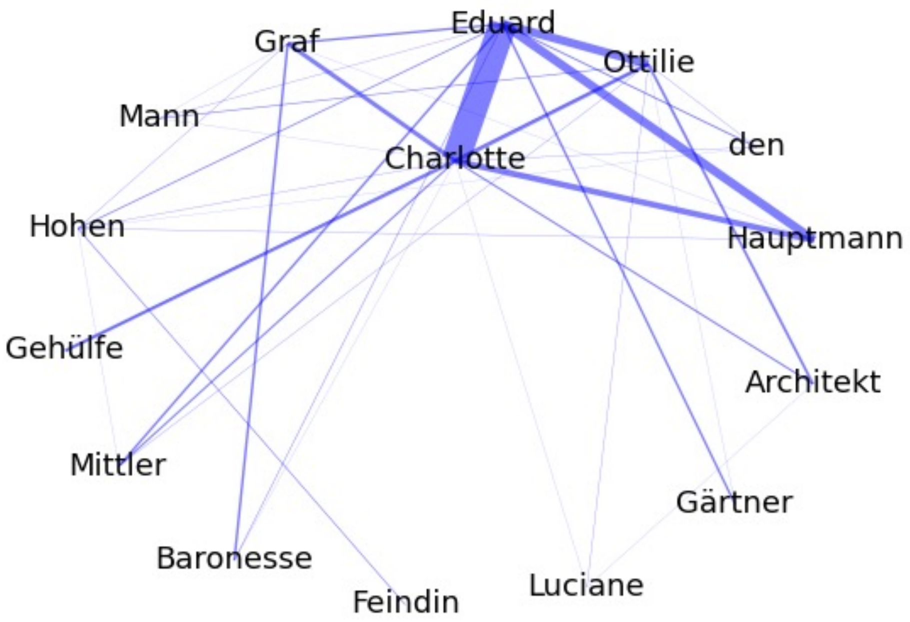


Figure source: Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77-84.

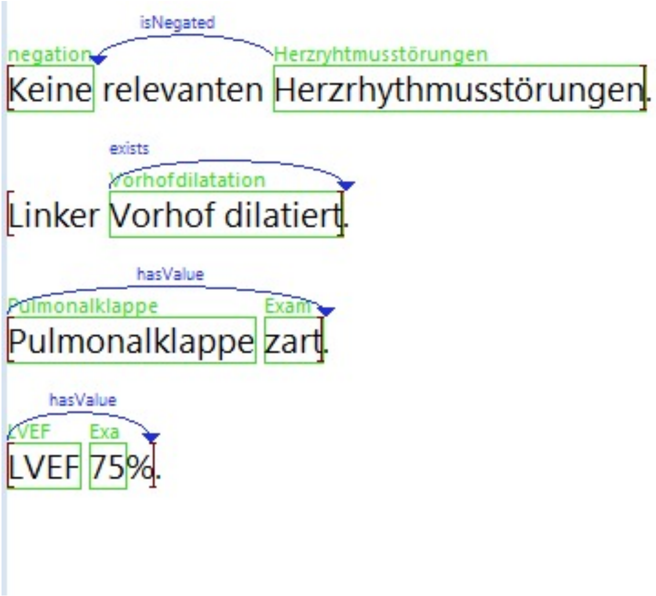
Language Processing at Chair VI

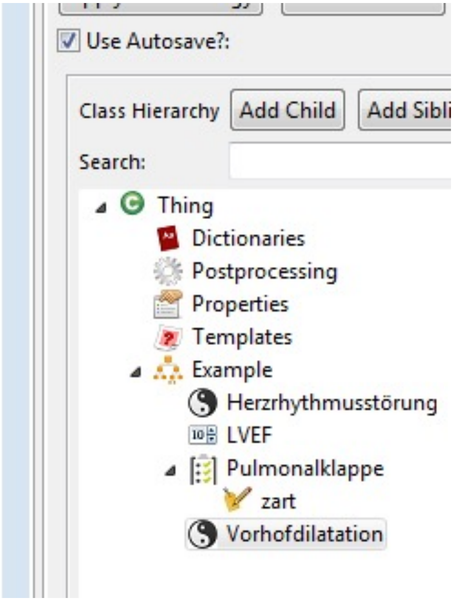
Creating character networks for novels:



Language Processing at Chair VI

Information extraction from medical reports:





Language Processing at Chair X

Automatic detection of novel genres:

- There are a lot of genres that aren't even properly defined yet!
- First successes:

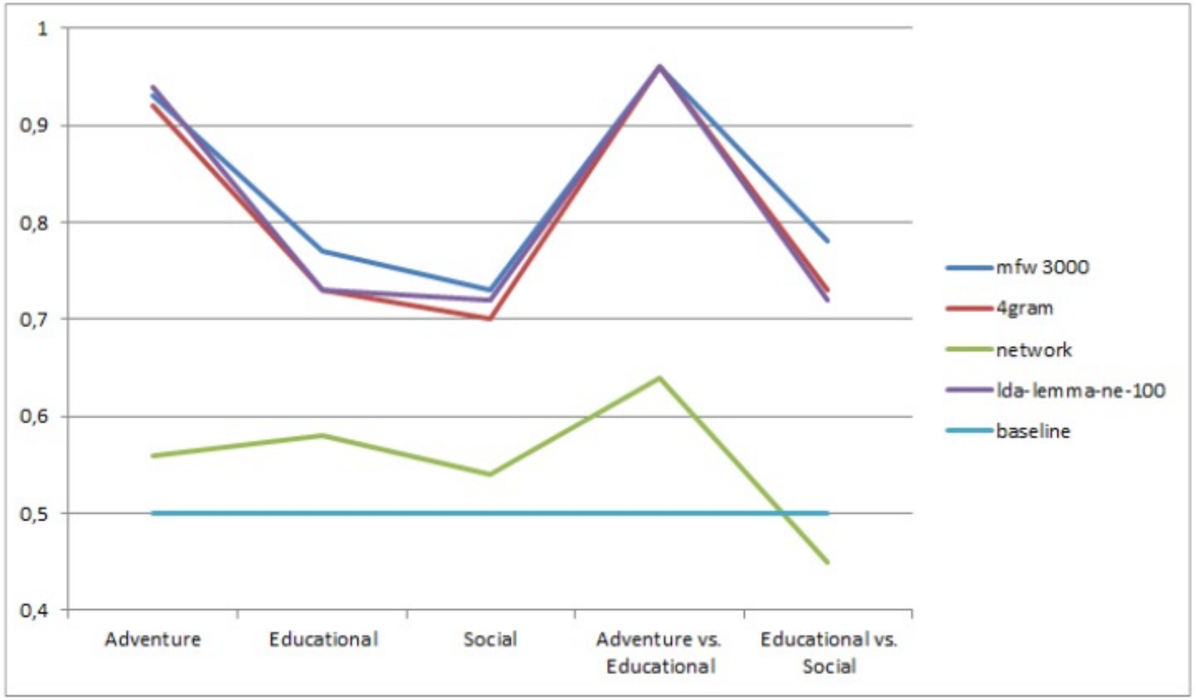
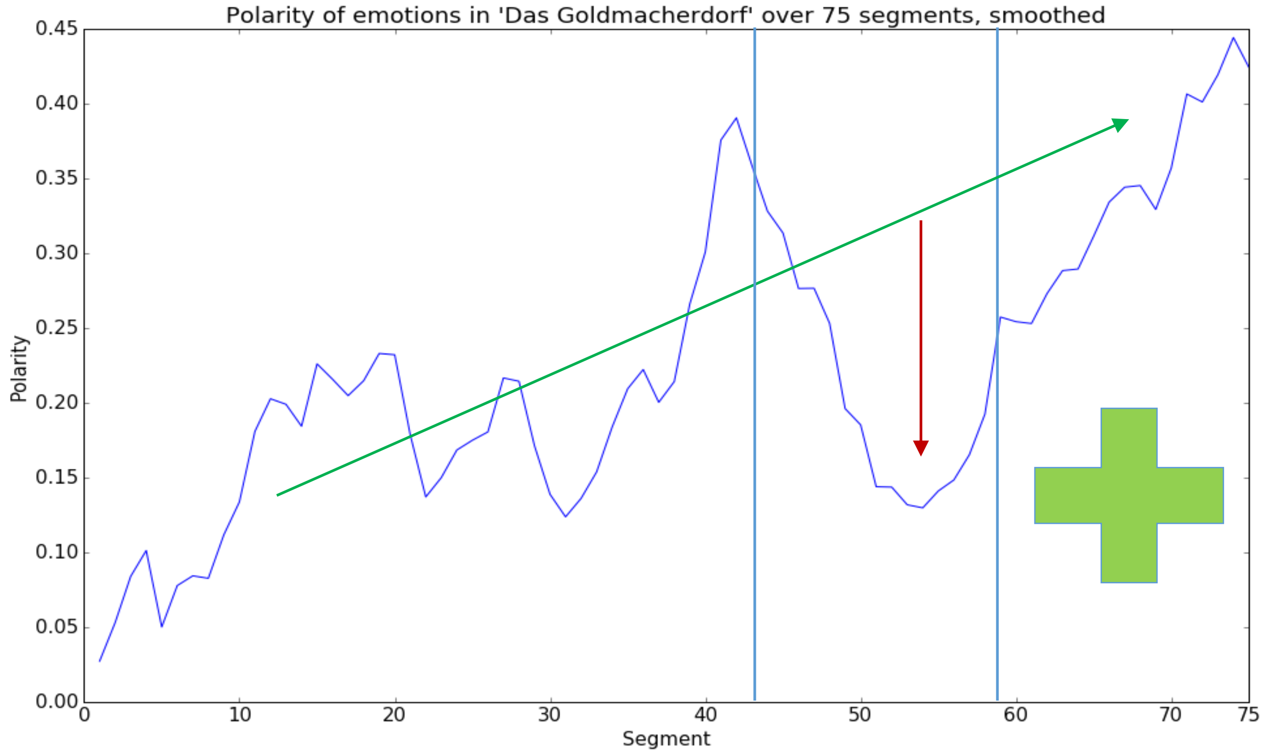


Fig. 2: Accuracy for different scenarios and feature sets including the majority vote baseline.

Language Processing at Chair X

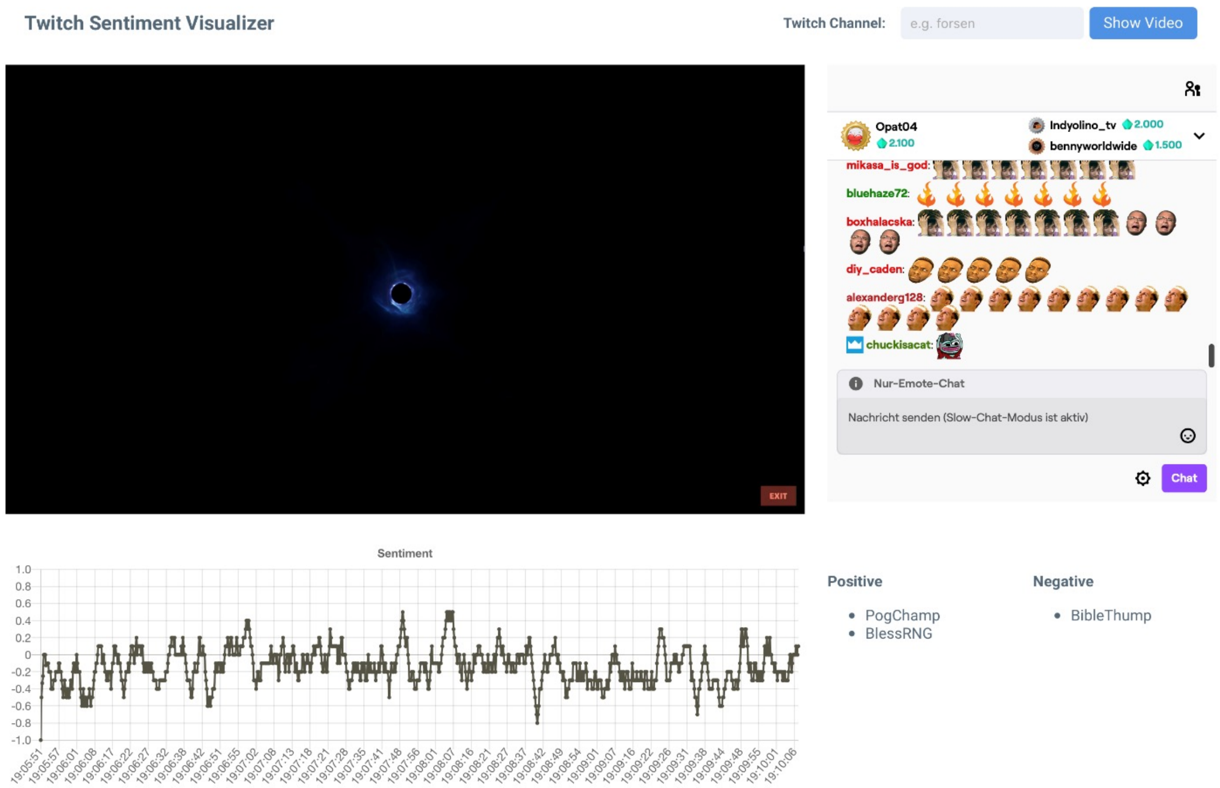
Sentiment analysis in novels, e.g., happy ending detection:

„Das Goldmacherdorf“ by Heinrich Zschokke



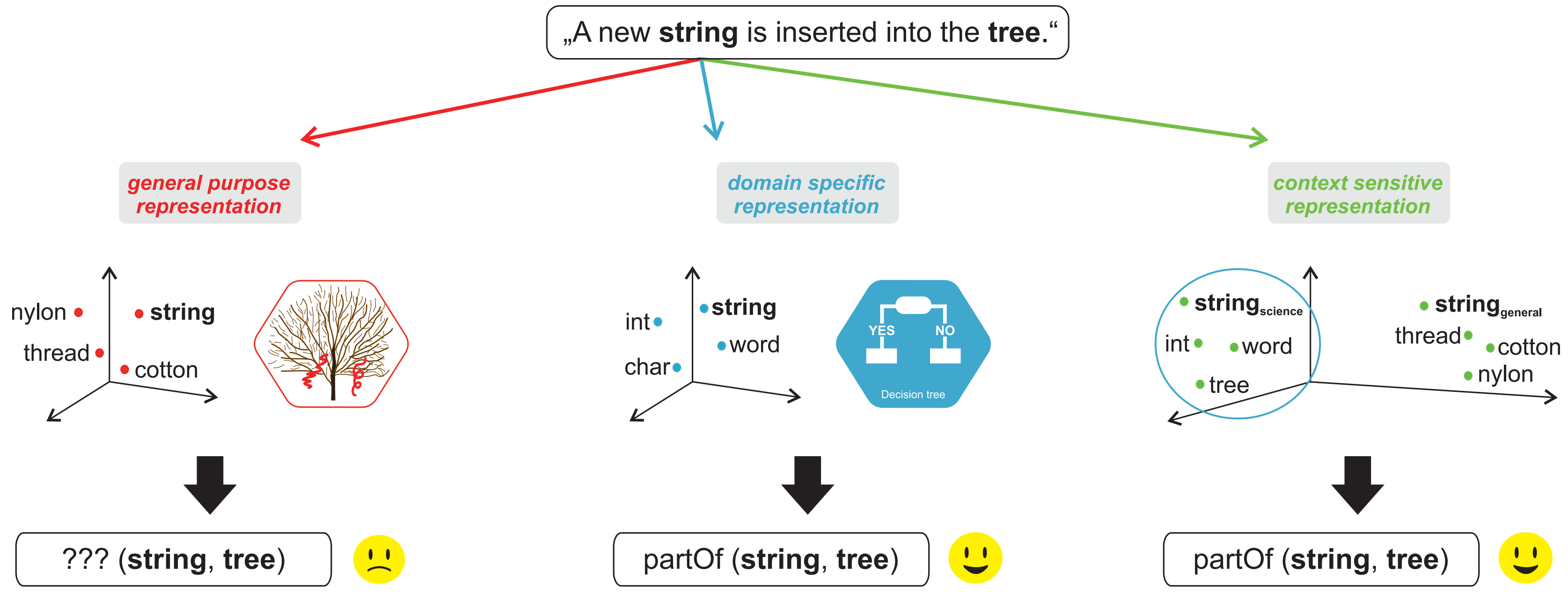
Language Processing at Chair X

Sentiment analysis on twitch.tv



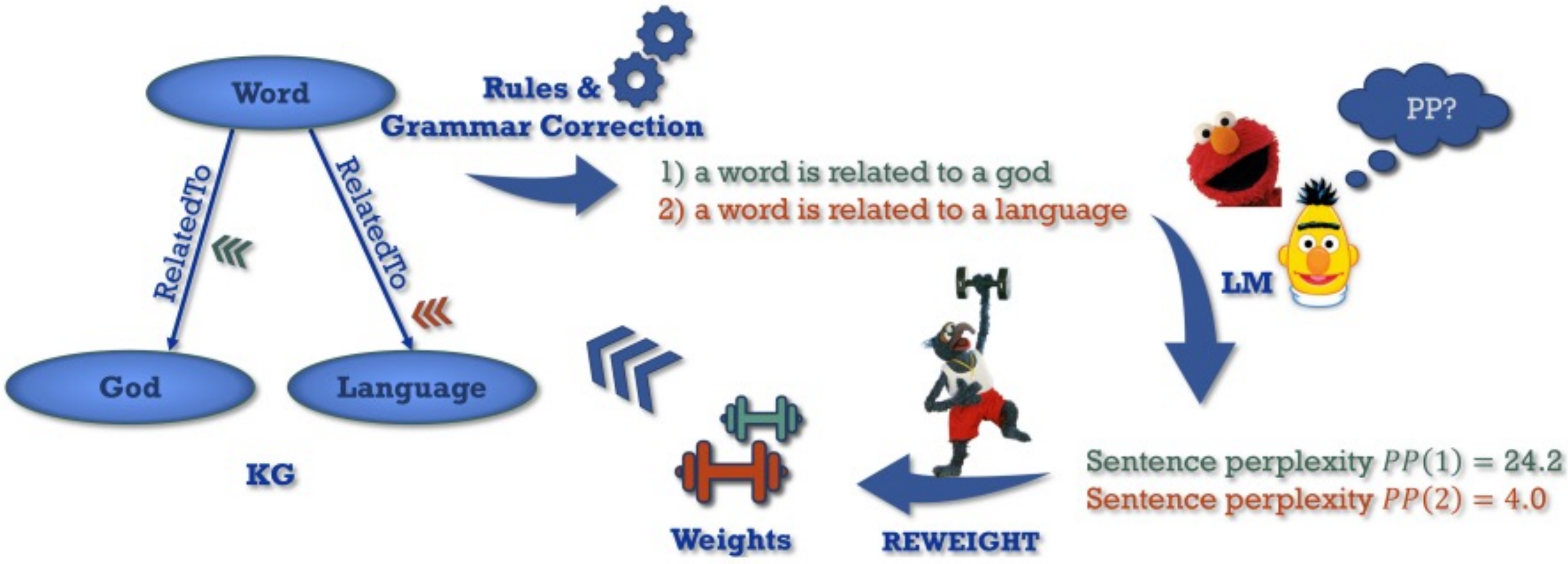
Language Processing at Chair X

Relation classification



Language Processing at Chair X

Atomatic relation weighting in knowledge graphs



Language Processing at the Chair (Advertising)

- We offer BA, MA and internship (Praktika) for text mining (but also for other areas):
  → http://www.dmir.uni-wuerzburg.de/teaching/theses/

- HiWi positions
  → http://www.dmir.uni-wuerzburg.de/open-positions/

→ Relevant for all computer science and digital humanities programs

This lecture vs Machine Learning for NLP

- Both lecture can work indepently, even though they might deal with the same tasks!

    - This lecture contains the classical approaches, while MLNLP deals with Neural Networks
    - In this lecture you will learn how to deal with „structure“ (sequences, trees, clusters)
    - You will learn some of the most fundamental algorithms in computer science
        - E.g. The Viterbi or the CKY Algorithm
    - You will learn about the philosophies of Deep Learning without this lecture dealing with it!
    - You will be able to cast any new task into something you can operationalize on