A Flexible NLP Pipeline for Computational Narratology in Literary Text

Thomas Bögel, Jannik Strötgen, Christoph Mayer, Michael Gertz

Database Systems Research Group, Heidelberg University, Im Neuenheimer Feld 348, 69120 Heidelberg, Germany

Context of this Work

Computational Narratology

- context: Digital Humanities
- facilitate annotations from literary scientists
- support hypotheses [1]
- methods: Natural Language Processing

Temporal phenomena

- field of study in narratology
- temporal structure of literary texts
- examples: time shifts, order phenomena (e.g., prolepsis)

The heureCLÉA Project [4]

Cooperation

- BMBF-funded eHumanities project
- narratologists (Hamburg)
- computer scientists (Heidelberg)
- temporal phenomena in literary text

Goals

- collaborative annotation framework that automatically suggests annotations
- reduce manual annotation effort
- analysis of temporal aspects in narrative texts

Temporal Phenomena in Literary Texts

Temporal expressions

- less frequent in literary text (usually)
- can be extracted automatically
- HeidelTime: extraction of explicit temporal expressions [2]

Tense information

- tenses provide information about temporal structures
- shifts in tenses indicate order phenomena

Task: robust annotation of tenses in narrative texts

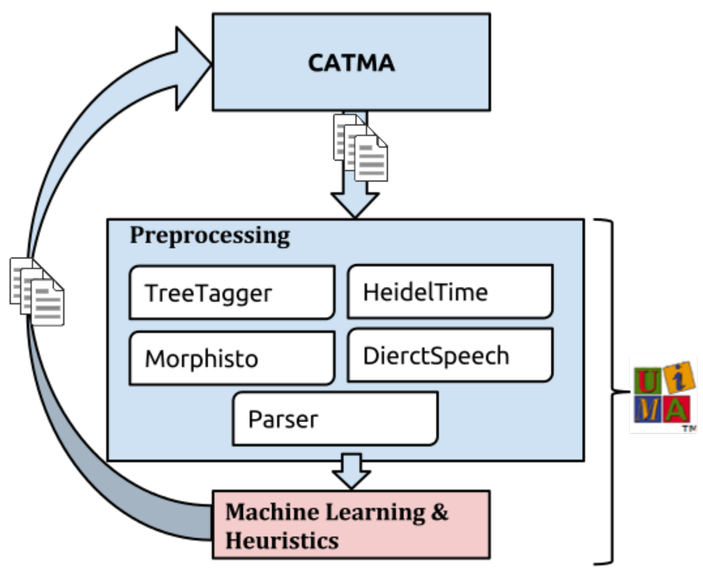
Prior work

- laborious manual annotations
- automatic systems focus on English
- no existing system for German tense annotation

Data set

- German narrative texts (20^{th} century)
- manual annotation by literary scientists
- tagset: narratological aspects

NLP Pipeline Architecture



Component Description and Tools

NLP components

- POS tagging: TreeTagger
- morphology: Morphisto
- time expressions: HeidelTime
- syntactic parsing: Parzu & Stanford parser

CATMA interface

- CATMA: collaborative annotation platform
- flexible CATMA $\leftrightarrow$ UIMA interface
- tailored to narratologists

Machine learning interface

- feature extraction and machine learning
- interchangeable algorithms
- goal: predict annotations automatically

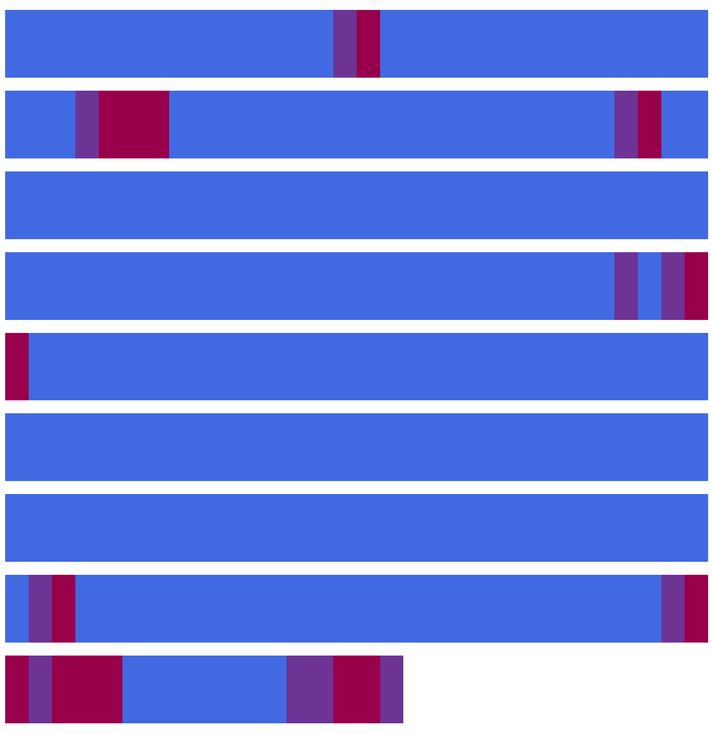
Feedback loop on predicted annotations

- manual corrections
- improvement of future predictions (ML)

Use Case: Tense Annotations

Extracting temporal clusters [3]

- temporal cluster: all tokens governed by the same verb
- approach based on morphological features & heuristics
- exploitation of tense markers (e.g., auxiliaries)
- rule set for combinations of morphological features
- heuristic for sentences with unknown tense
- evaluation: comparison to manual annotations
- high inter-annotator agreement ($\kappa > 0.8$)



Visualization of temporal clusters in "Meine erste Liebe" by Ludwig Thoma.

Evaluation Results

tense	correctly tagged verbs
present	93.10
preterite	95.73
perfect	96.43
pluperfect	84.71
future	90.00

$\Rightarrow$ **reliable and robust prediction of tense clusters**

Ongoing work

- machine learning based system for additional annotations, e.g., narrative levels
- hybrid, self-improving system: heuristics + machine learning

References

[1] I. Mani: Computational Narratology. **The living handbook of narratology**. http://www.lhn.uni-hamburg.de/article/computational-narratology, 2013.

[2] J. Strötgen and M. Gertz: **Multilingual and Cross-domain Temporal Tagging.** *Language Resources and Evaluation*, 47(2), 269–298, 2014.

[3] T. Bögel, J. Strötgen and M. Gertz: **Computational Narratology: Extracting Tense Clusters from Narrative Texts** *LREC*, 2013.

[4] The heureCLÉA Project: http://www.heureclea.de/.

Contact Information:

Thomas Bögel

boegel@uni-hd.de

http://dbs.ifi.uni-heidelberg.de/

This work was presented at **1. Jahrestagung der Digital Humanities im deutschsprachigen Raum (DHd)**, March 25–28, 2014, Passau, Germany.