Gleiche Daten, unterschiedliche Erkenntnisziele?

Zum Potential vermeintlich widersprüchlicher Zugänge zur Textanalyse

Universität Hamburg

Evelyn Gius

Jan Christoph Meister

Janina Jacke

Marco Petris

Universität Heidelberg

Thomas Bögel

Jannik Strötgen

Michael Gertz

DHd 2015 – Graz, Österreich

25. Februar 2015

Übersicht

1. Unterschiedliche Zugänge zur Textanalyse in heureCLÉA
2. Erweiterung des traditionellen Zugangs zu literaturwissenschaftlicher Textanalyse
3. Transparente Entscheidungsprozesse im Bereich der NLP
4. Fazit

Übersicht

1. Unterschiedliche Zugänge zur Textanalyse in heureCLÉA
2. Erweiterung des traditionellen Zugangs zu literaturwissenschaftlicher Textanalyse
3. Transparente Entscheidungsprozesse im Bereich der NLP
4. Fazit

heureCLÉA – einige Eckdaten

Ziel

Entwicklung einer „digitalen Heuristik“: automatische Annotationsvorschläge für die Analyse literarischer Erzähltexte

Vorgehen

1. Kollaborative manuelle Annotation eines Textkorpus mithilfe narratologischer Zeit-Kategorien
2. Analyse und Reproduktion der Annotationen durch einen kombinierten Zugang aus regelbasiertem Vorgehen und Machine Learning
3. Integration der digitalen Heuristik als neues Modul in die Textanalyseplattform CATMA

Datenqualität – disziplinäre Unterschiede

Manuelle Annotation

Interpretations*pluralismus* vs. *Noise*

Automatisierung

Transparenz vs. Qualität der Vorhersage

Übersicht

1. Unterschiedliche Zugänge zur Textanalyse in heureCLÉA
2. Erweiterung des traditionellen Zugangs zu literaturwissenschaftlicher Textanalyse
3. Transparente Entscheidungsprozesse im Bereich der NLP
4. Fazit

Erweiterung des traditionellen Zugangs zu literaturwissenschaftlicher Textanalyse

Ziel

Abmilderung der Noise-Problematik

Maßnahme

close(r) reading durch kollaborative, computergestützte Analysen

1. Dokumentation zugrundeliegender Annahmen durch Annotation
2. Erarbeitung von Annotationsguidelines (Angabe von Umfang, unmarkiertem Fall, Indikatoren, Tagging-Routine und Beispielen)
3. Überprüfung der Analyse durch kollaborativen Zugang und Diskussion strittiger Entscheidungen

Annotation Guidelines

Tagstring	**Unmarkierter Fall**
• Textabschnitte – Mindestgröße: Teilsatz	• Chronologisches Erzählen
Indikatoren auf der Textoberfläche • Zeitausdrücke, die Vorzeitigkeit, Gleichzeitigkeit oder Nachzeitigkeit ausdrücken • Tempuswechsel	
Tagging-Routine 1. Annotation aller nicht chronologisch dargestellten Textpassagen als Prolepse, Analepse, Simullepse oder Achronie. 2. Bei Anachronien: Spezifizierung von Umfang und Reichweite. 3. Bei Achronien: Spezifizierung der Verknüpfungsart.	
Beispiele • chronologisches Erzählen: „von ohngefähr erhob sie das Auge und traf mit dem blauesten Strahle in seinen Blick. Er ward wie von einem Blitz durchdrungen. Sie strauchelte, und so schnell er auch hinzusprang, konnte er doch nicht verhindern, daß sie nicht kurze Zeit in der reizendsten Stellung knieend vor seinen Füßen lag“ (Der Pokal) • Analepse: „Jetzt sah man, was geschehen war: der Hansjörg hatte sich am mittleren Gelenk den Zeigefinger der rechten Hand abgeschossen“ (Die Kriegspfeife) • Prolepse: „Zwanzig Jahre lang habe ich den Tod auf den Tag herbeigezogen, der in einer Stunde beginnen wird [...]“ (Der Tod) • Simullepse: „'Ich bin nicht allein', sagte ich […]. Dabei preßte sich mein Arm, der die Decke über ihren Kopf gelegt hatte, krampfhaft auf jene Stelle, wo ich den Mund vermutete [...]“ (Die Schutzimpfung) • Achronie: „Vorliebe empfindet der Mensch für allerlei Gegenstände. Liebe, die echte, unvergängliche, die lernt er – wenn überhaupt – nur einmal kennen.“ (Krambambuli)	

Abb. 1: Annotation von Ordnung, Zusammenfassung aus den Guidelines (Gius & Jacke, 2014)

Der erweiterte hermeneutische Zirkel

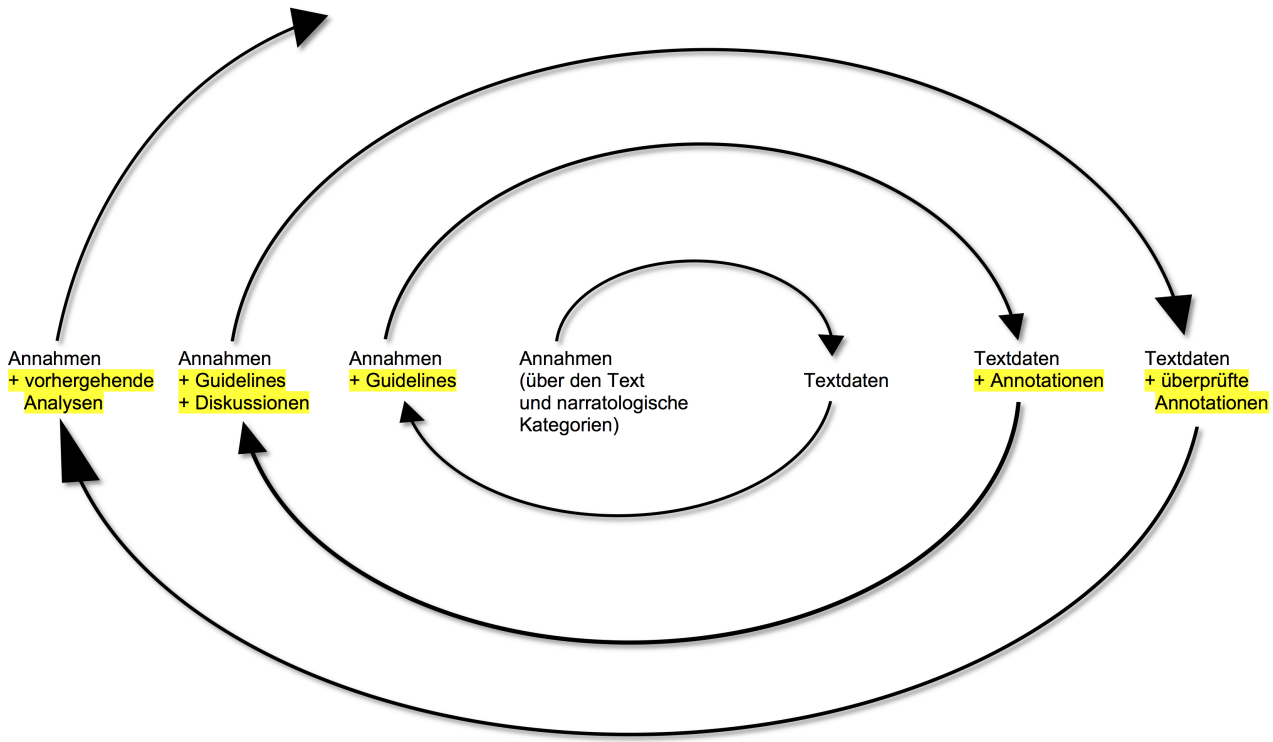


Abb. 2: Der erweiterte hermeneutische Zirkel

Übersicht

1 Unterschiedliche Zugänge zur Textanalyse in heureCLÉA

2 Erweiterung des traditionellen Zugangs zu literaturwissenschaftlicher Textanalyse

3 Transparente Entscheidungsprozesse im Bereich der NLP

4 Fazit

Automatische Vorhersage von Annotationen

Motivation

- Vorhersage von Phänomenen mit zunehmendem Komplexitätsgrad
  - Entlastung der Annotatorinnen bei Oberflächenphänomenen
  - Ableitungen von Regeln für komplexere Phänomene

Hybridansatz: Heuristik + Machine Learning (ML)

Heuristiken

- Keine Trainingsdaten
- Kodierung von Domänenwissen
- Nutzerspezifisch
- Aber: aufwändige Modellierung

Machine Learning

- Automatisches Lernen von Regeln
- Beliebige Komplexität (Spezialfälle)
- Aber: Notwendigkeit von Trainingsdaten

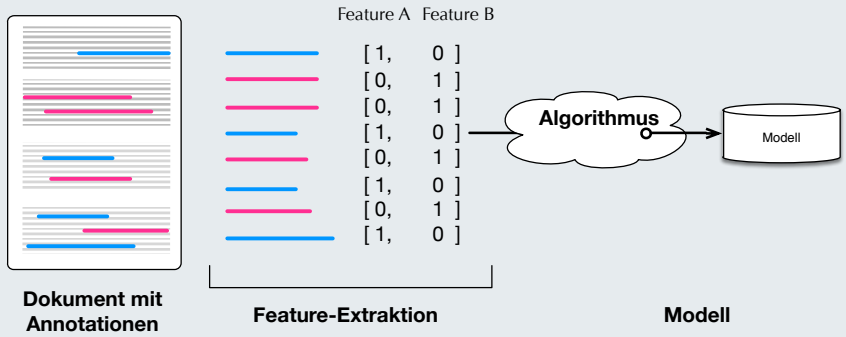
Realisierung von Machine Learning

Anforderungen an den ML-Prozess

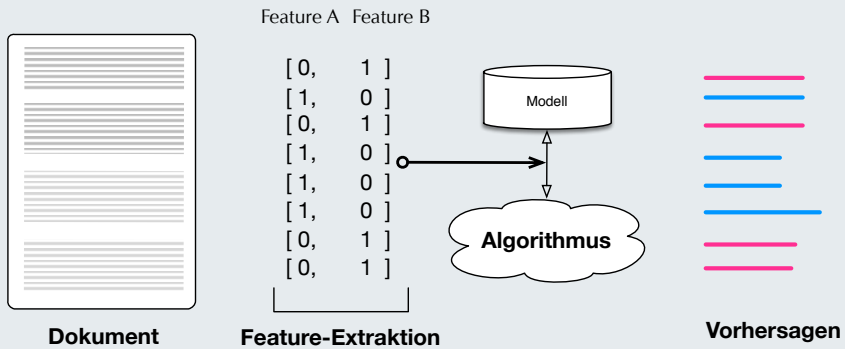
- *Transparenz* von Modell und Algorithmus
- Streitbare Annotationen erfordern streitbare Modelle
- Sichtbare Entscheidungsprozesse unabdingbares Kriterum zur *Akzeptanz* automatischer Annotationen
- Möglichkeit zur *Anpassung* an eigene Entscheidungskriterien
- *Qualität* der Vorhersage

Funktionsweise von Machine Learning

Training von Modellen

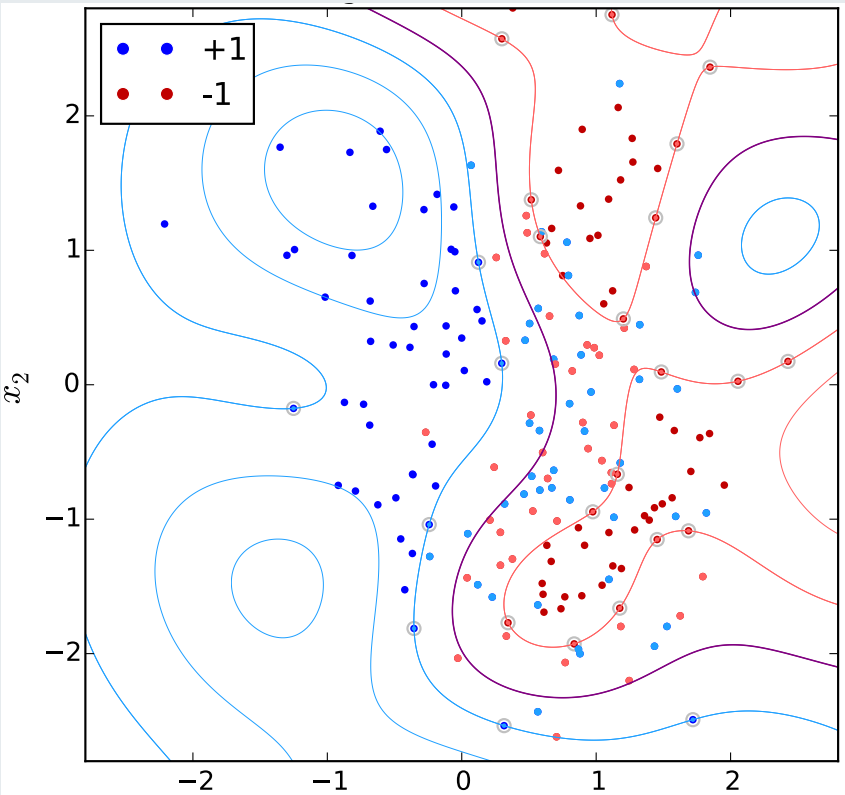


Vorhersage von Annotationen



Eigenschaften von ML-Modellen

Support Vector Machine (Vapnik, 1998)

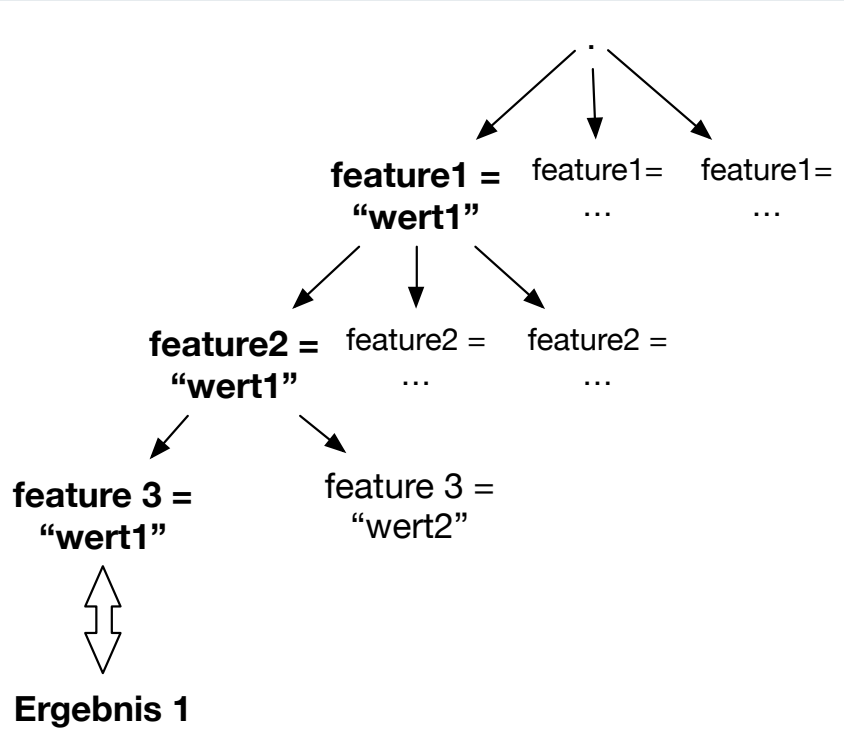


Eigenschaften

- Standard "state-of-the-art"-System
- Gute Ergebnisse
- Intransparenz, Daten als Vektoren
- Entscheidungsgrundlage nicht nachvollziehbar

Eigenschaften von ML-Modellen

Entscheidungsbäume (Quinlan, 1986)



Eigenschaften

- Konkreter Weg zur Entscheidung *nachvollziehbar*
- Maximale *Transparenz*
- Qualität der Vorhersage
- Ermittlung von Featurekonstellationen bei falschen Vorhersagen

Parametrisierung von ML-Modellen

Motivation für Parametrisierung

- Keine absolute Wahrheit für komplexe Annotationen
- Beispiel: Erzählgeschwindigkeit
- Annotationen u.a. von Denkschule abhängig
- Widerspruch zu maschinellem Lernen!

Parametrisierung

1. Initiale Vorhersage
2. Korrektur und Parametrisierung des Modells
3. Finale Vorhersage

Realisierung von Machine Learning

Anforderungen an den ML-Prozess

- ✓ *Transparenz* von Modell und Algorithmus
- ✓ Streitbare Annotationen erfordern streitbare Modelle
  → *Visualisierbares und interpretierbares Modell*
- ✓ Sichtbare Entscheidungsprozesse unabdingbares Kriterum zur *Akzeptanz* automatischer Annotationen
  → *Nachvollziehbare Einzelkriterien*
- ✓ Möglichkeit zur *Anpassung* an eigene Entscheidungskriterien
  → *Parametrisierung des Modells*
- ✓ *Qualität* der Vorhersage

Übersicht

1 Unterschiedliche Zugänge zur Textanalyse in heureCLÉA

2 Erweiterung des traditionellen Zugangs zu literaturwissenschaftlicher Textanalyse

3 Transparente Entscheidungsprozesse im Bereich der NLP

4 Fazit

Zusammenfassung

Textanalyse

- Herausforderung: disziplinär unterschiedliche Auffassungen zur Textanalyse (noise-Problematik)
- Adaption traditioneller Textanalyse durch kollaboratives close reading und Einführung von Annotationsguidelines → Erweiterung des hermeneutischen Zirkels

Automatische Vorhersage

- Herausforderung: Notwendige Transparenz von Vorhersagemodellen
- Flexible Sprachverarbeitungspipeline
- Verwendung transparenter maschineller Lernverfahren und Parametrisierung von Modellen

Referenzen I

Gius, Evelyn, & Jacke, Janina. 2014. *Zur annotation narratologischer kategorien der zeit. guidelines zur nutzung des catma-tagsets*. Hamburg.

Quinlan, J. Ross. 1986. Induction of decision trees. *Machine learning*, **1**(1), 81–106.

Vapnik, Vladimir. 1998. *Statistical learning theory. 1998*. Wiley, New York.