Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

IDS
LEIBNIZ-INSTITUT FÜR
DEUTSCHE SPRACHE

Gefördert durch
DFG Deutsche Forschungsgemeinschaft

Annelen Brunner

ZUR AUTOMATISCHEN ERKENNUNG VON REDEWIEDERGABE-FORMEN – WORK IN PROGRESS

06.07.2019, 2. forTEXT-Expertenworkshop

Mitglied der
Leibniz-Gemeinschaft

WARUM WIEDERGABE ERKENNEN?

- Art und Menge von Rede- (und/oder Gedanken-)Wiedergabe ist ein wichtiger Aspekt des Erzählstils und Teil sehr vieler narratologischer Beschreibungssysteme
  - ➢ Untersuchung von Unterschieden über die Zeit hinweg / zwischen Genres / zwischen Autoren usw.
- In Kombination mit Sprecherzuordnung:
  - ➢ Beziehungsnetze
  - ➢ Auswertungen zu Stil/Themen usw., die Figuren zugeordnet werden können

Jenseits der Literaturwissenschaft:

- Zuordnung von Aussagen zu Quellen (z.B. EMM NewExplorer; https://emm.newsexplorer.eu)
- Textzusammenfassung, Information extraction etc.

WIEDERGABE VON REDE, GEDANKEN UND GESCHRIEBENEM – EINE TYPOLOGIE

Direkte Wiedergabe	Er sagte: **„Ich habe Hunger."**
Indirekte Wiedergabe	Er sagte, **er habe Hunger.**
Erzählte Wiedergabe	**Er beklagte sich über seinen Hunger.**
Freie indirekte Wiedergabe (Erlebte Rede) (Vermischung Merkmale Erzähler/Figurenrede; meist für Gedankenwiedergabe)	Er war ratlos. **Wo sollte er jetzt nur etwas zu essen herbekommen?**
Redebericht (konjunktivische Wiedergabe ohne Rahmenformel)	Sie wusste Rat. **Es gebe gute Restaurants in der Nähe.**

WIEDERGABE VON REDE, GEDANKEN UND GESCHRIEBENEM – EINE TYPOLOGIE

Direkte Wiedergabe	Er sagte: **„Ich habe Hunger."**
Indirekte Wiedergabe	Er sagte, **er habe Hunger.**
Erzählte Wiedergabe	**Er beklagte sich über seinen Hunger.**
Freie indirekte Wiedergabe (Erlebte Rede) (Vermischung Merkmale Erzähler/Figurenrede; meist für Gedankenwiedergabe)	Er war ratlos. **Wo sollte er jetzt nur etwas zu essen herbekommen?**
Redebericht (konjunktivische Wiedergabe ohne Rahmenformel)	Sie wusste Rat. **Es gebe gute Restaurants in der Nähe.**

nicht-direkt

WIEDERGABE VON REDE, GEDANKEN UND GESCHRIEBENEM – EINE TYPOLOGIE

Was gibt es schon?

Was machen wir gerade?

Direkte Wiedergabe	Er sagte: **„Ich habe Hunger."**
Indirekte Wiedergabe	Er sagte, **er habe Hung**
Erzählte Wiedergabe	**Er beklagte sich über seinen**
Freie indirekte Wiedergabe (Erlebte Rede) (Vermischung Merkmale Erzähler/Figurenrede; meist für Gedankenwiedergabe)	Er war ratlos. **Wo sollte er jetzt nur etwas zu essen herbekommen?**
Redebericht (konjunktivische Wiedergabe ohne Rahmenformel)	Sie wusste Rat. **Es gebe gute Restaurants in der Nähe.**

ERKENNUNG DIREKTER WIEDERGABE – ANFÜHRUNGSZEICHEN: DIE OFFENSICHTLICHE LÖSUNG?

Anführungszeichen - Vorteile

- Sehr einfach zu implementierende Strategie
- Sehr erfolgreich – wenn Anführungszeichen konsistent gesetzt sind

Anführungszeichen - Nachteile

- Anführungszeichen ≠ direkte Wiedergabe
- Viele Typen von Anführungszeichen (sprach- und druckabhängig)
- Häufige Fehler bei der Setzung und Kodierung von Anführungszeichen
- Weglassen von Anführungszeichen geschieht auch bewusst (gerade in literarischen Texten)

→ Erkennung über Anführungszeichen erzeugt stark schwankende Ergebnisse (perfekt bis völliges Versagen)

Er sagte: **„Ich habe Hunger."**

ERKENNUNG DIREKTER WIEDERGABE – EINIGE WERKZEUGE

Über Anführungszeichen

Stanford CoreNLP Quote: Erkennt verschiedene Anführungszeichen-Varianten

→ https://stanfordnlp.github.io/CoreNLP/quote.html

GutenTag: Tool zu Taggen von Texten aus dem Gutenberg-Korpus; identifiziert im Text verwendete Anführungszeichen und markiert Text dazwischen

→ http://www.cs.toronto.edu/~jbrooke/gutentag

(beide mit Sprecherzuordnung)

Ohne Anführungszeichen

Tu/Krug/Brunner (2019): **Regelbasierte** Erkennung direkter Wiedergabe ohne Verwendung von Anführungszeichen (Accuracy 0,8-0,85 auf fiktionalen Texten, 0.59 auf nicht-fiktionalen Texten)

Jannidis et al (2018): Erkennung direkter Wiedergabe ohne Verwendung von Anführungszeichen mit **DeepLearning** (Accuracy bis zu 0,9 auf fiktionalen Texten)

DAS REDEWIEDERGABE-PROJEKT

DFG-gefördertes Kooperationsprojekt zwischen IDS Mannheim (A. Brunner, S. Engelberg, N.D.T. Tu) und Universität Würzburg (F. Jannidis, L. Weimer), Laufzeit 2017-2020

Redewiedergabe-Korpus

- deutschsprachiges, historisches Korpus (1840-1920)
- fiktionale und nicht-fiktionale Texte
- detaillierte manuelle Annotation von Wiedergabe von Reden, Gedanken und Geschriebenem in direkter, indirekter, frei-indirekter und erzählter Form

Beta-Release: **github.com/redewiedergabe**

Redewiedergabe-Erkenner

- Erkenner für historische Daten (auch nicht-fiktional)
- Erkenner auch für nicht-direkte Formen

Work in Progress

Veröffentlichung geplant im Frühjahr 2020

WIEDERGABE VON REDE, GEDANKEN UND GESCHRIEBENEM – EINE TYPOLOGIE

Direkte Wiedergabe	Er sagte: **„Ich habe Hunger."**
Indirekte Wiedergabe	Er sagte, **er habe Hung**
Erzählte Wiedergabe	**Er beklagte sich über seinen**
Freie indirekte Wiedergabe (Erlebte Rede) (Vermischung Merkmale Erzähler/Figurenrede; meist für Gedankenwiedergabe)	Er war ratlos. **Wo sollte er jetzt nur etwas zu essen herbekommen?**
Redebericht (konjunktivische Wiedergabe ohne Rahmenformel)	Sie wusste Rat. **Es gebe gute Restaurants in der Nähe.**

Was machen wir gerade?

WORK IN PROGRESS:
FREIE INDIREKTE WIEDERGABE

Er war ratlos. **Wo sollte er jetzt nur etwas zu essen herbekommen?**

- Vermischung von Merkmalen der Erzähler- und Figurenrede
- Typischerweise Gedankenwiedergabe; vor allem in literarischen Texten
- zunehmend verbreitet seit dem Beginn des 20. Jahrhunderts
- Kontextabhängig (Perspektivenverschiebung), keine ‚harten' Indikatoren

Herangehensweise

- Spezialkorpus: Moderne Heftchenromane und Krimis
- DeepLearning auf manuell annotierten Daten (FLAIR library, contextual string embeddings)

FREIE INDIREKTE WIEDERGABE

Auswertung auf einem Korpus mit 22 Ausschnitten aus **Heftchenromanen**
(je ca. 1000 Tokens, ca. 13% der Sätze mit FI)

	F1	Precision	Recall	Accuracy
Regelbasiert (Baseline)	0,37	0,57	0,27	0,88
DL-Modell	**0,45**	**0,78**	**0,32**	**0,9**

Und wie schwer ist es für Menschen? Vergleich zweier unabhängiger Annotatoren:

Menschliche Annotatoren	0,7	0,73	0,67	0,93

FREIE INDIREKTE WIEDERGABE

Auswertung auf dem **Erzähltextkorpus** (Brunner 2015)
Korpus mit 13 historischen Erzähltexten (1787-1913)
ca. 57.000 Tokens, 4,5% der Sätze mit FI, stark konzentriert auf einen Text

	F1	Precision	Recall	Accuracy
RandomForest-Modell (2015)	0,41	0,61	0,31	0,96
DL-Modell	**0,52**	**0,65**	**0,43**	**0,96**

→ Unser Modell liefert vergleichbare Ergebnisse für historische Daten
→ Deutliche Verbesserung zu dem Modell von 2015 erkennbar

FREI-INDIREKTE WIEDERGABE - FEHLERANALYSE

- Wenn Texte keine Anführungszeichen enthalten, wird leicht direkte Wiedergabe als FI erkannt
- Problem ist vor allem, dass Fälle überhaupt nicht gefunden werden (Recall), insbesondere Sätze, die wenig/keine auffallenden Merkmale für FI aufweisen

Und Gräfin Sophie , die diesen Empfang gab, hatte Gwen obendrein geraten, einen großen Bogen um ihren geschätzten Stiefsohn zu machen.	Erzählertext	Korrekt erkannt
Vermutlich war die ältere Dame von ihrem Beschützerinstinkt getrieben.	FI: Gwens Gedanken (Kontext)	Nicht erkannt
Kein Wunder, bei so einem gut aussehenden, charmanten Exemplar!	FI: Gwens Gedanken (mit FI-Merkmal Ausrufezeichen)	Korrekt erkannt

Vielen Dank!
Fragen?

Homepage: redewiedergabe.de

Github: github.com/redewiedergabe