UNIVERSITÄT
ZU KÖLN

Einleitung

HS Anwendungen der Computerlinguistik

Nils Reiter
nils.reiter@uni-koeln.de

17. Oktober 2024

INSTITUT FÜR
DIGITAL HUMANITIES
UNIVERSITÄT ZU KÖLN

Drei Fragen

- Wer sind Sie? (Name, anderes Fach, Lieblingsessen, -hobby oder -videospiel, ...)
- Was erwarten Sie von dieser Veranstaltung?
  Gibt es etwas was sie gerne behandeln/lernen möchten?
- Worüber haben Sie zuletzt gelacht?

Section 1

Organisatorisches

Computerlinguistik im B.A. Informationsverarbeitung

- Modul **Grundlagen der Computerlinguistik** (früher: Computerlinguistische Grundlagen)
  - Computerlinguistische Grundlagen (Seminar, Winter, Hermes)
    - Linguistische Grundlagen, Annotation
  - Sprachverarbeitung (Vorlesung + Übung, Sommer, Reiter und Pagel)
    - Quantitative Eigenschaften von Sprache, Machine Learning
- Modul **Anwendungen der Computerlinguistik** (früher: Angewandte Linguistische Datenverarbeitung)
  - Deep Learning (Übung, Winter, Nester oder Pagel)
    - Deep Learning
  - Anwendungen der Computerlinguistik (Hauptseminar, Winter, Reiter)
    - Experimente in der Computerlinguistik und darüberhinaus; wo kommen Fortschritt und Erkenntnis her?

Aufbaumodul (AM) 1: Anwendungen der Computerlinguistik

früher: Angewandte Linguistische Datenverarbeitung

- Modul
  - 12 Leistungspunkte
  - 5.–6. Studiensemester
  - "Die Modulnote bildet $48\,\%$ der Fachnote."

Aufbaumodul (AM) 1: Anwendungen der Computerlinguistik

früher: Angewandte Linguistische Datenverarbeitung

- Modul
  - 12 Leistungspunkte
  - 5.–6. Studiensemester
  - "Die Modulnote bildet $48\,\%$ der Fachnote."
- Bestandteile
  - Hauptseminar: dieses hier (Do., 16:00–17:30)
  - Übung: Deep Learning (Do., 12:00–13:30)
  - Modulprüfung: Hausarbeit

Aufbaumodul (AM) 1: Anwendungen der Computerlinguistik

früher: Angewandte Linguistische Datenverarbeitung

- Modul
  - 12 Leistungspunkte
  - 5.–6. Studiensemester
  - "Die Modulnote bildet $48\,\%$ der Fachnote."
- Bestandteile
  - Hauptseminar: dieses hier (Do., 16:00–17:30)
  - Übung: Deep Learning (Do., 12:00–13:30)
  - Modulprüfung: Hausarbeit

Lehrveranstaltung	Kontaktzeit	Selbststudium
Hauptseminar	30	60
Übung	30	60
Modulprüfung	–	180

Lernziele

- Umgang mit computerlinguistischer Forschungsliteratur
- Vertiefung vorhandener CL-Kenntnisse
- Verständnis dafür, welche Rolle NLP in den DH spielt oder spielen kann
- Big picture-Überblick, wie man eigene Experimente durchführt

Ablauf

- Material
  - Plan und Übersicht (öffentlich): `https://uni.koeln/TP78H`
  - Ilias (nicht-öffentlich): `https://uni.koeln/FK8CQ`

Ablauf

- Material
  - Plan und Übersicht (öffentlich): `https://uni.koeln/TP78H`
  - Ilias (nicht-öffentlich): `https://uni.koeln/FK8CQ`
- Studienleistung
  - Hausaufgaben (per Ilias abzugeben)
    - Bei Lektüreaufgaben: Drei Fragen zur Lektüre
  - Aktive Teilnahme in Gruppenarbeit

Praktische Experimente

- Extraktion von Paaren aus Begriffen mit zugehörigen Definitionen aus (englischsprachigen) Fließtexten
- Identifikation von Propaganda-Techniken in Überschriften von Nachrichtentexten
- Erkennung von Humor und ‘Offensivität’ in Tweets und Witzen

Modulprüfung

- Thema
  - Findung und Wahl: Ihre Aufgabe
  - Kann, muss aber nicht, etwas mit dem Seminar zu tun haben
  - Mit mir absprechen
- Praktischer Anteil: Offen.
  Beispiele: Experiment zur automatischen Identifikation eines Textphänomens, Annotationsexperiment, quantitativer Vergleich verschiedener Korpora, ...
- Am Ende: Hausarbeit von max. 4 S. Länge
- Brainstorming über Ideen für Modulprüfungsthemen am 16.01.

Section 2

Scientific Method

Texte
- Formal einheitlich

Falsifikation

Strecken nach Objektivität

geeignete Quellen

Quellen

Quellen-kritik

Nachvollziehbarkeit

Reproduzierbarkeit

Theorie vs. Praxis

Transparenz

What is the scientific method?

Ergebnisse hinterfragen

Kritik an gegenüber Ergebnissen

Ggf. unehrliche Methoden

Kontrolle durch Peer Review

Probleme

Contexts of Discovery and Justification

the well-known difference between the thinker's way of finding [a] theorem and his way of presenting it before a public may illustrate the difference in question. I shall introduce the terms context of discovery and context of justification to mark this distinction. *Reichenbach (1938, 6 f.)*

Contexts of Discovery and Justification

> *the well-known difference between the thinker's way of finding [a] theorem and his way of presenting it before a public may illustrate the difference in question. I shall introduce the terms context of discovery and context of justification to mark this distinction.* *Reichenbach (1938, 6 f.)*

- Context of discovery: How a hypothesis/statement is first thought of
  - Psychology, sociology, history, ...
  - No clearly defined process – *Heureka moments*
- Context of justification: How a hypothesis/statement is proven/validated/justified – and communicated
  - Epistemology, science philosophy, ...
  - Rational and logical processes

Contexts of Discovery and Justification

> *the well-known difference between the thinker's way of finding [a] theorem and his way of presenting it before a public may illustrate the difference in question. I shall introduce the terms context of discovery and context of justification to mark this distinction.* *Reichenbach (1938, 6 f.)*

- Context of discovery: How a hypothesis/statement is first thought of
  - Psychology, sociology, history, ...
  - No clearly defined process – *Heureka moments*
- Context of justification: How a hypothesis/statement is proven/validated/justified – and communicated
  - Epistemology, science philosophy, ...
  - Rational and logical processes
- ⚠ Way of discovering and way of showing some insight may differ

Gerstorfer (2020); Reichenbach (1938)

Context of Justification in Our Disciplines

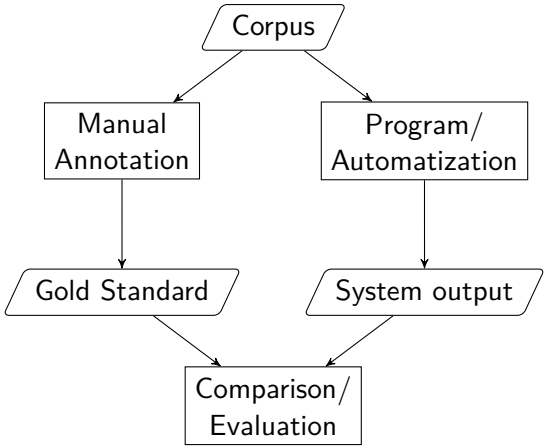
Computational Linguistics: Hypotheses around operationalization ideas

- E.g., "To detect parts of speech, sentence position of a word is important"
- E.g., "We can determine the genre of a play by looking at its end"

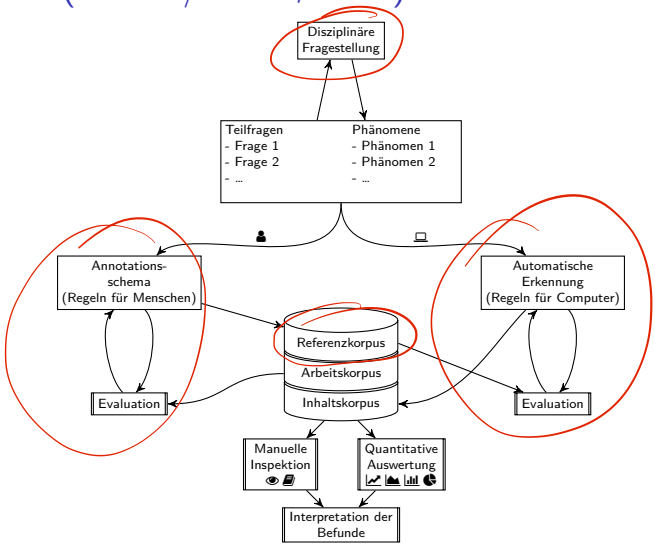
Digital Humanities: Hypotheses around artifact properties

- E.g., "The way people tell stories has changed from telling to showing in the 19th century"
- E.g., "Painters used colors differently after the impressionism"
- E.g., "Introduction of female characters in narratives is more focussed on appearance compared to male characters"

Computational Linguistics



Digital Humanities (Pichler/Reiter, 2020)



Experiment

Different uses of the word

- Contrastive to 'theoretical': "Let's see what happens"
- Contrastive to 'hermeneutic': "Let's look at data systematically"
- Following *scientific* standards
  - Falsification: "Let's see if we can rule out the opposite of what we want to show"
  - Validation: "Let's see if we can show some effect with statistical significance"

Experiment

Ingredients

- Independent variable(s): Manipulated by researchers
- Dependent variable(s): Measuring goal
- Hypothesis: Statement about the relation between independent and dependent variable(s)

Causation

- Assuming we have shown a positive correlation between blackness of a person's hair and their preference for coffee
- Does not mean that black-haired people like coffee *because* they have black hair
- A third, unknown variable, can cause both black hair and coffee preference
- ⚠ Correlation is not the same as causation

Conducting the Experiment

- How many examples do we need?
- Best case: All – but still no proof for a causal relation
  - ⚠ Not realistic

Conducting the Experiment

- How many examples do we need?
- Best case: All – but still no proof for a causal relation
  - ⚠ Not realistic
- Representative sample
  - Smaller, but with similar proportion of relevant properties than the entire population
  - Relevant properties: Difficult to know
  - Approximation through random samples

Experiments in Computational Linguistics

- NLP does not use these terms explicitly
- But underlying concepts motivate many decisions and best practices

Experiments in Computational Linguistics

- NLP does not use these terms explicitly
- But underlying concepts motivate many decisions and best practices
- Hypothesis: This (setting of an) NLP system works better than that (setting)
- 'Setting' includes
  - Features
  - Parameters and hyperparameters
  - Training corpora
  - Supporting resources
  - Annotation schema
  - Data structures

Section 3

Next Week

Next Week

Dong Nguyen/Maria Liakata/Simon DeDeo/Jacob Eisenstein/David Mimno/Rebekah Tromble/Jane Winters (2020). "How We Do Things With Words: Analyzing Text as Social and Cultural Data". In: *Frontiers in Artificial Intelligence* 3. ISSN: 2624-8212. DOI: 10.3389/frai.2020.00062. URL: https://www.frontiersin.org/articles/10.3389/frai.2020.00062

Hausaufgabe 1 (bis 30.10., 23:55)

Nguyen et al. (2020) lesen. Drei Fragen in Ilias einreichen.

References I

- Gerstorfer, Dominik (2020). "Entdecken und Rechtfertigen in den Digital Humanities". In: *Interdisziplinäre(s) Arbeiten in der CRETA-Werkstatt*. Ed. by Nils Reiter/Axel Pichler/Jonas Kuhn. Berlin, Boston: De Gruyter, pp. 107–124. ISBN: 978-3-11-069397-3. DOI: doi:10.1515/9783110693973-005. URL: https://doi.org/10.1515/9783110693973-005 (visited on 10/09/2024).
- Nguyen, Dong/Maria Liakata/Simon DeDeo/Jacob Eisenstein/David Mimno/Rebekah Tromble/Jane Winters (2020). "How We Do Things With Words: Analyzing Text as Social and Cultural Data". In: *Frontiers in Artificial Intelligence* 3. ISSN: 2624-8212. DOI: 10.3389/frai.2020.00062. URL: https://www.frontiersin.org/articles/10.3389/frai.2020.00062.
- Pichler, Axel/Nils Reiter (2020). "Reflektierte Textanalyse". In: *Reflektierte Algorithmische Textanalyse. Interdisziplinäre(s) Arbeiten in der CRETA-Werkstatt*. Ed. by Nils Reiter/Axel Pichler/Jonas Kuhn. Berlin: De Gruyter, pp. 43–60. DOI: 10.1515/9783110693973-003.

References II

 Reichenbach, Hans (1938). *Experience and Prediction. An Analysis of the Foundations and theStructure of Knowledge*. Chicago: Chicago University Press.