UNIVERSITÄT
ZU KÖLN

Einleitung

HS Anwendungen der Computerlinguistik

Nils Reiter
nils.reiter@uni-koeln.de

12. Oktober 2023

INSTITUT FÜR
DIGITAL HUMANITIES
UNIVERSITÄT ZU KÖLN

Drei Fragen

- Wer sind Sie? (Name, 2. Fach, Lieblingswitz, -hobby, -essen oder -videospiel, ...)
- Was erwarten Sie von dieser Veranstaltung?
  Gibt es etwas was sie gerne behandeln/lernen möchten?
- Wie geht's Ihnen?

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)
    - Quantitative Eigenschaften von Sprache, Machine Learning
- Modul **Anwendungen der Computerlinguistik** (früher: Angewandte Linguistische Datenverarbeitung)
  - Deep Learning (Übung, Winter, Nester)
    - 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

- Lesen und verstehen computerlinguistischer Forschungsliteratur
- Vertiefung vorhandener CL-Kenntnisse
- Verständnis dafür, welche Rolle NLP in den DH spielt
- Planung und Durchführung eigener Experimente

Ablauf

- Material
  - Plan und Übersicht (öffentlich): `https://uni.koeln/76ZZC`
  - Ilias (nicht-öffentlich): `https://uni.koeln/SJTRF`

Ablauf

- Material
  - Plan und Übersicht (öffentlich): `https://uni.koeln/76ZZC`
  - Ilias (nicht-öffentlich): `https://uni.koeln/SJTRF`
- Studienleistung
  - Hausaufgaben (per Ilias abzugeben)
    - Bei Lektüreaufgaben: Drei Fragen zur Lektüre
  - Aktive Teilnahme an eigenem Experiment

Praktische Experimente

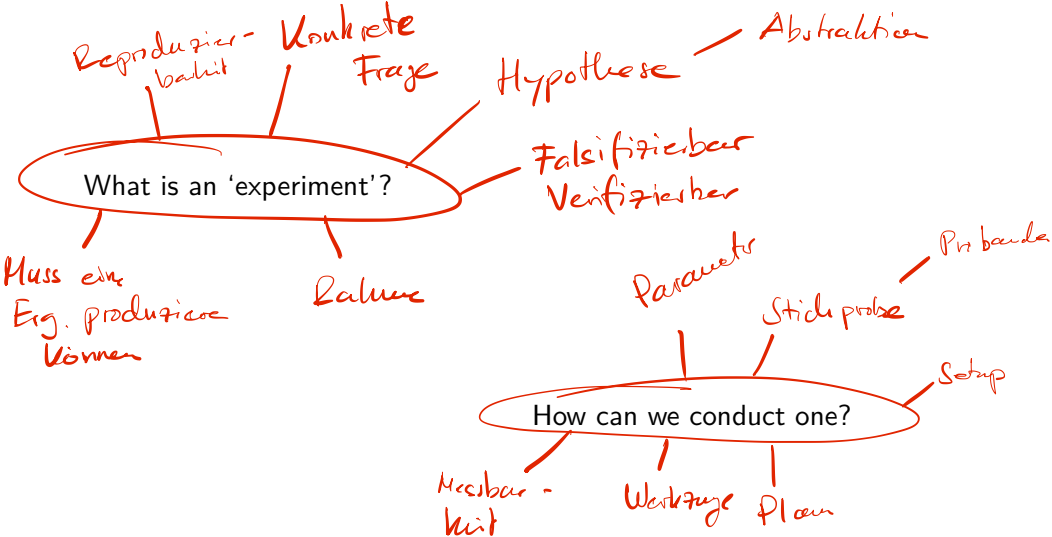
- Extraktion von Paaren aus Begriffen mit zugehörigen Definitionen aus (englischsprachigen) Fließtexten
- Identifikation von Propaganda-Techniken in Überschriften von Nachrichtentexten
- Erkennung von Humor und Offensivität in Tweets und Witzen

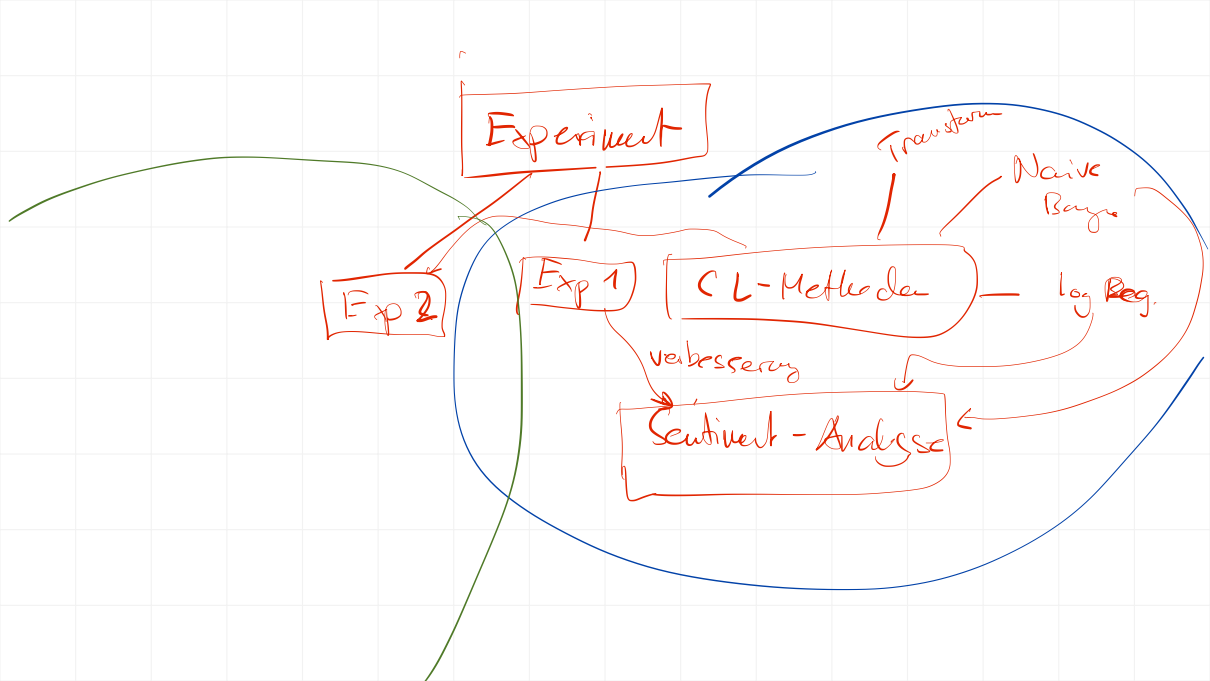
Modulprüfung

- 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 18.01.

Section 2

Experiments


Reproduzierbarkeit
Konkrete Frage
Hypothese
Abstraktion
What is an ‘experiment’?
Falsifizierbar
Verifizierbar
Muss ein Ereignis produzieren können
Kausal
Parameter
Probanden
Stichprobe
How can we conduct one?
Setup
Messbarkeit
Werkzeuge
Plan


Experiment
Exp 2
Exp 1
CL-Methoden
Transform
Naive Bayes
log Reg.
Verbesserung
Sentiment-Analyse

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: "Let's see if we can rule out the opposite of what we want to show"

Experiment

Ingredients

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

Experiment

Example (Goal: People with black hair like coffee)

Experiment

Example (Goal: People with black hair like coffee)

- Hypothesis: There is a positive correlation between blackness of a person's hair and their preference for coffee
  - Independent variable: Hair blackness (in: percent)
  - Dependent variable: How much they like coffee (in: number of pots per week)

Experiment

Example (Goal: People with black hair like coffee)

- Hypothesis: There is a positive correlation between blackness of a person's hair and their preference for coffee
  - Independent variable: Hair blackness (in: percent)
  - Dependent variable: How much they like coffee (in: number of pots per week)
- This is not without alternatives!
- Hypothesis: If a person has black color, they like coffee more than if not
  - Independent variable: Hair color (as a nominal value, e.g. 1=black, 0=other)
  - Dependent variable: How much they like coffee (in: number of pots per week)

Conducting the Experiment

Options for Measuring/Controlling/Manipulating Independent Variable

Conducting the Experiment

Options for Measuring/Controlling/Manipulating Independent Variable

- Take 1 person, dye their hair, count coffee over many weeks
- Take n persons with various hair colors, count coffee for one week
- Take n persons with various natural hair colors, count coffee for one week
- Genetically engineer n babies, such that they grow various hair colors, wait until they're grown up, count coffee for one week
- ...

Conducting the Experiment

Options for Measuring the Dependent Variable

Conducting the Experiment

Options for Measuring the Dependent Variable

- ... count coffee pots over many weeks
- ... count liters of coffee over many weeks
- ... count percentage of caffeine in blood/urine over many weeks
- ...

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

n

- How many people do we need?
- Best case: All – but still no proof for a causal relation
  - ⚠ Not realistic
  - Dead black haired people cannot be observed anymore, more black haired people will be born

Conducting the Experiment

n

- How many people do we need?
- Best case: All – but still no proof for a causal relation
  - ⚠ Not realistic
  - Dead black haired people cannot be observed anymore, more black haired people will be born
- Representative sample
  - Smaller, but with similar proportion of relevant properties than the entire population
  - Relevant properties: Difficult to know
  - Approximation through random samples

Section 3

Next Week

Next Week

Klaus Krippendorff (2019). *Content Analysis: An Introduction to its Methodology*. 4th. Sage, Kapitel 2

Hausaufgabe 1 (bis 18.10., 23:55)

Krippendorff (2019, 2.4) lesen. Drei Fragen in Ilias einreichen.

References I

 Krippendorff, Klaus (2019). *Content Analysis: An Introduction to its Methodology*. 4th. Sage.