

Einleitung

HS Experimentelles Arbeiten in der Sprachverarbeitung

Nils Reiter

`nils.reiter@uni-koeln.de`

13. Oktober 2022

Was ist ein Experiment? Was ist Experiment in der Computerlinguistik?

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
 - ▶ Experimentelles Arbeiten in der Sprachverarbeitung (Hauptseminar, Winter, Reiter)
 - ▶ Experimente in der CL; 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 NLP-technischer Forschungsliteratur
- ▶ Vertiefung vorhandener NLP-Kenntnisse
- ▶ Planung und Durchführung eigener Experimente

Ablauf

▶ Material

- ▶ Plan und Übersicht (öffentlich): <https://uni.koeln/BQG2T>
- ▶ Ilias (nicht-öffentlich): <https://uni.koeln/AVYXF>

Ablauf

- ▶ Material
 - ▶ Plan und Übersicht (öffentlich): <https://uni.koeln/BQG2T>
 - ▶ Ilias (nicht-öffentlich): <https://uni.koeln/AVYXF>
- ▶ Studienleistung
 - ▶ Hausaufgaben (per Ilias abzugeben)
 - ▶ Bei Lektüreaufgaben: Drei Fragen zur Lektüre
 - ▶ Aktive Teilnahme

Praktische Experimente

► Detecting and Rating Humor and Offense

```
1 35,Learn from the scars of others,0,,0.05  
2 53,Why does Kim Jong Un love books. Because he's the glorious Reader,1,1.68,0,0.8
```

Praktische Experimente

► Detecting and Rating Humor and Offense

```
1 35,Learn from the scars of others,0,,0.05
2 53,Why does Kim Jong Un love books. Because he's the glorious Reader,1,1.68,0,0.8
```

► Patronizing and Condescending Language Detection

```
1 982_@@24299294__refugee_in__Bangla hosts now critics of Rohingya refugees__0
2 7341____@14684361__refugee_in__Egypt billionaire offers to buy Mediterranean island f
```

Praktische Experimente

▶ Detecting and Rating Humor and Offense

```
1 35,Learn from the scars of others,0,,0.05
2 53,Why does Kim Jong Un love books. Because he's the glorious Reader,1,1.68,0,0.8
```

▶ Patronizing and Condescending Language Detection

```
1 982_@@24299294__refugee_in__Bangla hosts now critics of Rohingya refugees__0
2 7341____@14684361__refugee_in__Egypt billionaire offers to buy Mediterranean island f
```

▶ Named Entity Recognition and Classification on Historical Newspapers

1 Neben_0_____
2 der_0_____
3 Branntweinteuerfrage_0_____
4 wird_0_____
5 den_0_____
6 Reichstag_B-ORG_____Q160208_____
7 auch_0_____
8 noch_0_____
9 [...]
10 ein_0_____
11 Geetz_0_____
12 über_0_____
13 die_0_____
14 Verorgung_0_____
15 der_0_____
16 Hinterbliebenen_0_____
17 des_0_____
18 Reichsheeres_B-ORG_____Q313422_____
19 und_0_____
20 der_0_____
21 Kaierlichen_B-ORG_____Q156649_____
22 Marine_I-ORG_____Q156649_____
23 beschäftigen_0_____
24 ._0_____EndOfSentence

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
- ▶ 'Letzte' Übung vor der Bachelor-Arbeit
- ▶ Brainstorming über Ideen für Modulprüfungsthemen am 22.12.

Section 2

Experiments

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

Experiments in Natural Language Processing

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

Experiments in Natural Language Processing

- ▶ 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

Rrubaa Panchendrarajan/Nazick Ahamed/Brunthavan Murugaiah/Prakhash Sivakumar/Surangika Ranathunga/Akila Pemasiri (2016). “Implicit Aspect Detection in Restaurant Reviews using Cooccurrence of Words”. In: *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. San Diego, California: Association for Computational Linguistics, pp. 128–136. DOI: 10.18653/v1/W16-0421. URL: <https://www.aclweb.org/anthology/W16-0421>

Daniel Preoțiuc-Pietro/Mihaela Gaman/Nikolaos Aletras (2019). “Automatically Identifying Complaints in Social Media”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 5008–5019. DOI: 10.18653/v1/P19-1495. URL: <https://www.aclweb.org/anthology/P19-1495.pdf>