Experiments in Computational Linguistics

HS Sprachtechnologie für eine bessere Welt (Winter term 2022/23)

Nils Reiter,
nils.reiter@uni-koeln.de

January 17, 2023

INSTITUT FÜR
DIGITAL HUMANITIES
UNIVERSITÄT ZU KÖLN

Section 1

Introduction

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)

Experiments in Computational Linguistics

- Independent variable(s): Aspects of the machine learning system/algorithm
  - E.g., the features used or the way data is preprocessed
- Dependent variable(s): Performance on test data
  - E.g., f-score
- Hypothesis: Using feature set X increases the performance
  - E.g., context information is useful to assign part of speech tags

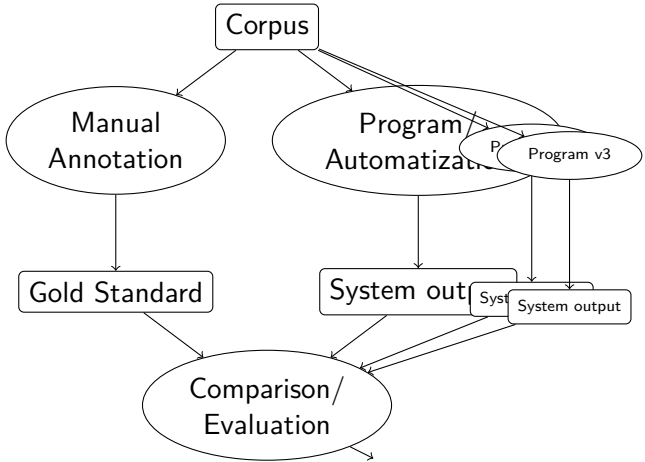
Experiments in Natural Language Processing

- 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

Experiments



	P	R	F
Programm	5.7	9.2	...
v2	9.9	16.7	...
v3	15.3	21.8	...

Experiments

- Reproducibility
- Hypotheses about the operationalisation of language/text phenomena

Example

- Position within a sentence is indicative for the part of speech
- Meaning of a word depends on its context
- The protagonist of a play is the character who talks the most

Checkliste zu NLP-Experimenten zur Klassifikation

Stand: 20. Dezember 2022

Hinweise

- Wenn Ihr Task kein Klassifikationstask ist, dann ist dieser Fragebogen nicht für Sie.
  - Keine Klassifikationstasks sind z.B. Generierung oder Übersetzung.
- Markieren Sie alle Punkte die Sie planen umzusetzen.
- Machen Sie einen Termin in der Sprechstunde, wenn Ihnen Punkte unklar sind.
- Der Fragebogen ist keine Prüfung, sondern dient als Hilfestellung, bei der Experimentplanung an alles zu denken, auf Ideen zu kommen und ggf. die richtigen Fragen zu stellen. Diese klären wir dann idealerweise im Gespräch.
- Der Fragebogen repräsentiert eine Planung. Gewisse Abweichungen von der Planung sind normal und zu erwarten.
- Bei Fragen schreiben Sie gerne eine E-Mail an `nils.reiter@uni-koeln.de` oder melden Sie sich gerne zu einer Sprechstunde an. Wenn Sie rein technische Fragen zur Infrastruktur haben, schreiben Sie bitte an `apinfo-admin@uni-koeln.de`.

Der Task

1. Die Aufgabe heißt: _____
2. Es handelt sich um ☐ Textklassifikation, ☐ Sequence Labeling, oder ☐ Sonstiges: _____
3. Die zu klassifizierenden Instanzen sind: _____
4. Es gibt _____ Kategorien/Klassen.
5. Einer Instanz kann ☐ genau eine oder ☐ mehrere Klassen zugewiesen werden.

Die Daten

1. Annotierte Daten ☐ liegen bereits vor oder ☐ müssen noch erstellt werden.
2. In den Daten sind _____ Instanzen (von o.g. Typ) annotiert.
3. Die Klassen sind
   - ☐ gleichverteilt (d.h. jede Klasse ist ungefähr gleich häufig)
   - ☐ unterschiedlich verteilt, und zwar: _____

Die Annotationen

Nur relevant, wenn neue Daten annotiert werden sollen. (Frage Task.1)

1. Annotationsrichtlinien
   - ☐ Ich verwende die folgenden, bereits existierenden Annotationsrichtlinien: _____
     - Mit diesen wurde ein Inter-Annotator-Agreement von ___ erzielt (Metrik: _____).
   - ☐ Ich schreibe neue Annotationsrichtlinien.

2. Annotator:innen
   - ☐ Ich annotiere selbst.
   - ☐ Ich rekrutiere Annotator:innen aus meinem Freundes-/Bekanntenkreis.
   - ☐ Ich sammle Annotationen über eine Umfrage, z.B. mittels LimeSurvey.
   - ☐ Ich sammle Annotationen über crowd sourcing.
3. Annotationsworkflow
   - ☐ Annotator:innen treffen eine Annotationsentscheidung auf der Basis eines Kontextes von ___☐ Wörtern, ☐ Sätzen, ☐ Zeilen, ☐ Absätzen, ☐ _____ oder ☐ sie verwenden den gesamten Text als Kontext.
   - ☐ Sie können dabei außerdem die folgenden Wissensquellen verwenden: ☐ Wikipedia, ☐ Lexika, ☐ Wörterbücher
4. Anforderungen an Annotationssoftware
   - ☐ Annotator:innen müssen Spannen selbst markieren können.
   - ☐ Annotator:innen müssen neue Kategorien oder Labels ergänzen können.

Die Baseline

- ☐ Weil die Klassen ungleich verteilt sind, bietet sich eine majority baseline an. Diese erzielt eine Accuracy von _____ %.
- ☐ Weil die Klassen gleich verteilt sind, bietet sich eine random baseline an. Diese erzielt eine Accuracy von _____ %.
- ☐ Eine weitere mögliche Baseline ist: _____ Diese erzielt eine Accuracy von _____ %.
- ☐ Eine weitere mögliche Baseline ist: _____ Diese erzielt eine Accuracy von _____ %.

Das Experiment

1. Ich möchte das folgende oder die folgenden Verfahren verwenden:
   - ☐ Entscheidungsbaum / Decision Tree (DT)
   - ☐ Naive Bayes
   - ☐ Support Vector Machines (SVM)
   - ☐ Logistic Regression
   - ☐ Neural Networks (NN)
     - ☐ Feed-Forward Neural Networks
     - ☐ Convolutional Neural Networks
     - ☐ Recurrent Neural Networks
     - ☐ Transformer-Architektur (BERT & co.)
   - ☐ Sonstige: _____
2. Ich möchte die folgenden Features verwenden
   - ☐ Metadaten: _____
   - ☐ Inhaltsdaten, z.B. aus Texten:
     - ☐ Worthäufigkeiten (von allen Wörtern), auch bekannt als bag of words
     - ☐ Häufigkeiten von Wörtern aus folgenden Wortlisten: _____
     - ☐ Embeddings (z.B. Word Embeddings)
     - ☐ Sequenzielle Information (d.h. Klassifikationsergebnisse für Elemente davor oder danach)
     - ☐ N-Gram-Häufigkeiten, mit $N \leq$ _____
     - ☐ Thematische Informationen aus einem Topic Model (z.B. Latent Dirichlet Allocation, LDA)
3. Meine Features haben die folgenden Datentypen:
   - ☐ Numerisch: _____ (Anzahl Features) ☐ Kategorial: _____ (Anzahl Features)
4. Testdaten
   - ☐ Ich teile meinen o.g. Datensatz selbst in Trainigs- und Testdaten auf. _____% der Instanzen werden als Trainingsdaten verwendet.
   - ☐ Ich verwende N-fold cross validation, mit $N =$ _____.
   - ☐ Trainings- und Testdaten sind bereits aufgeteilt, z.B. weil es Daten aus einem shared task sind.
5. Ich variiere und vergleiche
   - ☐ die Größe des Trainingsdatensatzes (z.B. 100, 1000, 10 000 Instanzen für den Trainingsdatensatz)
   - ☐ die Menge an oder Art von Features die verwendet werden (z.B. inhaltliche vs. sprachliche Features)
   - ☐ das Verfahren als solches oder Parameter davon (z.B. NN vs. SVM)
   - ☐ die Vorverarbeitung (z.B. Groß- und Kleinschreibung)
6. Meine Hypothese ist: _____

Die Auswertung und Evaluation

1. Ich verwende die Evaluationsmetrik(en)
   ☐ Accuracy ☐ Precision ☐ Recall ☐ F-Measure ☐ Area under curve (AUC)
   ☐ Sonstige: _____
   ☐ Meine Testdaten sind stark unbalanciert (Frage Daten.3), daher verwende ich die Metriken in der Mikro- und Makro-Average-Variante.
2. Für meine Fehleranalyse inspiziere ich _____ Instanzen manuell.

Die praktische Umsetzung

1. Ich verwende die Programmiersprache
   - ☐ Python
   - ☐ Java
   - ☐ R
   - ☐ _____
2. Hardware-Ausstattung und Vorkenntnisse
   - ☐ Ich verfüge über einen Computer
     - ☐ der auch mal über Nacht durchlaufen kann, wenn eine Berechnung etwas länger dauert.
     - ☐ der eine GPU mit CUDA-Unterstützung hat oder ein Mac mit M1/M2-Prozessor ist.
     - ☐ der ausreichend freien Plattenspeicher hat.
   - ☐ Ich möchte Berechnungen auf einem Server der Universität laufen lassen.
     - ☐ Ich kann mich per SSH auf einem Server einloggen.
     - ☐ Ich weiß wie ich auf einer Kommandozeile ein Programm laufen lasse.

Section 2

Evaluation and Quality Assurance in NLP

Motivation

- Impact of NLP methods/applications on the world
- How to systematically consider that in research?

Established Evaluation in NLP

Intrinsic

- Comparison with a gold standard
- Precision/recall/f-score/accuracy/...
- Goal: Replicate the gold standard

Established Evaluation in NLP

Intrinsic

- Comparison with a gold standard
- Precision/recall/f-score/accuracy/...
- Goal: Replicate the gold standard

Extrinsic

- Integrate your method into a larger system
- Evaluate this larger system
  - Against a gold standard
  - Post-hoc with human evaluators

Established Evaluation in NLP

Intrinsic

- Comparison with a gold standard
- Precision/recall/f-score/accuracy/...
- Goal: Replicate the gold standard

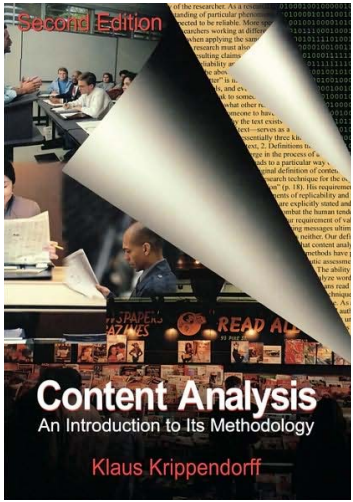
Extrinsic

- Integrate your method into a larger system
- Evaluate this larger system
  - Against a gold standard
  - Post-hoc with human evaluators

Further considerations, often not discussed in NLP

- Is this method *significantly* better than an alternative?
- Comparison against one or more baseline(s)
- Cost-benefit-analysis (e.g., learning curve)
- Are statements derived from this „valid“?

Quality Assurance in Social Sciences



Klaus Krippendorff (2004). *Content Analysis: An Introduction to its Methodology*. 2nd. Los Angeles, California, USA: Sage

Two aspects of quality:

- Reliability
- Validity

Assumptions

- Data: Measurement results
- No strict separation between manual annotation and automatic prediction
- Content analysis: Analyzing large volumes of text (or data) with an interest in (some aspect of) the content
  - I.e.: Not purely methodological interest
  - Not an interest in the words, but the meaning of the words

Reliability

[...] content analysts must be confident that their data (a) have been generated with all conceivable precautions in place against known pollutants, distortions, and biases, intentional or accidental, and (b) mean the same thing for everyone who uses them. Reliability grounds this confidence empirically.
[...] a research procedure is reliable when it responds to the same phenomena in the same way regardless of the circumstances of its implementation. (Krippendorff, 2004, 211)

Reliability Designs

- Three types of reliability (and three ways to test it)
- Generate ‚reliability data' – in addition to the data whose reliability is in question

Reliability	Designs	Causes of Disagreements	Strength
Stability	test-retest	intraobserver inconsistencies	weakest
Reproducibility	test-test	intra+inter	medium
Accuracy	test-standard	intra+inter+deviations from standard	strongest

Tabelle: Types of Reliability (Krippendorff, 2004, 215)

Reliability Designs

- Three types of reliability (and three ways to test it)
- Generate ‚reliability data' – in addition to the data whose reliability is in question

Reliability	Designs	Causes of Disagreements	Strength
Stability	test-retest	intraobserver inconsistencies	weakest
Reproducibility	test-test	intra+inter	medium
Accuracy	test-standard	intra+inter+deviations from standard	strongest

Tabelle: Types of Reliability (Krippendorff, 2004, 215)

- NLP and reliability: ✔

Validity

> *Validation provides compelling reasons for taking the results of scientific research seriously. It can serve as the ground for developing theories and the basis of successful interventions.*
> *[...]*
> *A measuring instrument is considered valid if it measures what its user claims it measures.* *(Krippendorff, 2004, 313)*

- Validity is not a quantitative test we can apply
- Validity is the result of a process of validation, an argumentation
  - A researcher makes arguments for the validity
  - An audience may be skeptical about some or all of them
- Krippendorff differentiates several sources of validity or ways of validating

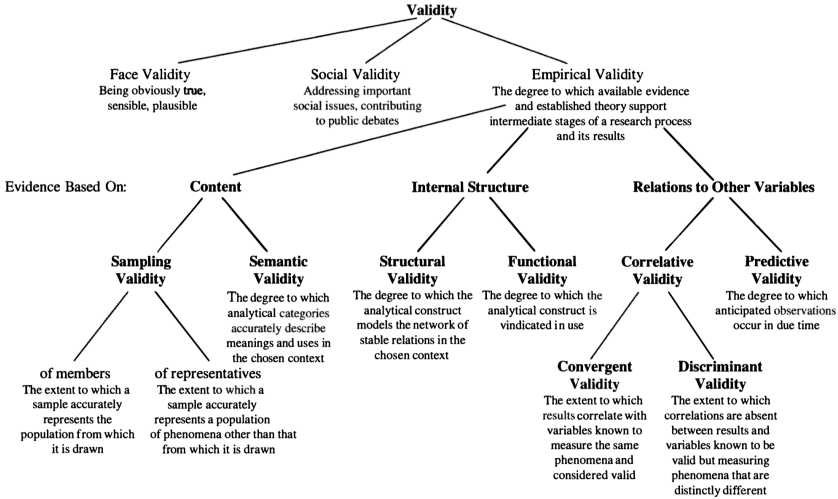
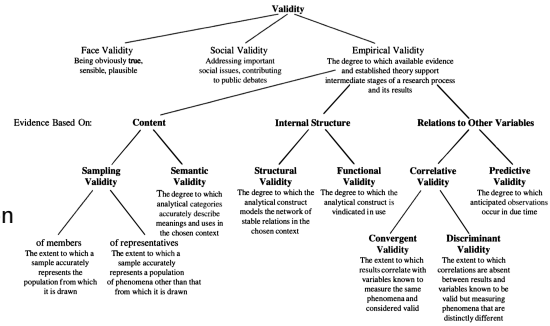


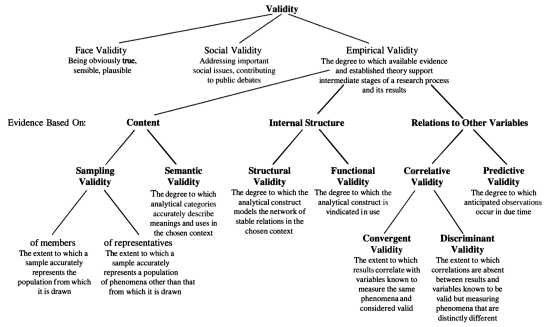
Abbildung: A Typology of Validation Efforts (Krippendorff, 2004, 319)

Social and Face Validity



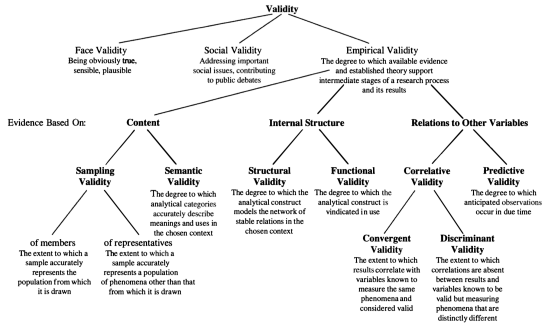
- Face Validity
  - Obvious or common truth
  - „an individual's judgment with the assumption that everyone else would agree with it“ (Krippendorff, 2004, 314)
- Social Validity
  - Research is valuable for the society
  - „[S]ocial validity of content analysis studies is often debated, negotiated, and a matter of public concern“ (Krippendorff, 2004, 314)

Evidence Based on Content



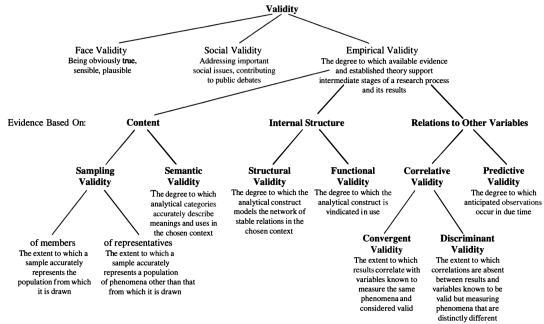
- Sampling validity
  - Does the sample we investigate accurately represent the population we want to say something about?
  - Ideally: Use sampling methods that ensure representativeness
  - Reality: Not always controllable
- Semantic validity
  - To which extent do the categories we investigate correspond to the meanings of the text?
  - Do our categories represent the meaning of interest at all?

Evidence Based on Internal Structure



- Structural Validity
  - Does the structure of the content analysis (i.e., the different components together) accurately represent the domain?
- Functional Validity
  - Is the analytical construct established and useful?
  - „[O]ne must demonstrate that its analytical constructs [...] are useful over time and in many empirical situations.“(Krippendorff, 2004, 332)

Evidence Based on Relations to Other Variables



- Correlative Validity
  - Validity ‚travels' along high correlations
  - If two variables are highly correlated, and one's validity is established, the other is considered valid as well
- Predictive Validity
  - To which degree do content analysis methods accurately predict events, identify properties etc.?
  - How well can we actually predict previously unknown things?

Reliability and Validity

> *whereas reliability provides assurances that particular research results can be duplicated [...] validity provides assurances that the claims emerging from the research are borne out in fact.* *(Krippendorff, 2004, 212)*

- Unreliability limits the chance of validity

Reliability and Validity

> *whereas reliability provides assurances that particular research results can be duplicated [...] validity provides assurances that the claims emerging from the research are borne out in fact.* *(Krippendorff, 2004, 212)*

- Unreliability limits the chance of validity
- Reliability does not guarantee validity

Reliability and Validity

> *whereas reliability provides assurances that particular research results can be duplicated [...] validity provides assurances that the claims emerging from the research are borne out in fact.* *(Krippendorff, 2004, 212)*

- Unreliability limits the chance of validity
- Reliability does not guarantee validity
- In the pursuit of high reliability, validity tends to get lost
  Example: Merritt (1966)
  - Study about rising national consciousness in 13 American colonies
  - No operationalization of „national sentiment“
  - Instead: Enumerate and count American place-names

Group Exercise

- Let's make this more concrete
- What are examples related to NLP with low and high validity?
- Find examples for as many validity types as possible

Validity Types

Sampling, semantic, structural, functional, correlative, predictive