Verlässlichkeit der Annotation

Proseminar Linguistische Annotationen

Benjamin Weitz

SS 10

17. Juni 2010

Inhaltsverzeichnis

1. Generalizability Theory
   - Überblick
   - Die Theorie
   - Anwendungsbeispiel
   - Planung einer G-Studie
   - G-Theorie vs. andere Agreement-Maße
   - Zusammenfassung
2. Einfluss von Annotationsqualität auf maschinelles Lernen
   - Einleitung
   - Die Methode
   - Auswertung
   - Fazit
   - Zusammenfassung

Wofür wird Generalizability Theory benutzt?

- Manuelle Annotation: Hauptinformationsquelle in vielen computerlinguistischen Projekten
- *aber:* oft nicht zuverlässig genug
- *deshalb:* Methode zum Bestimmen der Faktoren, die die Annotationsqualität beeinflussen
- **Generalizability Theory (G-Theory):** systematische Bestimmung von Ursachen für Annotator-Disagreement

Basisinformationen

- Zuverlässigkeit in G-Theory: Größe der Varianz in Annotationen
  - kleinere Gesamtvarianz $\rightarrow$ größere Zuverlässigkeit
- Zuverlässigkeit beeinflusst von verschiedenen **Facetten** (Faktoren)
  - sind (einzeln und in Interaktion) verantwortlich für die Varianz
  - *z.B.:* persönliches Verhalten der Annotatoren, Änderungen in den Annotationstools, Zeitdruck, Änderung des Lohns, Änderungen im Annotationsschema
  - können jeweils das Verhalten des Annotators und somit das Disagreement zwischen den Annotatoren beeinflussen
  - Jede der Facetten hat einen bestimmten Einfluss auf die Zuverlässigkeit der Annotation
- **Aufgabe der G-Theory:** Einfluss der einzelnen Facetten isolieren und die Größe dieses Einflusses bestimmen

Designs

- zwei mögliche Designs: **crossed** und **nested**
  - **crossed:** Messungen für jede mögliche Kombination von Facettenwerten
    - z.B. 2 Facetten: items (z.B. Phrase, Phon,...) und Annotatoren
    - Jedes item wird von jedem Annotator annotiert
    - *also:* Jeder Wert der Facette item wird in Kombination mit jedem Wert der Facette Annotator gemessen
  - **nested:** Messungen nur für eine Untermenge der möglichen Kombinationen
    - z.B. wenn nur manche Annotatoren bestimmte Objekte annotiert haben (wie bei unserer Annotation: SALSA vs. GermaNet)
  - crossed Designs brauchen mehr Messungen, liefern aber auch mehr Informationen
    $\rightarrow$ Sollten bevorzugt werden, um ein Bild aller möglichen Einflüsse zu erhalten

Berechnung der Komponenten der Varianz

- Bei crossed Designs:
  - Varianz = Summe der Varianz der einzelnen Facetten und ihrer Interaktion
  - 3 Facetten a, b, c:
    $\sigma^2(X_{abc}) = \sigma^2_a + \sigma^2_b + \sigma^2_c + \sigma^2_{ab} + \sigma^2_{ac} + \sigma^2_{bc} + \sigma^2_{abc,e}$
  - e: Fehlervarianz
- Bei nested Designs:
  - einige Facetten können nicht einzeln behandelt werden, weil sie von anderen Facetten abhängen
  - 3 Facetten a, b, c,
    c wird mit verschiedenen Werten von b kombiniert
    $\sigma^2(X_{abc}) = \sigma^2_a + \sigma^2_b + \sigma^2_{ab} + \sigma^2_{c,cb} + \sigma^2_{ac,abc,e}$
  - Varianz für c kann nicht einzeln bestimmt werden
- 7 vs. 5 Summanden $\rightarrow$ unterstreicht, dass crossed Designs mehr Informationen liefern

Interpretation der Komponenten der Varianz

- Gesamtvarianz: Summe der einzelnen Varianzkomponenten $\rightarrow$ Gesamtvarianz = 100%
- relative Größe einer Komponente im Bezug auf die Gesamtvarianz: Maß für den Beitrag dieser Komponente zur Verlässlichkeit
  - z.B. Facette die 60% der Gesamtvarianz ausmacht: Hauptquelle für Unzuverlässigkeit im Gegensatz zu einer Facette die nur 5% der Varianz erklärt
  - große Varianz bei Annotator-Facette: Varianz durch systematische Unterschiede im Annotationsverhalten der einzelnen Annotatoren
  - große Varianz bei Schema-Facette: Varianz durch systematische Unterschiede in den Kategorien
  - große Varianz bei Coder-Schema-Interaktion: Varianz durch systematische Unterschiede, wie die Annotatoren die Kategorien benutzen
  - große Varianz bei item-Facette: einige Materialien sind schwieriger zu annotieren als andere
- Nach Identifikation: Maßnahmen zur Behebung treffen

Re-Analyse von Shriberg and Lof (1991)

- Untersuchten die Genauigkeit von breiter und enger phonetischer Transkription
- Vier Facetten:
  - Annotationsschema (Typ von Konsonant)
  - Granularität (breit vs. eng)
  - Material (fortlaufende Sprache vs. Artikulationstest)
  - Annotationsteam

Re-Analyse von Shriberg and Lof (1991)

Effect	df	Variance components estimates		Percentage of total variance
Consonant (C)	23	234.86877		25.70
Granularity (G)	1	312.80278		34.23
Team (T)	3	3.70906		0.41
Material (M)	1	0.0	[−3.08672]*	–
CG	23	99.18526		10.85
CT	69	0.0	[−8.25984]*	–
CM	23	45.80498		5.01
GT	3	0.0	[−1.12263]*	–
GM	1	0.0	[−6.05138]*	–
TM	3	0.0	[−1.74740]*	–
CGT	69	3.84207		0.42
CGM	23	111.61108		12.12
CTM	69	57.64646		6.31
GTM	3	6.04318		0.66
CGTM,*e*	36	38.23065		4.18
		913.74429		99.99

- Normalerweise wird angenommen: Varianz hauptsächlich durch Annotatorverhalten
- hier: team-Facette nur für kleinen Anteil der Varianz verantwortlich
  - auch in Interaktion mit anderen Faktoren
- *also:* Keine großen Unterschiede in Annotationsqualität der Teams

Re-Analyse von Shriberg and Lof (1991)

Effect	df	Variance components estimates		Percentage of total variance
Consonant (C)	23	234.86877		25.70
Granularity (G)	1	312.80278		34.23
Team (T)	3	3.70906		0.41
Material (M)	1	0.0	[−3.08672]*	–
CG	23	99.18526		10.85
CT	69	0.0	[−8.25984]*	–
CM	23	45.80498		5.01
GT	3	0.0	[−1.12263]*	–
GM	1	0.0	[−6.05138]*	–
TM	3	0.0	[−1.74740]*	–
CGT	69	3.84207		0.42
CGM	23	111.61108		12.12
CTM	69	57.64646		6.31
GTM	3	6.04318		0.66
CGTM,*e*	36	38.23065		4.18
		913.74429		99.99

- Hauptverantwortliche Faktoren für die Varianz: Granularität und Konsonantentyp
- Material: kein großer Einfluss
  - *aber:* relevant in Interaktion mit Konsonant und Granularität
- **Unverlässlichkeit hauptsächlich wegen Granularität und Schema, nicht durch Eigenheiten der Annotatoren**

Re-Analyse von Shriberg and Lof (1991)

- Welche Werte der Facetten sind für Disagreement verantwortlich?
- Kann durch Eingabedaten (Agreement-Daten) festgestellt werden
  - Granularität: geringeres Agreement bei enger Transkription (64,15%) als bei breiter (89,46%)
  - Konsonanten: kritische Phoneme (z.B. /ð/, /ʃ/) vs. unkritische (z.B. /j/, /b/)
  - CG-Interaktion: Disagreement bezüglich Konsonanten bei engen Transkriptionen höher als bei breiten
  - CGM-Interaktion: Artikulationstests höheres Disagreement
- *also:* Gute Auswahl von Annotatoren und Training
- *aber:* Für hohe Verlässlichkeit:
  - breite Transkriptionen bevorzugen
  - Training für problematische Konsonanten und mit Artikulationstestdaten

Was muss man bei der Planung beachten?

- Qualität hängt von richtiger Auswahl der Faktoren ab, die die Situation der Annotatoren vollständig und genau beschreiben
- relevante, aber nicht offensichtliche Faktoren können leicht übersehen werden
- statistische Ergebnisse können dafür aber ein Indiz sein (hohe Fehlervarianz)
- Faktorenanzahl nicht begrenzt
- *aber:* mehr Faktoren $\rightarrow$ komplexer
- zu wenige Faktoren $\rightarrow$ Fehlende Daten für Interpretation
- keine Regel, ab wann eine Komponente als „unwichtig“ interpretiert wird $\rightarrow$ Faustregel: 8%

Vergleich mit anderen Agreement-Maßen

- häufig benutzte Agreement-Maße: Übereinstimmungsrate und Kappa-Statistik
- messen die Gesamtverlässlichkeit
- G-Theorie: zum Vergleich des Einflusses verschiedener Facetten (und deren Interaktionen) auf die Verlässlichkeit
- *aber:* gibt keine Hinweise, welche Werte für die Verlässlichkeit verantwortlich sind ($\rightarrow$ Daten analysieren)
- *Deshalb:* **Übereinstimmungsrate und G-Theorie sollten nicht als konkurrierend, sondern als ergänzend angesehen werden**
- **Kappa-Statistik: erster Schritt um eine Vorstellung über das Disagreement zu bekommen $\rightarrow$ G-Theorie: zweiter Schritt zum Feststellen der Gründe für Unverlässlichkeit**

Zusammenfassung

- Dient dazu, herauszufinden, welche Facetten für Disagreement verantwortlich sind
- Analyse mehrerer Facetten gleichzeitig: Gründe für Unverlässlichkeit können besser verstanden werden und es können geeignete Gegenmaßnahmen getroffen werden
- Die Daten von Shriberg and Lof wurden reanalysiert
  - enge Transkripitonen unzuverlässiger als breite
  - Vermutung, dass der Grund dafür nicht bei den Annotatoren, sondern beim Schema liegt, konnte bestätigt werden

Inhaltsverzeichnis

1. Generalizability Theory
   - Überblick
   - Die Theorie
   - Anwendungsbeispiel
   - Planung einer G-Studie
   - G-Theorie vs. andere Agreement-Maße
   - Zusammenfassung
2. Einfluss von Annotationsqualität auf maschinelles Lernen
   - Einleitung
   - Die Methode
   - Auswertung
   - Fazit
   - Zusammenfassung

Wann ist eine Annotation verlässlich?

- oft wird eine Zuverlässigkeit handannotierter Daten von 0,8 in der Computerlinguistik als geeignet angesehen
- Zuverlässigkeit von 0,67 - 0,8: tolerabel
- Ursprung: Werk von Krippendorff (1980)
- benutzt eine spezielle Statistik: α-Statistik
- es wird angenommen, dass man diesen Wert auch als Faustregel für κ-Statistik benutzen kann

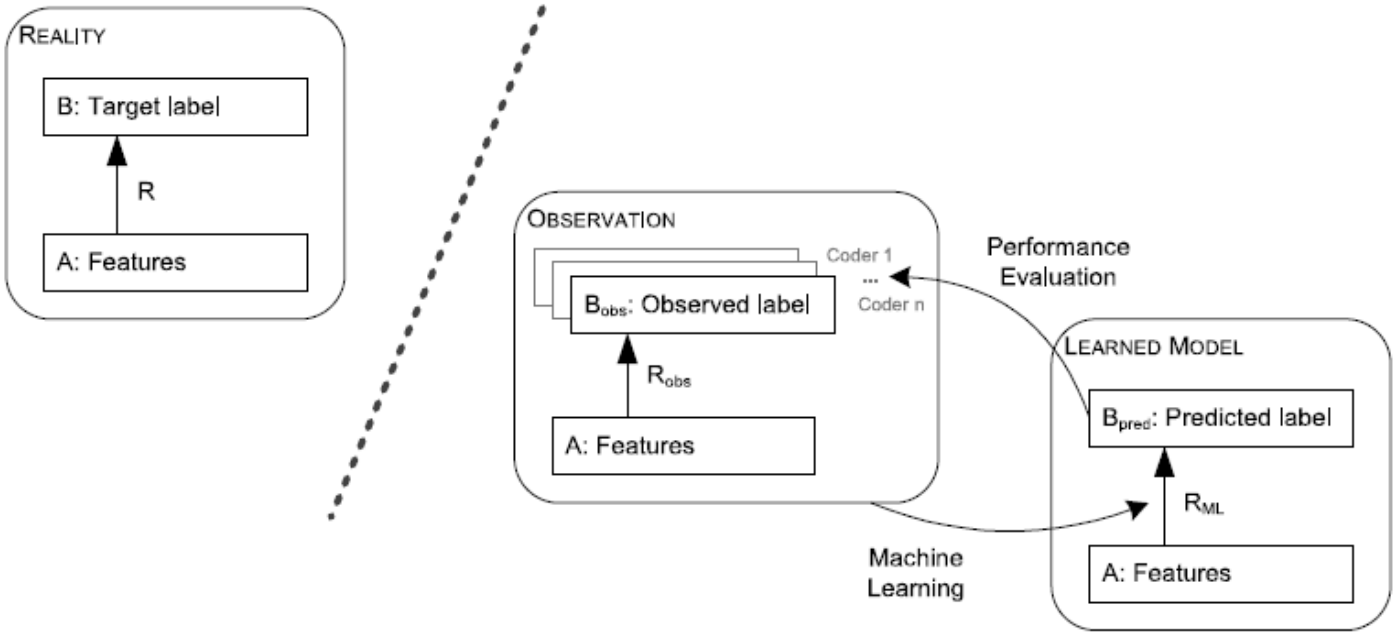
Stimmt das wirklich?

- heutzutage werden computerlinguistische Daten ganz anders verwendet als damals
- nämlich als Trainings- und Testmaterial für automatische Klassifizierer, anstatt zur Korrelationsberechnung zweier Variablen
- hier ist der Wert 0,8 nicht geeignet, weil Disagreement auf diese Classifier eine andere Wirkung hat als auf Korrelationen

Auch die Art des Disagreements ist wichtig.

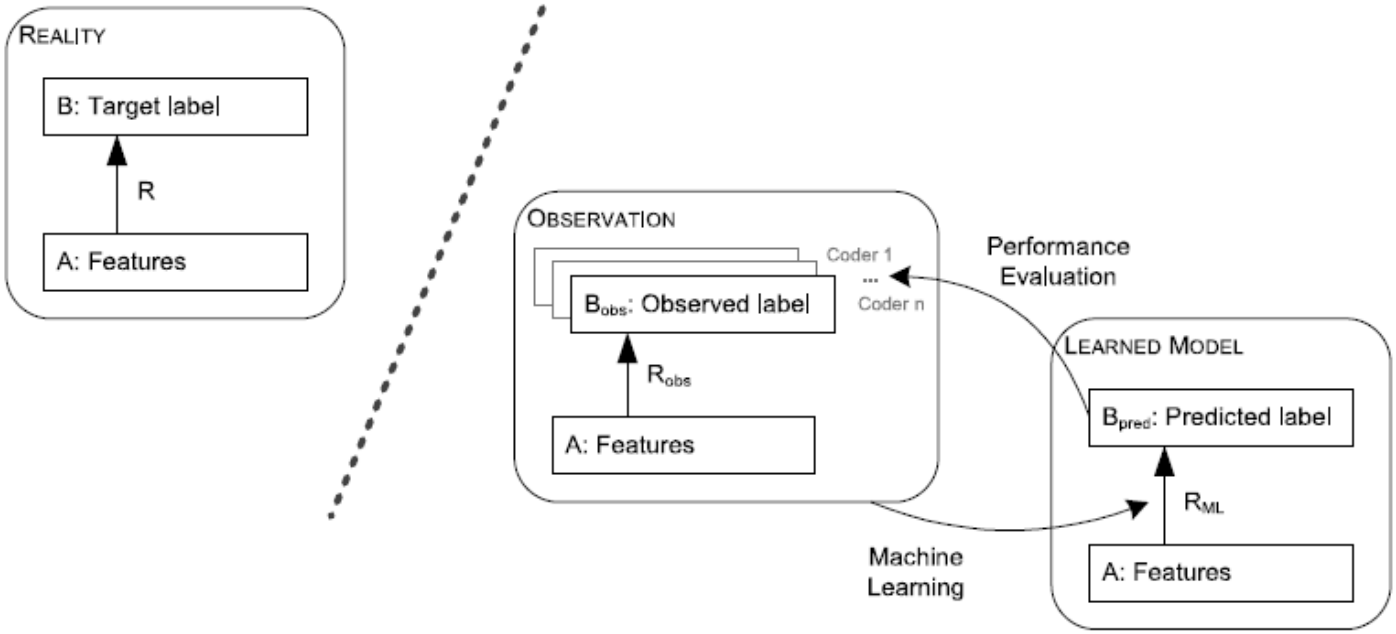
- Unterschiede zwischen zufälligen Störungen und Mustern
- Algorithmen für maschinelles Lernen sollen Muster erkennen und vorhersagen können
- *also:* zufällige, unsystematische Abweichungen in den annotierten Daten unwichtig
- *Im Gegensatz dazu:* systematische Muster: gefährlich $\rightarrow$ Classifier lernt sie
- **Auch Daten mit geringer Verlässlichkeit können zum maschinellen Lernen benutzt werden, wenn es sich um random Noise/zufällige Störungen handelt**
- **Bei Patterns/Mustern: Auch Daten, die ein hohes Agreement aufweisen, können zu falschen Ergebnissen führen**

Modell einer Annotation



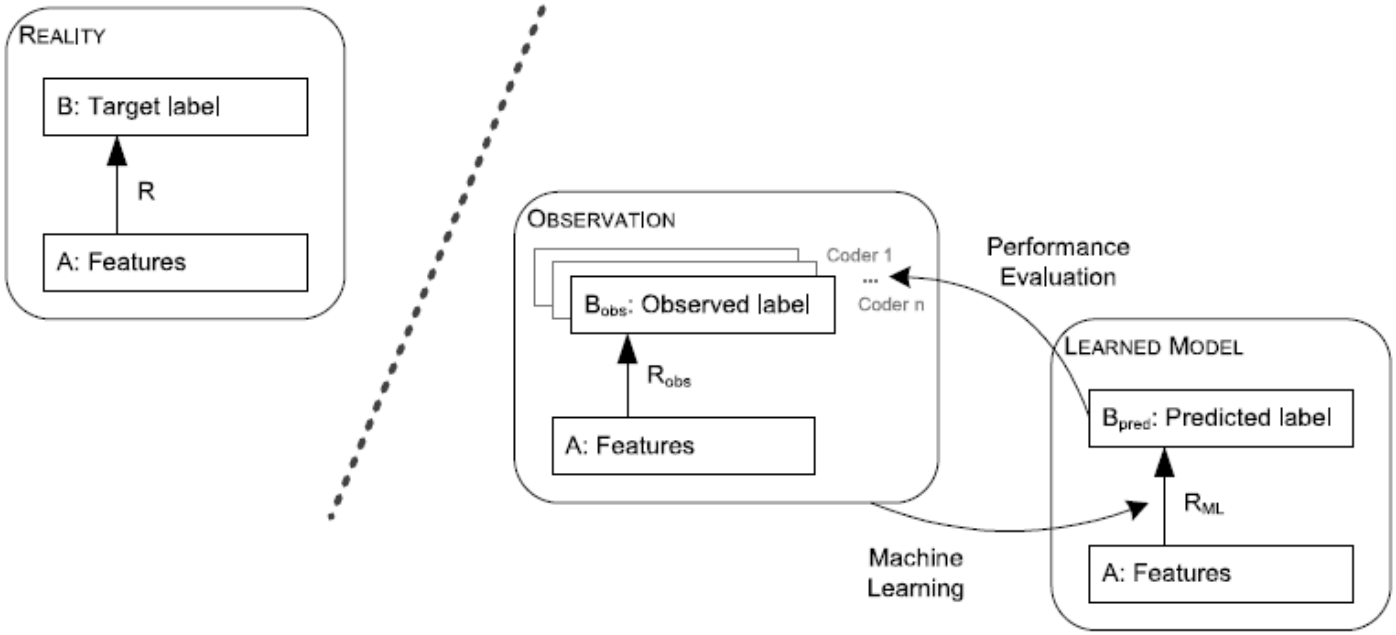
- Grafik zeigt Relation zwischen einem Feature (Eigenschaft) **A** und einem Klassen-Label **B**
- Häufige Aufgabe in der Computerlinguistik: annotierte Labels aufgrund von Features lernen

Modell einer Annotation



- links: **R:** wirkliche Beziehung zwischen Feature **A** und Label **B**
- theoretisch: nur ein richtiges Label für ein Feature
- in der Praxis: Annotatoren annotieren möglicherweise unterschiedliche Labels $\mathbf{B_{obs}}$ für das selbe Feature (mitte) $\rightarrow$ beobachtete Daten

Modell einer Annotation



- rechts: Classifier
- benutzt die aufgrund der beobachteten Daten gelernte Beziehung $\mathbf{R_{ML}}$, um das Label $\mathbf{B_{pred}}$ zu bestimmen
- Bei Disagreement hängt Auswahl der Daten zur Erstellung des Classifiers vom Projekt ab
  - *aber:* $\mathbf{R_{ML}}$ ist immer von den beobachteten Daten beeinflusst

Simulation einer Annotation

- Für Verlässlichkeit möchte man wissen, wie sich $\mathbf{R}$ und $\mathbf{R_{ML}}$ unterscheiden
- Problem: Ohne die realen Daten kann man nicht feststellen, wie gut $\mathbf{R}$ von $\mathbf{R_{ML}}$ beschrieben wird
- Man kann es nur mit den beobachteten Daten vergleichen
- *Deshalb:* Simulation der realen Welt $\rightarrow$ Unterschiede zwischen beobachteter und realer Performance können festgestellt werden

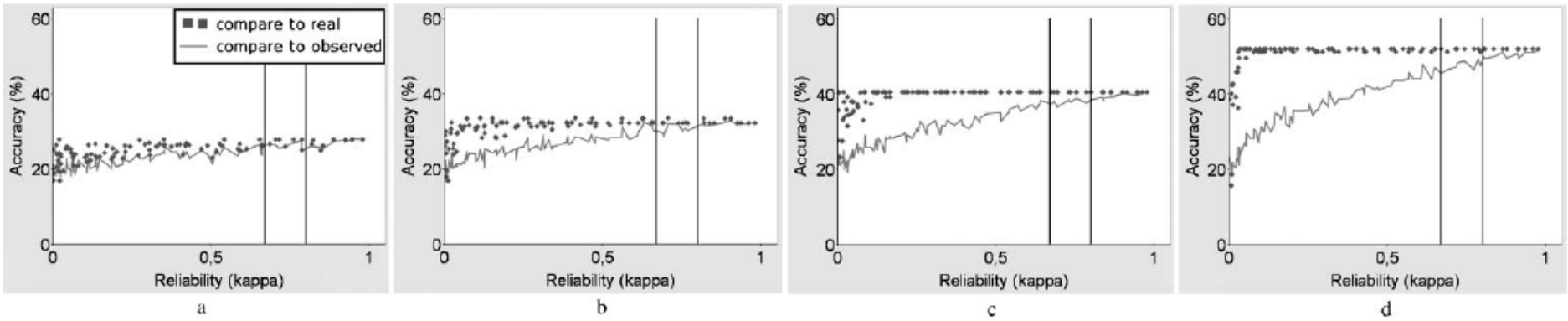
Simulation einer Annotation

- Bayes'sches Netz, um die „realen" Daten zu simulieren
  - 3000 samples von Features **A** und zugehörigen Labels **B**
  - ein einziges Feature mit 5 Werten
  - 5 mögliche Labels
  - Häufigkeit der Labels variiert zwischen 17 und 25%
- Modifizierung der Daten, um die beobachteten (annotierten) Daten $\mathbf{B_{obs}}$ zu simulieren
- Mit 2000 samples dieser Daten wird ein neuronales Netz trainiert
- Performanz des Netzwerks wird getestet
  - auf den restlichen 1000 samples von $\mathbf{B_{obs}}$
  - auf den selben 1000 samples der „realen" Daten

Simulation einer Annotation

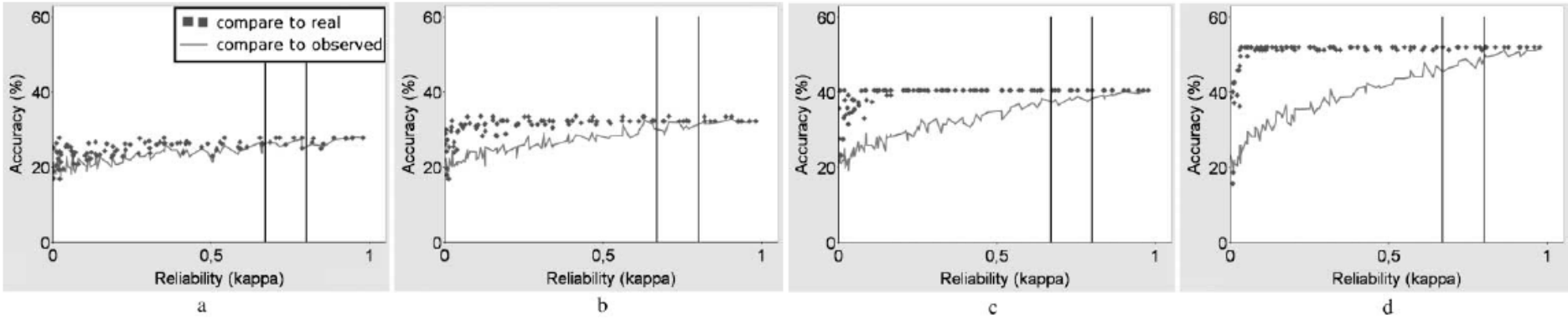
- Simulation muss in 3 Faktoren variiert werden
  - Stärke der Beziehung zwischen Features und Labels $\rightarrow$ Änderung der Übergangswahrscheinlichkeiten im Bayes'schen Netzwerk
  - Größe des Disagreements in $\mathbf{B_{obs}}$ $\rightarrow$ 200 Versionen: $\kappa = 0$ bis $\kappa = 1$
  - Art des Disagreements $\rightarrow$ Random Noise vs. überproportional häufiger Gebrauch eines Labels (Muster)

Bei Random Noise



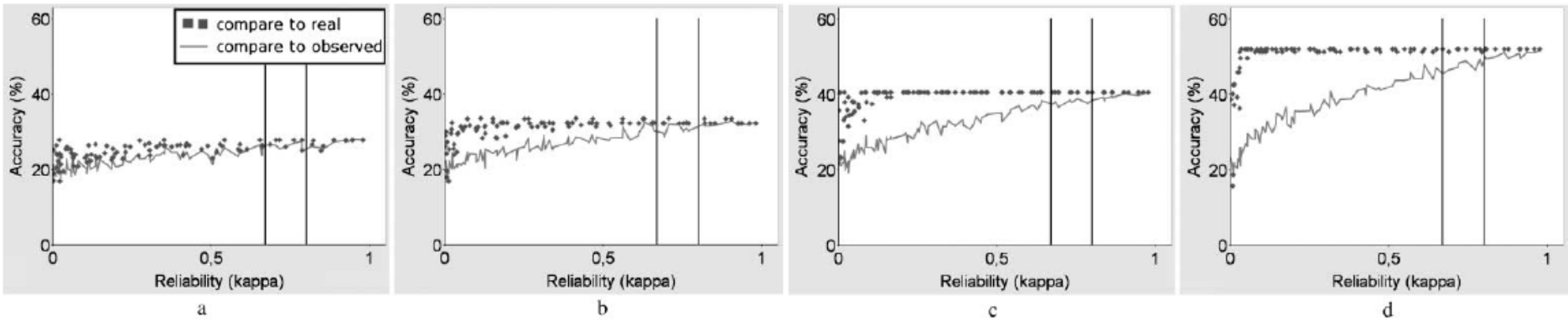
- Grafik zeigt Performance des Netzwerks wenn Random Noise vorliegt
- a: schwache Beziehung bis d: sehr starke Beziehung zwischen Features und Labels
- y-Achse: Akkuratheit (Prozentzahl, wie oft das Netzwerk Instanzen richtig klassifiziert hat)
- x-Achse: κ-Werte (Agreement)
  - schwarze Linien: $\kappa = 0{,}67$ und $\kappa = 0{,}8$

Bei Random Noise



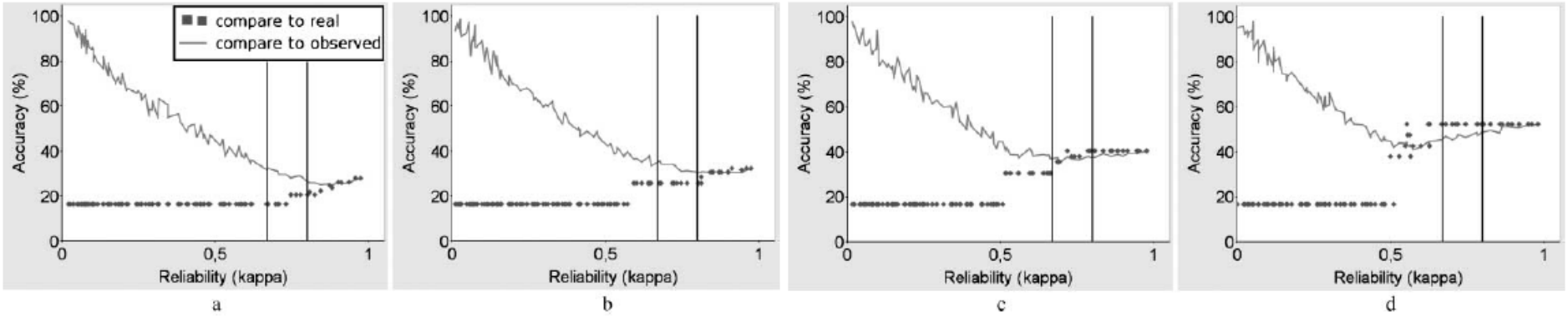
- Linie: Akkuratheit bezüglich der beobachteten Version der Test-Daten (wie man es normalerweise macht)
  - höheres $\kappa \rightarrow$ höhere Genauigkeit
  - $\kappa = 0$ (keine Übereinstimmung $\rightarrow$ zufällige Auswahl der Labels): Genauigkeit = 20 % (zu erwarten wenn Classifier die Labels zufällig auswählt)
  - höherer Zusammenhang zwischen Feature und Label $\rightarrow$ bessere Gesamtperformance

Bei Random Noise



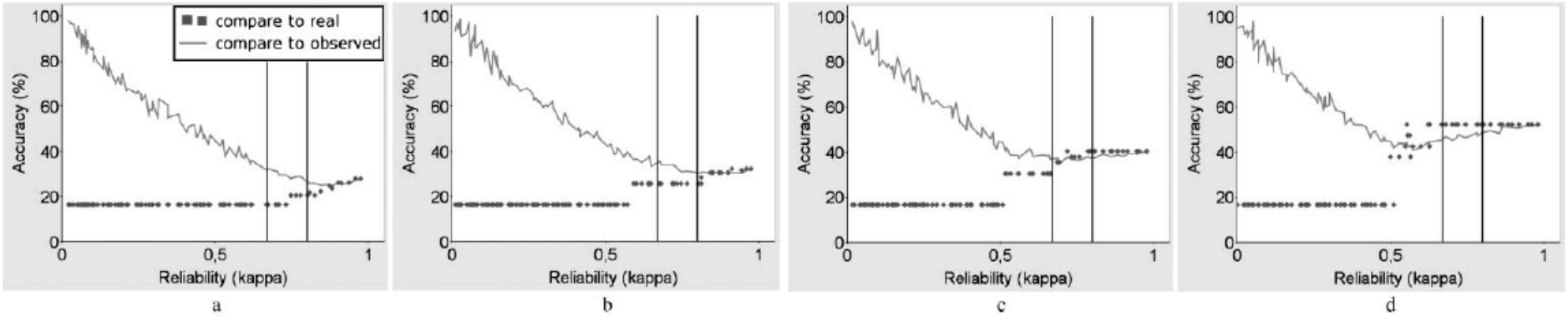
- Punkte: Akkuratheit bezüglich der echten Version der Daten
  - höhere Genauigkeit als bei beobachteten Daten
    - Classifier klassifiziert Instanzen richtig, die in den beobachteten Daten falsch waren
    - je stärker die Beziehung zwischen Feature und Label in den echten Daten, umso stärker ist der Effekt ausgeprägt
- **Classifier kann Fehler wegen Annotation Noise ignorieren, weil sie keine Muster enthalten, die er lernen kann**

Bei Patterns



- Grafik zeigt Performance des Netzwerk wenn ein Pattern (überproportional häufiger Gebrauch eines Labels) vorliegt
- komplett anderer Effekt als bei Annotation Noise
  - Performance auf beobachteten Daten sehr hoch (fälschlicherweise) (bis zu bestimmten κ-Werten)
  - macht Sinn:
    - κ niedrig bei stark ausgeprägter Überbenutzung eines Labels $\rightarrow$
    - Classifier übernimmt dieses Muster
    - Bei Test mit beobachteten Daten: Übereinstimmung nicht durch richtige Klassifizierung, sondern weil der Classifier das selbe Label überbenutzt wie der Annotator

Bei Patterns



- Bei starker Beziehung zwischen Features und Labels (d): Performance auf beobachteten Daten fälschlicherweise hoch nur unter etwa $\kappa = 0{,}55$
- *aber:* bei moderater oder starker Beziehung: schon bei eigentlich für tolerabel gehaltenen Größen von κ
- bei sehr schwacher Beziehung: sogar bei $\kappa > 0{,}8$
- **Classifier übernimmt Fehlerpatterns**

Was bedeutet das?

- Problem in den aktuellen Methoden der Computerlinguistik, da Überbenutzung eines Labels ein häufig gemachter Fehler ist
- in echter Annotation: Sowohl Annotation Noise als auch Muster
- *außerdem:* Beziehungsstärken variieren
- Macht Erkennen der Fehlergröße schwieriger, ändert aber nichts an dem Kern der Ergebnisse
- auch nicht auf κ festgelegt: wenn man stattdessen α benutzt: ähnliche Ergebnisse
- nicht nur bei maschinellem Lernen $\rightarrow$ andere statistische Daten werden auf eigene Weise durch Unterschiede zwischen Annotation Noise und Mustern beeinflusst

Was muss getan werden?

- statt eines Wertes für die Größe des Disagreements zu haben sollte man die Art des Disagreements bestimmen
- man sollte nach Patterns im Disagreement suchen
  - mögliche Lösung: Wenn **kein Pattern** vorhanden ist $\rightarrow$ **Assoziations- und Korrelationstests nur auf den Instanzen mit Disagreement** sollten Unabhängigkeit zeigen
  - *Problem:* Es kann sein, dass einige Patterns so nicht entdeckt werden, weil nicht genug Daten vorhanden sind
- Es muss eine Methode entwickelt werden, um Patterns im Disagreement zu finden
- In problematischen Fällen: eine solche Simulation machen und damit den Einfluss der unverlässlichen Daten auf Algorithmen für maschinelles Lernen abschätzen

Zusammenfassung

- klassische Agreementmaße reichen für viele Anwendungsgebiete nicht aus (z.B. maschinelles Lernen)
- trotz hohem Agreement können Fehler aufgrund von Patterns auftreten
- Andererseits können, wenn keine Patterns vorliegen, auch bei geringen Agreement-Raten gute Ergebnisse erzielt werden
- Man muss feststellen, welche Art von Disagreement vorliegt
- Dafür gibt es bis jetzt noch keine Methode

Quellenangabe

- Petra Saskia Bayerl and Karsten Ingmar Paul. Identifying Sources of Disagreement: Generalizability Theory in Manual Annotation Studies. *Computational Linguistics*, Volume 33, Issue 1 (March 2007).
- Dennis Reidsma and Jean Carletta. Reliability measurement without limits. *Computational Linguistics*, Volume 34, Issue 3 (September 2008).