Zuverlässigkeit von Daten

Caroline Sporleder

Computational Linguistics
Universität des Saarlandes

Sommersemester 2011

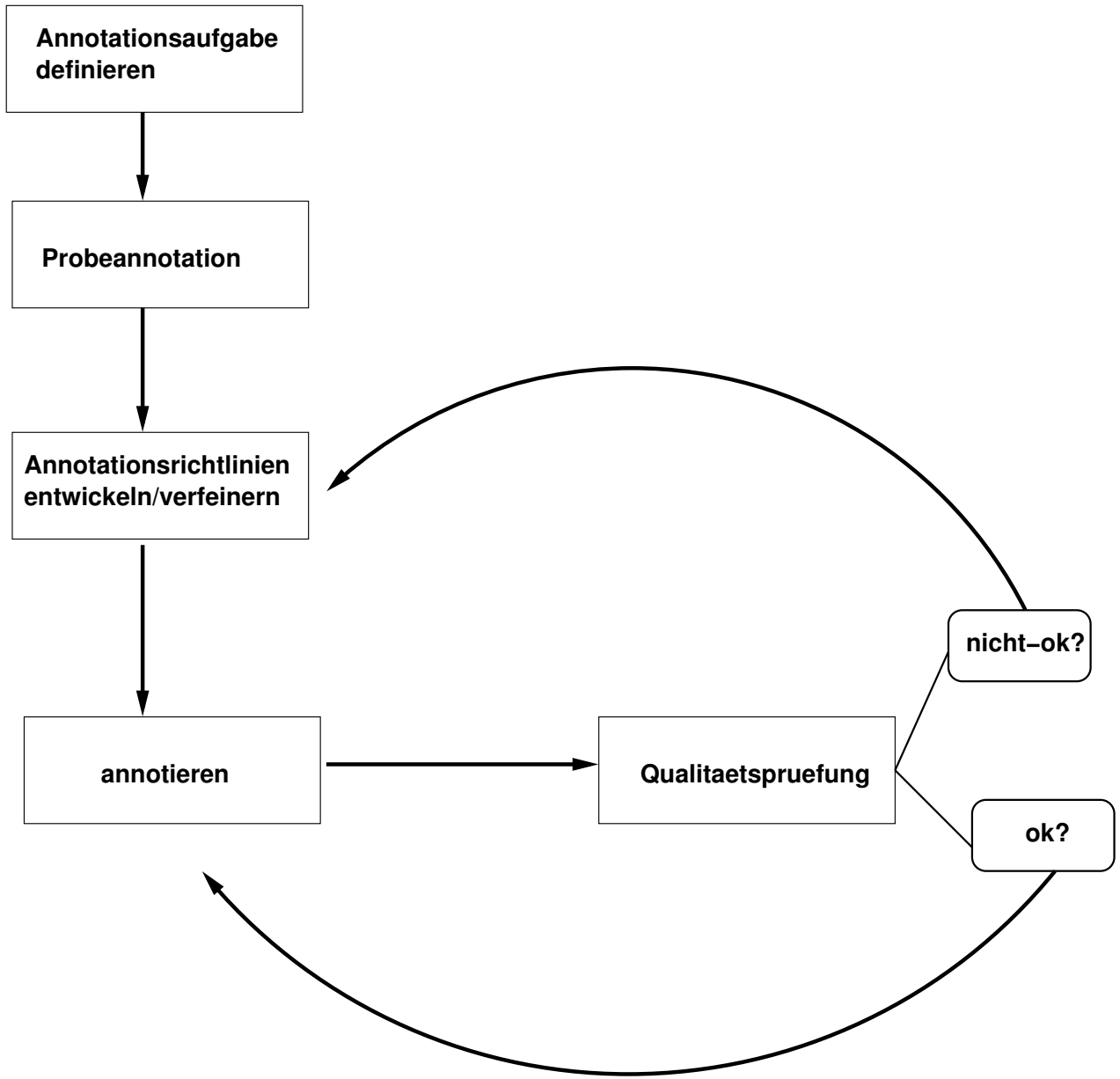
07.07.2011

Datenqualität ist wichtig

- Verlässlichkeit der ML-Ergebnisse hängt (auch) von der Verlässlichkeit der Daten ab
- Wie gut ist die Annotation?
  - Ist die Annotation linguistisch plausibel?
  - Ist die Annotation detailliert genug (nicht zu simplistisch, der Aufgabe angemessen)?
    Bsp: jeweils nur zwei Bedeutungen pro Wort vs. jede Bedeutungsnuance
  - Ist die Annotation verlässlich (d.h. reproduzierbar und kein Zufallsprodukt)?

⇒ **Verlässlichkeit und Detailliertheit stehen oft in einem Spannungsverhältnis**

Annotationsprozess



Annotationszuverlässigkeit

Inter-Annotator Agreement

Daten (oder eine Teilmenge davon) werden von mehr als einem Annotator annotiert und Übereinstimmung wird gemessen.

Intra-Annotator Agreement

Ein und derselbe Annotator annotiert einen Teil der Daten zweimal (mit mehreren Wochen Abstand). Danach wird die Übereinstimmung berechnet.

Inter-Annotator Agreement (IAA)

$$IAA = \frac{\#\ bereinstimmende\ \ Annotationen}{\#\ \ alle\ \ Annotationen}$$

Problem

Zufallsübereinstimmung (chance agreement) wird nicht rausgerechnet. Bei einfachen Annotationsentscheidungen (z.B. nur zwei Kategorien) ist ein hohes IAA zu erwarten, selbst dann, wenn die Annotatoren 'unintelligent' annotieren.

Cohens Kappa (1)

$$Kappa = \frac{(A-E)}{(1-E)}$$

A = Anzahl der übereinstimmenden Annotationsentscheidungen
E = Anzahl der (durch Zufall) erwarteten übereinstimmenden Annotationsentscheidungen

Beispiel

- 100 Instanzen: 50 mit Gold-Label X, 50 mit Gold-Label Y
  $\Rightarrow E_{random} = 0.5$
- ein Annotator, der per **Zufallsentscheid** annotieren würde, würde im Durchschnitt 50% der Instanzen korrekt annotieren
  $\Rightarrow A_{random} = 0.5$
  $\Rightarrow kappa = \frac{0.5-0.5}{1-0.5} = 0$; $IAA = 50\%$
- ein **geübter Annotator** mit 70% Übereinstimmung
  $\Rightarrow A_{gebt} = 0.7$
  $\Rightarrow kappa = \frac{0.7-0.5}{1-0.5} = \frac{0.2}{0.5} = 0.4$; $IAA = 0.7$

Cohens Kappa (2)

Probleme / Einschränkungen

- Bestraft Ungleichgewicht der Daten (gerechtfertigt?).
- Welcher Kappa-Wert ist gut genug? (In NLP wird $kappa > 0.6$ oft als 'gut' bis 'sehr gut' angesehen.)
- Cohens Kappa ist nur auf zwei Annotatoren anwendbar, Fleiss' Kappa für mehr als zwei Annotatoren.

Entwicklung eines Goldstandards

Als Goldstandard (engl. gold standard) bezeichnet man die Annotation, die als verlässlich und definitiv angesehen wird.

Maßnahmen zur Qualitätssicherung (1)

Annotationsrichtlinien

- dienen den Annotatoren als Leitfaden und machen die Annotation nachvollziebar
- detailliert aber nicht zu detailliert (Warum?
)

Maßnahmen zur Qualitätssicherung (1)

Annotationsrichtlinien

- dienen den Annotatoren als Leitfaden und machen die Annotation nachvollziebar
- detailliert aber nicht zu detailliert (Warum? Annotatoren sollen ihrer linguistischen Intuition folgen)

Maßnahmen zur Qualitätssicherung (2)

Vorwissen / Training der Annotatoren

- wer annotiert (Laien, Studenten vom Fach, Experten)?
- wieviel Training (sowohl zu wenig als auch zu viel kann schlecht sein)?

Maßnahmen zur Qualitätssicherung (3)

Verlässlichkeit durch mehrfache Annotation

- Adjudikation (Entscheidung durch Experten / durch Diskussion), wenn Annotatoren nicht übereinstimmen
- ggf. auch doppelte Adjudikation und Meta-Adjudikation, falls Adjudikatoren nicht übereinstimmen
- meist weniger empfehlenswert: einfacher Mehrheitsentscheid bei abweichenden Annotationen

Neue Wege der Datenannotation

- Games with a Purpose (GWAPs)
- Amazon Mechanical Turk und ähnliche Dienste

⇒ Daten werden von Laien annotiert
⇒ linguistisch komplexe Annotationsaufgaben müssen in geeigneter Weise vereinfacht werden
⇒ Qualitätssicherung besonders wichtig (oft durch massive Redundanz)

Aus der ML-Sicht

Bei Lernexperimenten auch immer die Zuverlässigkeit der Daten hinterfragen

- Wie schwierig ist die Annotationsaufgabe?
- Geben die Autoren IAA/kappa an?
- Wenn 'ja', wie verlässlich sind diese Werte (z.B. Datengröße)?
- Wer hat annotiert (Laien, Studenten vom Fach, Experten, die Autoren selber)?
- Gibt es (veröffentlichte) Annotationsrichtlinien? Wie detailliert und nachvollziehbar sind diese?
- Wieviel Training gab es?
- Wie wurde der Goldstandard erstellt (Qualitätssicherung)?

Zusammmenfassung

Was ihr gelernt haben solltet:

- Was ist bei der Annotation von Daten zu beachten?
- Wie kann die Qualität von annotierten Daten beurteilt werden?
- Welche Maße gibt es um Inter-Annotator-Agreement zu messen (Vor- und Nachteile)?

Bibliographie

J. Cohen (1960):
A coefficient of agreement for nominal scales.
In: Educational and Psychological Measurements, pp 37-46.

J. Carletta (1996):
Assessing agreement on classification tasks: The kappa statistic.
In: Computational Linguistiscs, 22:2, pp. 249-254.

J.L. Fleiss (1971):
Measuring nominal scale agreement among many raters.
In: Psychological Bulletin, 76:5, pp. 378-382.