

Die Repräsentation und Auflösung von ambigen Wortbedeutungen in der Computerlinguistik

Evaluierung

PD Dr. Sabine Schulte im Walde

Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart

27. März 2010

- Wie gut ist ein (computerlinguistisches) Modell?
- Evaluierungsmöglichkeiten:
 - Introspektion
 - intrinsische, Modell-spezifische Evaluierung
 - mehrfache, unabhängige Bewertungen
 - Vergleich gegen einen Gold Standard
 - extrinsische Evaluierung als Teil einer Anwendung

- ① menschliche Bewertungen
- ② intrinsische Evaluierung: Perplexität
- ③ Gold-Standard-Evaluierung: Precision, Recall, F-Score
- ④ Pseudo-Desambiguierung
- ⑤ Evaluierung durch Anwendung

Menschliche Bewertungen

- Menschliche Bewertungen können für die Evaluierung eines Modells benutzt werden.
- Möglichkeiten:
 - Erzeugen eines Gold-Standards im Voraus oder
 - mehrfache, unabhängige Bewertungen der Modell-Leistung mit sich anschließender
 - Berechnung der Übereinstimmung
- Vorgaben: psycholinguistisch fundierte Daten-Sammlung oder Spezifikation der Modell-Aufgaben

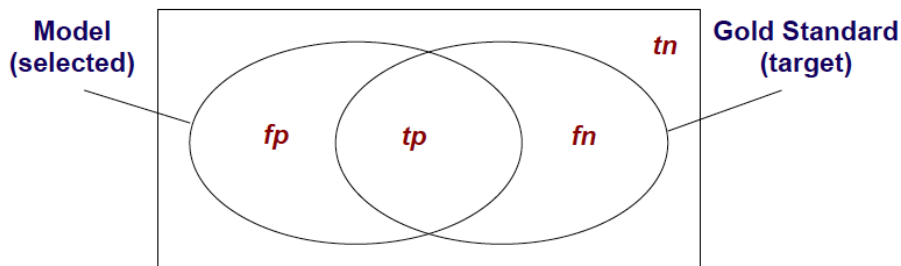
Menschliche Bewertungen: Übereinstimmung

- **Inter-Annotator-Agreement**,
z.B. Proportion der Übereinstimmung, κ
- Wie hoch ist die Übereinstimmung in Bezug auf die Aufgabe?
- Wie verlässlich und konsistent sind die menschlichen Bewertungen?
- Was ist die objektiv korrekte Lösung für die Aufgabe?
- Wie einfach/schwierig ist die Aufgabe?

Intrinsische Bewertung: Perplexität

- Wie gut beschreibt ein Modell die Daten?
- Perplexität \leftrightarrow Entropy
- Maß für die Unsicherheit bei der Modellierung
- Perplexität von k :
Unsicherheit bei einer Zufallsauswahl aus k Elementen
- nur anwendbar bei Wahrscheinlichkeitsmodellen

Gold-Standard-Evaluierung: True/False Positives/Negatives



tp: true positives

fp: false positives

tn: true negatives

fn: false negatives

	<i>actual</i>	
<i>system</i>	target	¬ target
selected	tp	fp
¬ selected	fn	tn

Gold-Standard-Evaluierung: Precision, Recall, F-Score

- **Precision**: Anteil der Einheiten, bei denen das Modell richtig liegt

$$P = \frac{tp}{tp + fp}$$

- **Recall**: Anteil der Zieleinheiten, die das Modell abdeckt

$$R = \frac{tp}{tp + fn}$$

- **F-Score**: Kompromiss zwischen Precision und Recall

$$F_1 \text{ (harmonisches Mittel): } F = \frac{2 * P * R}{P + R}$$

Pseudo-Desambiguierung

- Annotierte Daten sind teuer.
- Szenario für Pseudo-Desambiguierung:
 - automatische Erzeugung von Daten für die Evaluierung
 - Unterscheidung zwischen echten und nicht gesehenen Daten
- Beispiel: Word Sense Disambiguation
Desambiguierung zwischen Pseudo-Wörtern, z.B. *Tür/Banane*

Evaluierung durch Anwendung

- Modelle werden innerhalb einer natürlichsprachlichen Anwendung evaluiert.
- Beispiele:
 - Akquisition von Subkategorisierungsinformation → Parsing
 - Parsing → Akquisition von Subkategorisierungsinformation

- Modelle werden innerhalb einer natürlichsprachlichen Anwendung evaluiert.
- Beispiele:
 - Akquisition von Subkategorisierungsinformation → Parsing
 - Parsing → Akquisition von Subkategorisierungsinformation
 - semantisches Parsing → Word Sense Disambiguation
 - Word Sense Disambiguation → maschinelle Übersetzung

Referenzen: Evaluierung allgemein



Sidney Siegel and N. John Castellan.
Nonparametric Statistics for the Behavioral Sciences.
McGraw-Hill, Boston, MA, 1988.



Christopher D. Manning and Hinrich Schütze.
Foundations of Statistical Natural Language Processing.
MIT Press, Cambridge, MA, 1999.



Jacob Cohen.

A Coefficient of Agreement for Nominal Scales.

Educational and Psychological Measurement, 20:154–163, 1960.



Jean Carletta.

Assessing Agreement on Classification Tasks: The Kappa Statistic.

Computational Linguistics, 22(2):249–254, 1996.



Barbara Di Eugenio and Michael Glass.

The Kappa Statistic: A Second Look.

Computational Linguistics, 30(1):95–101, 2004.