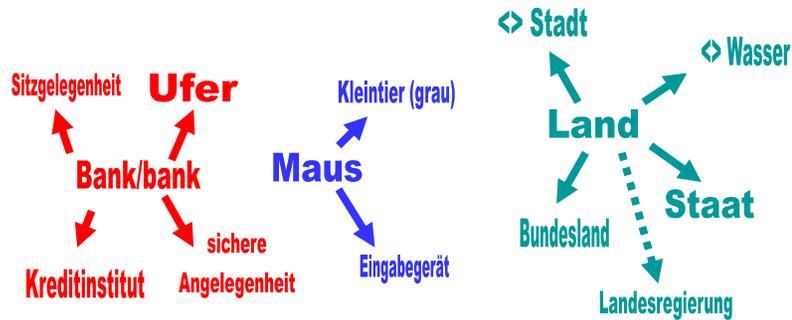


Kapitel 7 Word sense disambiguation



1

Methoden zur Wortsinnunterscheidung

- 3 verschiedene Methoden:
 - 7.2 supervised disambiguation
 - 7.3 dictionary-based disambiguation
 - 7.4 unsupervised disambiguation

1

Begriffe

- Upper Bound
obere Grenze für die Richtigkeit einer Sinnunterscheidung. Entspricht normalerweise den menschlichen Fähigkeiten
- Lower Bound
untere Grenze
Der Algorithmus sollte besser sein als raten

1

Begriffe

- Pseudowort

```
graph TD; A[bananadoor] --> B[banana]; A --> C[door];
```
- Wird verwendet, um Qualität eines Algorithmus zu testen

2 Supervised Disambiguation

- Symbole:
- w = ein Wort z.B. Bank
- s_i = i. Sinn z.B. Sitzgelegenheit
- c_i = i. Kontext z.B.:
... Der **ruinierte** Anzug, mit dem er sich auf die **frisch gestrichene Bank** setzte, hatte viel **Geld** gekostet. ...
- v_k = k. Wort, das den Kontext darstellt, also alle Wörter in c

2 Supervised Disambiguation

- Entscheide für s' ,
wenn $P(s'/c) > P(s_i/c)$ für $s_i \leftrightarrow s'$
- Baye's rule:
 $P(s/c) = P(s) * (P(c/s)/P(c))$
- → Entscheide für das s , für das
 $\log(P(s)) + \log(P(c/s))$
maximal wird
- Logarithmus weil weniger rechenintensiv

2 Supervised Disambiguation

- Naïve Baye's Assumption:
 $P(c|s) = \prod_{v \text{ in } c} (P(v|s))$
- $P(v|s)$ und $P(s)$ durch Zählen im Trainingsdatensatz (=Supervised)

2 Supervised Disambiguation

- v ist „good clue“ von c , wenn $P(v/c)$ größer als andere $P(v/c')$
- Also $P(\text{Geld}|\text{Kreditinstitut}) \gg P(\text{Geld}|\text{Sitzgelegenheit})$

3 Dictionary-Based Disambiguation

- Vergleich mit Lexikoneinträgen
- Im Lexikon steht meist ein Satz, der das Wort beschreibt
- Zwischen diesem Satz und dem Kontext wird die Schnittmenge gebildet
- Die Größe dieser Schnittmenge ist der Vergleichswert

3 Dictionary-Based Disambiguation

- Problem:
Wort, bzw. eine Bedeutung davon steht nicht im Lexikon
- Lösung: Adaptive Disambiguation
→ Ein neuer Thesaurus entsteht durch Kategorisieren von Wörtern.
Wörter die häufiger als zufällig nahe beieinanderstehen, müssen derselben Kategorie angehören.

4 Unsupervised Disambiguation

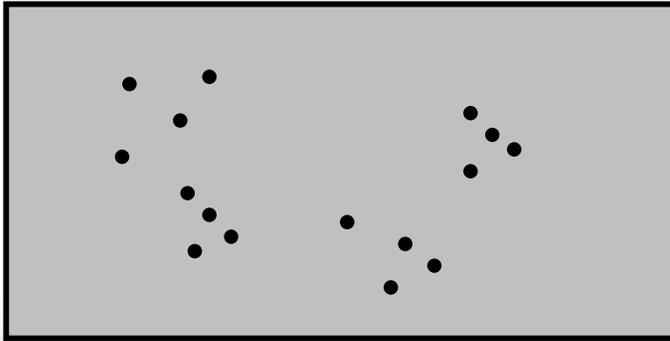
- Wenn einem nichts anderes einfällt...
- Also:
 - Wenn Wörterbücher unzureichend sind
 - Wenn Text zu fachspezifisch
 - Wenn Kein Corpus vorhanden

4 Unsupervised Disambiguation

- EM-Algorithmus
 - Eingabe
 - Zufällige Startwerte für $P(v/s)$ und $P(s)$
 - Anzahl der verschiedenen s
 - Ausgabe
 - Optimierte Werte für $P(v/s)$ und $P(s)$

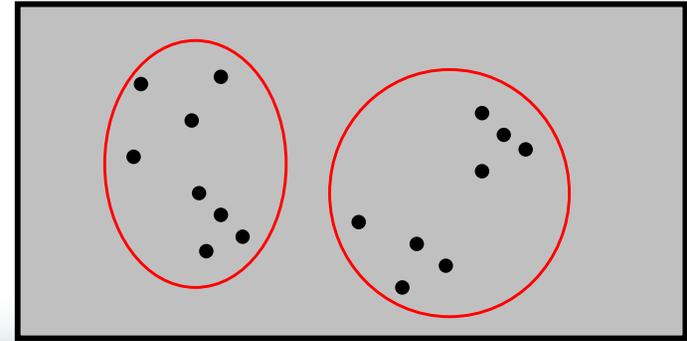
4 Unsupervised Disambiguation

- Das bedeutet: **Clustering**



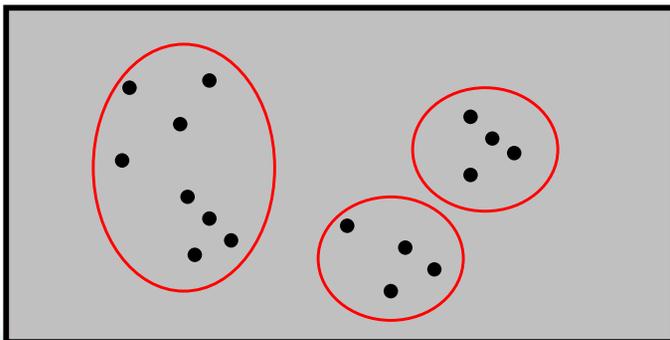
4 Unsupervised Disambiguation

- Ergebnis abhängig von der Anzahl der Kategorien



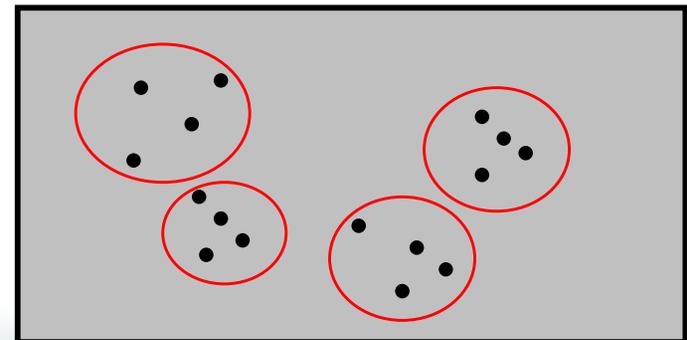
4 Unsupervised Disambiguation

- Ergebnis abhängig von der Anzahl der Kategorien



4 Unsupervised Disambiguation

- Ergebnis abhängig von der Anzahl der Kategorien



4

Unsupervised Disambiguation

- Ergebnis abhängig von der Anzahl der Kategorien

