n-gram models

- Erraten des folgenden Wortes
- die n-1 letzten Worte waren vorhanden
- Normalerweise n = 2,3,4
- größere n zu aufwendig
- Stemming zum verkleinern des Vokabulars

Wahrscheinlichkeits Abschätzung

- N Anzahl der Traningssätze (Worte)
- B Anzahl der Zielworte
- V Vokabelgröße
- w_1n n-gram w_1 ... w_n
- C(w_1...w_n) Häufigkeit des n-grams w_1n
- r Häufigkeit eines n-grams
- f(.) Häufigkeits Abschätzung eines Models

Maximum Likelihood Estimation

$$P_{MLE}(w_1 \ldots w_n) = C\left(\frac{w_1 \ldots w_n}{N}\right)$$

$$P_{MLE}(w_n / w1 \ldots w_{n-1}) = \frac{C(w_1 \ldots w_n)}{C(w_1 \ldots w_{n-1})}$$

Problem MLE

- Viele Kombinationen des Vokabulars treten nicht in erscheinung
- MLE berechnet für nicht aufgetretene Ereignisse eine Wahrscheinlichkeit von 0, obwohl es eine solche Kombination geben kann.

0-Wahrscheinlichkeit vermeiden

- Laplace's Gesetz
- Addieren von 1 zur Häufigkeit des n-grams
- Problem: da es viele ungesehene n-grams gibt, wird zu viel Wahrscheinlichkeitsraumes auf nicht gesehene n-grams verschwendet
- im Text als Beispiel 99,97%

$$P_{lap}(w_1..w_n)=\frac{C(w_1...w_n)+1}{N+B}$$

Lidstone's Gesetz

- Addieren eines geringeren Anteils
- Lambda meist 0.5

$$P_{Lid}(w_1 \ldots w_n) = \frac{C(w_1 \ldots w_n) + \lambda}{N + B\lambda}$$

Testen

- Nie auf dem Trainingssatz testen!
- Vor dem Training einen kleinen Anteil (5-10%) des Datensatzes als Testsatz zurückhalten.
- Auch sollte bei der Weiterentwicklung des Algorithmus nicht auf den Testdatensatz geschaut werden.

Varianz

- Testen auf einem Testdatensatz gibt nur ein Ergebnis.
- Mehrere Testdatensätze verwenden um verschiedene Ergebnisse zu erhalten und um sie vergleichen zu können.

t-Test verschiedene Systeme

- Wie schon aus Kapitel 5 bekannt, kann hier der t-Test verwendet werden um verschiedene Systeme miteinander zu vergleichen.

Cross-Validation

- Verschiedene Sätze als Trainingssätze und Testsätze verwenden.
- deleted estimation
- Leaving-one-out
  - Jeden Token einmal aus dem Trainingssatz entfernen und Ergebnis beobachten

Good-Turing Estimation

- weitere Methode zur Bestimmung der Wahrscheinlichkeitsabschätzung
- Verwendet frequencies of frequencies

Discounting

- Alternative zum addieren von Wahrscheinlichkeiten bei nicht gesehenen n-grams
- Abziehen eines Anteils der Wahrscheinlichkeit bei gesehenen n-grams

Probleme

- Finden der korrekten Wahrscheinlichkeiten eines n-grams
- Berücksichtigen von nicht vorhanden n-grams im Trainingssatz
- Über/Unterbewertung bestimmter n-grams