Musterlösungen zum 5. Übungsblatt

Sebastian Padó

30. Januar 2006

Aufgabe 5.1a

Konfusionsmatrix für NAdj:

	Echte NAdj	Echte Adj
Als NAdj klass.	900	10
Als Adj klass.	80	10

Evaluationsmaße für NAdj:

- Recall: 900 / (900+80) = 900/980 = 91.8%
- Precision: 900 / (900+10) = 900/910 = 98.9%
- F-Score: (2*0.918*0.989)/(0.918 + 0.989) = 0.952

Aufgabe 5.1b

Die Ergebnisse für NAdj sind fast perfekt, sehr viel besser als die Ergebnisse für Adj. Trotzdem ist der Klassifikator nicht gut geeignet: die Klasse Adj ist ja die "interessante" Klasse, und die Ergebnisse für diese Klasse sind sehr schlecht.

Aufgabe 5.2

- Es gibt zwei Klassen: Adj, NAdj
- Zwei Features, mit je zwei Werten. Die Größe des Ereignisraumes ist das Produkt der Wertanzahl aller Features, hier $2 \times 2 = 4$ Ereignisse.

- Wir haben alle Ereignisse einmal gesehen (100%). Das heisst, für jedes Ereignis auf neuen Daten können wir eine Entscheidung treffen (ob sie richtig ist oder nicht, ist eine ganz andere Frage).
- Für das erste Ereignis (selbst nicht Artikel, nächstes Wort nicht kapitalisiert) weist der Klassifikator die Klasse NAdj zu. Dasselbe tut er für alle anderen Ereignisse – die Klasse Adj wird nie zugewiesen. Das heißt, der Klassifikator ist völlig unbrauchbar.
- Für jedes Ereignis wird die zugewiesene Klasse danach bestimmt, welche am häufigsten mit diesem Ereignis gesehen wurde. Wenn es sehr seltene und sehr häufige Klassen gibt, ist es für die kleinere Klasse sehr schwer, überhaupt für irgendein Ereignis häufiger vorzukommen als die häufige Klasse.

Aufgabe 5.3

- Probabilistische Grammatiken weisen ihren Bäumen (Analysen) Wahrscheinlichkeiten. Diese Wahrscheinlichkeiten können im besten Fall (wenn es sich um eine "gute" Grammatik handelt) als Grammatikalität interpretiert werden, d.h. grammatische Sätze erhalten hohe, ungrammatische niedrige Wahrscheinlichkeiten
- Berechnung von Regelwahrscheinlichkeiten:
  - Für jedes Nonterminal N zählt man, wie oft es in der Baumbank vorkommt ($|N|$).
  - Für jede Regel der Form $N \rightarrow RS$, d.h. die N zur rechten Seite RS expandiert, zählt man, wie oft sie in der Baumbank vorkommt ($N \rightarrow RS$).
  - Die Wahrscheinlichkeit der Regel ist $P(RS|N) = \frac{|N \rightarrow RS|}{|N|}$.
- Zusammenhang zwischen Expansion einer Regel und ihrem lexikalischen Kopf. (Andere Beispiele: Grossvaterkategorie, etc.) am Beispiel VP.
  - Angenommen, es gibt nur intransitive und transitive Verben. Dann gibt es zwei Regeln $P(V|VP)$ und $P(VNP|VP)$. Diese Wahrscheinlichkeiten sollten unterschiedlich gross sein, je nachdem, ob das Verb V zu einem transitiven oder intransitiven Verb expandiert:
    - Intransitive Verben: $P(V|VP) > P(VNP|VP)$.
    - Transitive Verben: $P(V|VP) < P(VNP|VP)$.

    Das ist aber unmöglich. Grund: die Regeln wissen nicht, was "weiter unten" im Baum passiert.
  - Das Regelformat kann erweitert werden zu $P(RS|NT, Kopf)$. In Beispiel oben kann man dann schreiben:

    - Intransitive Verben: $P(V|VP, schlafen) > P(VNP|VP, schlafen)$.
    - Transitive Verben: $P(V|VP, studieren) < P(VNP|VP, studieren)$.
- Die neue Berechnung der Regelwahrscheinlichkeiten:
    - Für jedes Nonterminal N **mit dem lexikalischen Kopf** l zählt man, wie oft es in der Baumbank vorkommt ($|N_l|$).
    - Für jede Regel der Form $N_l \rightarrow RS$, d.h. die N_l zur rechten Seite RS expandiert, zählt man, wie oft sie in der Baumbank vorkommt ($|N_l \rightarrow RS|$).
    - Die Wahrscheinlichkeit der Regel ist $P(RS|N, l) = \frac{|N_l \rightarrow RS|}{|N_l|}$.
- Problem: Es gibt sehr viel mehr Nonterminale (je eins für ein altes Nonterminal und einen lexikalischen Kopf), und dementsprechend auch mehr Regeln: "sparse data".