Einführung in die Computerlinguistik

Übung 6

Jan Hendrik Dithmar
Matrikelnummer 2031259

Aufgabe 6.1

(a)

$$\text{Recall} = \frac{\text{True\_positives}}{\text{True\_positives} + \text{False\_negatives}} = \frac{900}{900 + 80} = \frac{900}{980} \approx 0.92$$

$$\text{Precision} = \frac{\text{True\_positives}}{\text{True\_positives} + \text{False\_positives}} = \frac{900}{900 + 10} = \frac{900}{910} \approx 0.99$$

$$\text{F-Score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \approx 0.95$$

(b)

Die Ergebnisse für die Klasse NAdj sind signifikant besser.

(c)

Meiner Meinung nach ist die Evaluation der Klasse Adj aussagekräftiger, da z.B. die Genauigkeit (Precision) in diesem Fall nur 11% beträgt, was ein schlechtes Ergebnis darstellt. Im Vergleich dazu liegt die Genauigkeit bei der anderen Klasse bei 99%, was zu der Annahme verleiten würde, dass das Modell für die Praxis brauchbar wäre, was meiner Meinung nach nicht der Fall ist.

Aufgabe 6.2

(a)

Es gibt 2 Klassen: Adj und NAdj.

(b)

Es gibt 2 Features: Selbst Artikel? und Nächstes Wort kapitalisiert?. Jedes dieser Features hat jeweils 2 Werte (wahr oder falsch).

(c)

Es gibt 4 mögliche Ereignisse:

- Selbst kein Artikel und nächstes Wort nicht kapitalisiert.
- Selbst kein Artikel und nächstes Wort kapitalisiert.
- Selbst Artikel und nächstes Wort nicht kapitalisiert.
- Selbst Artikel und nächstes Wort kapitalisiert.

Die Größe des Ereignisraums ist also 4.

(d)

Das Modell hat für den gesamten Ereignisraum Trainingsinstanzen gesehen, was bedeutet, dass die vollständige Abdeckung auch bei neuen Daten gewährleistet sein kann. (Ich vermute, dass eine vollständige Abdeckung auf Grund mancher Datensätze nicht immer möglich ist.)

(e)

Geht man z.B. von der ersten Zeile der Tabelle aus, so gibt es in diesem Fall mehr Elemente in NAdj als in der Adj-Klasse. Entscheidet man sich bei den Regeln für die Mehrheit der Fälle, so kommen folgende Regeln zu Stande:

- Selbst kein Artikel und nächstes Wort nicht kapitalisiert $\Rightarrow$ NAdj
- Selbst kein Artikel und nächstes Wort kapitalisiert $\Rightarrow$ NAdj
- Selbst Artikel und nächstes Wort nicht kapitalisiert $\Rightarrow$ NAdj
- Selbst Artikel und nächstes Wort kapitalisiert $\Rightarrow$ NAdj

Diese Regeln sind natürlich wenig plausibel, da einfach alles als NAdj erkannt wird.
Würde man bei den ersten zwei Regeln auf Adj schließen, weil deren Anzahl > 0 ist, so wäre das Modell trotzdem nicht plausibel.

(f)

Das Problem besteht beim naiven statistischen Modell darin, dass man nicht weiß, für welche Folgerung man sich entscheiden soll. Die Ursache liegt darin begründet, dass die Unterscheidung bei den Ereignissen nicht fein genug ist. Würde man hier besser differenzieren, so wären wohl auch die Ergebnisse eindeutiger.