Aufgabenblatt 5 — 21.11.2019

Einführung in die Computerlinguistik (WS19/20) — Abgabe bis 30.11.2019

Aufgabe 1) N-Gram Modelle Rechnen **Punkte: 10**

Betrachten Sie folgendes Trainingskorpus:[1]

- `<s> Auch Horus war der Sohn von Osiris </s>`
- `<s> Es war Seth der Osiris tötete </s>`
- `<s> War auch Horus tot </s>`

Beachten Sie während der ganzen Übung keine Groß- und Kleinschreibung. Modellieren Sie auch die Satzgrenzen <s> (Satzbeginn) und </s> (Satzende).

1. Erstellen Sie ein Unigramm- und ein Bigramm-Modell aus dem Trainingskorpus. Hierbei soll jedes Wort des Trainingskorpus Bestandteil des Vokabulars sein (inklusive der Satzgrenzen). (3 Punkte)

2. Berechnen Sie mit Ihrem Unigramm-Modell die Wahrscheinlichkeit für folgenden Satz ohne Smoothing:
   `<s> Seth tötete Osiris </s>`. (1 Punkt)

3. Berechnen Sie mit ihrem Bigramm-Modell die Wahrscheinlichkeit für folgenden Satz ohne Smoothing:
   `<s> War Seth der Sohn von Osiris </s>`.
   Was ist die Perplexität, wenn dies Ihr ganzes Testkorpus wäre? (2 Punkte)

4. Berechnen Sie mit ihrem Bigramm-Modell die Wahrscheinlichkeit für folgenden Satz (ohne Smoothing):
   `<s> Tötete Seth auch Horus </s>`
   Erklären Sie das Ergebnis. Hätte man ein anderes Ergebnis, wenn man die Modelle auf viel größeren Datenmengen trainiert hätte? (2 Punkte)

5. Berechnen Sie nun erneut die Wahrscheinlichkeit für den Satz:
   `<s> Tötete Seth auch Horus </s>`
   mit dem Bigramm-Modell. Verwenden Sie nun jedoch verschiedene Arten von Smoothing, die Sie in der Vorlesung kennengelernt haben: (2 Punkte)

   - Laplace Smoothing
   - Interpolationssmoothing (Wählen Sie $\lambda_1 = \lambda_2 = 0.5$) Hier müssen Sie die in der Vorlesung gesehene Formel für Interpolationssmoothing für Trigramme auf Interpolationssmoothing für Bigramme anpassen.

[1] Satzzeichen wurden hier ignoriert, deswegen gibt es im zweiten Satz kein Komma vor dem Relativsatz. Bitte benutzen Sie das Korpus genau wie vorgegeben.

Aufgabe 2) Ngram-Modelle Fragen **Punkte: 8**

1. Wir haben in der Vorlesung besprochen, dass die Inklusion von Satzanfängen und Satzenden es erlaubt, zu modellieren, ob ein Wort häufiger am Satzanfang/Satzende vorkommt als sonst. Ein Beispiel war, dass das Wort *and* wohl eher unwahrscheinlich am Satzanfang/Satzende ist. Geben Sie ein Beispiel für ein Wort oder eine Wortgruppe, welche häufiger am Satzanfang vorkommen als sonst. (1 Punkt)

2. Selbst in einem Unigrammmodell kann es sinnvoll sein, Satzanfänge und Satzenden zu modellieren. Welche Information geben Ihnen die Satzgrenzen da? (1 Punkt)

3. n-gram Modelle schauen nur eine sehr begrenzte Anzahl von Tokens zurück. Beschreiben Sie zwei linguistische Phänomene (im Deutschen oder Englischen), wo dies ein Problem darstellt. (2 Punkte)

4. Eine weitere Annahme von n-gram Modellen, die wir in der Vorlesung nicht besprochen haben, ist dass die Wahrscheinlichkeiten *stationär* sind, d.h. die Wahrscheinlichkeit, dass eine Wortsequenz vorkommt, wird als gleich angesehen, egal ob wir annehmen, dass wir uns im ersten, zweiten oder 500sten Satz eines Textes befinden. Warum trifft man wohl diese Annahme? Geben Sie zwei (grundlegend verschiedene) Beispiele, wo diese Annahme nicht zutrifft. (4 Punkte)

Aufgabe 3) Language Identification Fragen **Punkte: 2**

Wir haben in der Vorlesung das Modell von Cavnar und Trenkle kennengelernt, dass mittels Buchstaben-ngrams Sprachen erkennt. Im Normalfall wird das Leerzeichen auch als Buchstabe mitbenutzt. Begründen Sie, warum dies eine gute Idee ist. Geben Sie zwei verschiedenartige Begründungen. (2 Punkte)