Schriftliche Prüfung zur Übung
Statistische Methoden in der maschinellen Sprachverarbeitung
WS 2018/19
Dozent: Helmut Schmid

Sie haben für die Bearbeitung 60 Minuten Zeit.

Thema der Prüfung ist die Berechnung der geglätteten Wahrscheinlichkeiten eines N-Gramm-Sprachmodelles über Buchstaben. Die Länge der N-Gramme ist dabei nicht fest vorgegeben. Die Wahrscheinlichkeiten sollen mit der interpolierten Backoff-Methode geglättet werden. Die Häufigkeiten der Buchstaben-N-Gramme sind dabei bereits gegeben.

Aufgabe 1) Welche Python-Datenstruktur eignet sich zur Repräsentation der bereits berechneten Häufigkeiten?

Hinweis: Es gibt hier mehrere Möglichkeiten. Wählen Sie eine aus, mit der Sie dann die folgenden Aufgaben lösen. (1 Punkt)

Aufgabe 2) Schreiben Sie eine Funktion mit dem Namen **discount**, welche die N-Gramm-Häufigkeiten als Argument bekommt und den Discount zurückgibt.

Hinweis: Den Discount erhalten Sie, indem Sie N1 durch N1 plus 2 mal N2 teilen, wobei N1 die Zahl der N-Gramme mit Häufigkeit 1 ist. (4 Punkte)

Aufgabe 3) Schreiben Sie eine Funktion **estimate_prob**, welche die N-Gramm-Häufigkeiten als Argument bekommt und dann von jeder N-Gramm-Häufigkeit den Discount abzieht und das Ergebnis durch die Kontexthäufigkeit teilt. Die Kontexthäufigkeit eines N-Grammes **g** bekommen Sie, indem Sie die Häufigkeiten aller N-Gramme summieren, die sich höchstens im letzten Element von **g** unterscheiden.

Die Werte, die für jedes N-Gramm berechnet wurden, geben Sie in einer geeigneten Datenstruktur zurück. Diese Werte werden im Folgenden als *angepasste relative Häufigkeiten* bezeichnet. (10 Punkte)

Aufgabe 4) Schreiben Sie eine Funktion **compute_backoff_factors**, welche die Tabelle mit den angepassten relativen Häufigkeiten aus der vorherigen Aufgabe als Argument erhält und die Backoff-Faktoren für die verschiedenen Kontexte berechnet, indem Sie die Wahrscheinlichkeiten aller N-Gramme in der Tabelle summiert, die bis auf ein zusätzliches letztes Element mit dem Kontext-N-Gramm identisch sind. Die Summe wird dann von 1 abgezogen, um den Backoff-Faktor des Kontextes zu erhalten. Die Tabelle mit den Backoff-Faktoren geben Sie zurück. (6 Punkte)

Aufgabe 5) Schreiben Sie eine Funktion **get_prob**, welche ein N-Gramm als Argument erhält und die geglättete Wahrscheinlichkeit des letztes Elementes des N-Grammes gegeben die vorherigen Elemente zurückgibt. Dazu muss zu der angepassten relativen

Häufigkeit des N-Grammes das Produkt aus Backoff-Faktor und geglätteter Backoff-Wahrscheinlichkeit addiert werden. Die Backoff-Wahrscheinlichkeit wird dabei rekursiv mit derselben Funktion **get_prob** berechnet.

Die angepasste relative Häufigkeit jedes N-Grammes kann in dem globalen Dictionary **prob** nachgeschlagen werden. Diese Werte wurden in der vorherigen Aufgabe berechnet. Das Dictionary **prob** enthält aber auch die Werte für alle kürzeren N-Gramme. Die Backoff-Faktoren des Kontextes jedes N-Grammes können in einem Dictionary **backoff** nachgeschlagen werden. Die Rekursion zur Berechnung der geglätteten Wahrscheinlichkeiten soll bei Unigrammen enden.

Geben Sie bei dieser Aufgabe zusätzlich an, welche Defaultwerte die beiden Dictionaries prob und backoff liefern sollten, falls ein Key nicht definiert ist.

(9 Punkte)

(30 Punkte insgesamt)

Viel Erfolg!