

Sprachverarbeitung: Musterlösung zur Übung 23

Statistische Sprachmodellierung

Vorbemerkung: In dieser Übung wird angenommen, dass die Wortfolgen der Sprachen L_1 und L_2 sehr lang sind und deshalb die Randeffekte bei den N-Gram-Sprachmodellen vernachlässigt werden können (d.h. $P(v_i v_j) \approx 0$ für $v_j = \text{END}$).

Aufgabe 1: N-Gram-Sprachmodelle einfacher Sprachen

Bei den vorliegenden einfachen Sprachen ist ein N-Gram-Sprachmodell dann angemessen, wenn N minimal ist, so dass die Sprache (im Sinne der Statistik) vollständig bestimmt ist. Die gesuchten N-Gram-Sprachmodelle sind demnach:

- a) Die Sprache L_1 kann mit einem Bigram-Sprachmodell beschrieben werden, weil die einzige Bedingung für das momentane Wort ist, dass es nicht gleich sein darf wie das Wort davor, also $P(w_k = v_i | w_{k-1} = v_i) = 0$ ist. Die Bigram-Wahrscheinlichkeiten $P(v_i | v_i)$ müssen also null sein. Alle übrigen bedingten Wahrscheinlichkeiten $P(v_j | v_i)$ mit $i \neq j$ sind gleich. Sie lassen sich mit der Bedingung $\sum_j P(v_j | v_i) = 1$ bestimmen zu $P(v_j | v_i) = \frac{1}{|V|-1} = \frac{1}{3}$, wobei $|V|$ die Vokabulargrösse ist.

Weil in der Sprache L_1 für keines der Wörter v_i aus dem Vokabular $V = \{v_1, v_2, \dots, v_{|V|}\} = \{a, b, c, d\}$ eine Einschränkung besteht, gilt: $P(v_i) = \frac{1}{|V|} = \frac{1}{4}$ (wird in Aufgabe 2a gebraucht um $P(v_i v_j)$ zu berechnen).

- b) Aus den Angaben zur Sprache L_2 ergeben sich die folgenden Gleichungen:

$$P(x) = \frac{3}{2}P(y) \tag{1}$$

$$P(xy) + P(yx) = 3 [P(xx) + P(yy)] \tag{2}$$

Bei einer Sequenz aus “x” und “y” kann zudem die Anzahl der Wortpaare “xy” und “yx” höchstens um eins verschieden sein. Für eine lange Sequenz gilt deshalb:

$$P(xy) = P(yx) \tag{3}$$

Zudem gilt für sich gegenseitig ausschliessende Ereignisse e_i , die den gesamten Ereignisraum abdecken, dass $\sum_i P(e_i) = 1$ ist, also:

$$P(x) + P(y) = 1 \tag{4}$$

$$P(xx) + P(xy) + P(yx) + P(yy) = 1 \tag{5}$$

$$P(x|x) + P(y|x) = 1 \tag{6}$$

$$P(x|y) + P(y|y) = 1 \tag{7}$$

Ferner gelten aufgrund der Produktregel die folgenden Gleichungen:

$$P(xx) = P(x|x)P(x) \tag{8}$$

$$P(xy) = P(y|x) P(x) \quad (9)$$

$$P(yx) = P(x|y) P(y) \quad (10)$$

$$P(yy) = P(y|y) P(y) \quad (11)$$

Diese 11 Gleichungen enthalten 10 (unbekannte) Wahrscheinlichkeiten. Das Gleichungssystem ist jedoch nicht überbestimmt. Es ist einfach einzusehen, dass sich Gleichung (5) aus der Addition der Gleichungen (8) bis (11) ergibt.

Das Gleichungssystem kann von Hand oder mit der Matlab-Funktion `solve` (vergl. Matlab-Skript `ueb23_1b.m`) gelöst werden. Die Lösung lautet dann:

$$P(x) = \frac{3}{5}$$

$$P(y) = \frac{2}{5}$$

$$P(xx) = \frac{9}{40}$$

$$P(xy) = \frac{3}{8}$$

$$P(yx) = \frac{3}{8}$$

$$P(yy) = \frac{1}{40}$$

$$P(x|x) = \frac{3}{8}$$

$$P(x|y) = \frac{15}{16}$$

$$P(y|x) = \frac{5}{8}$$

$$P(y|y) = \frac{1}{16}$$

Die Sprache L_2 kann somit durch ein Bigram-Sprachmodell beschrieben werden.

Aufgabe 2: Perplexität einfacher Sprachen

- a) Die Perplexität ist die mittlere Wortverzweigungsrate. Da in der Sprache L_1 auf das Wort v_i jedes andere Wort mit derselben Wahrscheinlichkeit folgen kann, ist die Perplexität $\mathcal{Q} = |V| - 1 = 3$.

Selbstverständlich kann man die Perplexität auch berechnen, indem zuerst die Entropie $H(L_1)$ nach Formel (210) im Skript ermittelt wird:

$$\begin{aligned} H(L_1) &= - \sum_{i,j} P(v_i v_j) \log_2 P(v_j | v_i) \\ &= -12 P(v_i v_j) \log_2 P(v_j | v_i) \quad i \neq j \\ &= -12 P(v_j | v_i) P(v_i) \log_2 P(v_j | v_i) \\ &= -12 \frac{1}{|V|-1} \frac{1}{|V|} \log_2 \frac{1}{|V|-1} = -12 \frac{1}{3} \frac{1}{4} \log_2 \frac{1}{3} = \log_2 3 \end{aligned}$$

Damit ist die Perplexität der Sprache L_1 : $\mathcal{Q}(L_1) = 2^{H(L_1)} = 3$.

- b) Um die Perplexität von L_2 zu ermitteln, kann wiederum die Entropie berechnet werden:

$$\begin{aligned} H(L_2) &= - \sum_{i,j} P(v_i v_j) \log_2 P(v_j | v_i) \\ &= -P(xx) \log_2 P(x|x) - P(xy) \log_2 P(y|x) - P(yx) \log_2 P(x|y) - P(yy) \log_2 P(y|y) \\ &= 0.7076 \end{aligned}$$

Damit ist die Perplexität $\mathcal{Q}(L_2) = 2^{0.7076} = 1.6331$.

Aufgabe 3: Schätzen von N-Gram-Wahrscheinlichkeiten

Die Musterlösung der Funktion `ngram = estim_ngram_probs(tr_data, vocsize, N, smval, dbg)` ist im Directory `Uebung23/Loesung/` zu finden.

Aufgabe 4: Sprachidentifikation mittels N-Gram

Warum kann aufgrund der Kreuzentropie auf die Sprache geschlossen werden? Die Kreuzentropie gibt an, wie gut eine Buchstabenfolge zum Ngram passt. Sie passt umso besser, je kleiner der Wert der Kreuzentropie ist.

Erwartungsgemäss beschreiben Ngrams mit zunehmendem N die sprachspezifischen Eigenheiten von Buchstabenfolgen besser. Deshalb machen die Unigram-Modelle bei der Sprachidentifikation viele Fehler, während die Trigram-Modelle viel häufiger korrekt entscheiden. Dies trifft beispielsweise für das Wort “Sprachidentifikation” zu:

`ueb23_4('sprachidentifikation', 1, 1)` → Englisch

`ueb23_4('sprachidentifikation', 2, 1)` → Französisch

`ueb23_4('sprachidentifikation', 3, 1)` → Deutsch