Lösungen für 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.

Lösung zu 1)

1.

Wort	Anzahl
Auch	2
Horus	2
war	3
der	2
Sohn	1
von	1
Osiris	2
Seth	1
es	1
tötete	1
tot	1
$< s >$	3
$< /s >$	3

Korpusgröße $N = 23$. Vokabular $V = 13$

Wort 1 in der Zeile[2]

Wort	Horus	war	der	Sohn	von	Osiris	Seth	es	tötete	auch	tot	<s >	</s >
Horus	0	1	0	0	0	0	0	0	0	0	1	0	0
war	0	0	1	0	0	0	1	0	0	1	0	0	0
der	0	0	0	1	0	1	0	0	0	0	0	0	0
Sohn	0	0	0	0	1	0	0	0	0	0	0	0	0
von	0	0	0	0	0	1	0	0	0	0	0	0	0
Osiris	0	0	0	0	0	0	0	0	1	0	0	0	1
Seth	0	0	1	0	0	0	0	0	0	0	0	0	0
Es	0	1	0	0	0	0	0	0	0	0	0	0	0
tötete	0	0	0	0	0	0	0	0	0	0	0	0	1
auch	2	0	0	0	0	0	0	0	0	0	0	0	0
tot	0	0	0	0	0	0	0	0	0	0	0	0	1
<s >	0	1	0	0	0	0	0	1	0	1	0	0	0

2. $p = \frac{3}{23} \cdot \frac{1}{23} \cdot \frac{1}{23} \cdot \frac{2}{23} \cdot \frac{3}{23}$

3. $p = p(<s>)p(war|<s>)p(Seth|war)p(der|Seth)p(Sohn|der)p(von|Sohn)p(Osiris|von)p(</s>|Osiris) = \frac{3}{23} \cdot \frac{1}{3} \cdot \frac{1}{3} \cdot \frac{1}{1} \cdot \frac{1}{2} \cdot \frac{1}{1} \cdot \frac{1}{1} \cdot \frac{1}{2}$

   Eine Alternativversion (wie auch im Buch) ist $p(< s >)$ nicht zu berücksichtigen.

   Cross-entropy ergibt sich aus $-\frac{1}{8} \log p$ und Perplexität dann aus $2^{Cross-Entropie}$.

[2]Das Satzendzeichen wurde hier in der *Zeile* ignoriert, da darauf nichts folgt, und also dann vernachlässigt werden kann, wenn man nur Satzwahrscheinlichkeiten abschätzen will. In den Spalten ist es natürlich vorhanden. Sieht man stattdessen das ganze Korpus als aneinandergereihte Sätze, dann folgt auf jedes Satzendzeichen ein Satzanfangszeichen.

4. $P = p(<s>)p(toetete|<s>)p(Seth|toetete)p(auch|Seth)p(Horus|auch)p(</s>|Horus) = \frac{3}{23} \cdot \frac{0}{3} \cdot \frac{0}{1} \cdot \frac{0}{1} \cdot \frac{2}{2} \cdot \frac{0}{2} = 0$

   Wahrscheinlich hätte man auf größeren *ägyptologischen* Daten auch ein Auftreten der fehlenden Bigramme gesehen, da es sich im Beispiel um häufige Unigramme an sich handelt. Somit würde sich dann eine Wahrscheinlichkeit > 0 ergeben. Allerdings sind auch in großen Korpora immer noch viele theoretisch mögliche Bigramme Null und das Korpus müsste auch domänenspezifisch sein.

5. - $p_{+1} = p_{+1}(<s>) \cdot p_{+1}(toetete|<s>) \cdot p_{+1}(Seth|Toetete) \cdot p_{+1}(auch|Seth) \cdot p_{+1}(Horus|auch) \cdot p_{+1}(</s>|Horus)$
     $= \frac{3+1}{23+13} \cdot \frac{0+1}{3+13} \cdot \frac{0+1}{1+13} \cdot \frac{0+1}{1+13} \cdot \frac{2+1}{2+13} \cdot \frac{0+1}{2+13}$
   - mit $\lambda_1 = 0.5$ und $\lambda_2 = 0.5$ und ungesmoothter Unigrammwahrscheinlichkeit von $p(<s>)$:

     $p_{\text{Intpol}} = p(<s>)p_{intpol}(toetete|<s>) \cdot p_{intpol}(Seth|toetete) \cdot p_{intpol}(auch|Seth) \cdot p_{intpol}(Horus|auch) \cdot p_{intpol}(</s>|Horus)$
     $= \frac{3}{23} \cdot (0.5 \cdot \frac{0}{3} + 0.5 \cdot \frac{1}{23}) \cdot (0.5 \cdot \frac{0}{1} + 0.5 \cdot \frac{1}{23}) \cdot (0.5 \cdot \frac{0}{1} + 0.5 \cdot \frac{2}{23}) \cdot (0.5 \cdot \frac{2}{2} + 0.5 \cdot \frac{2}{23}) \cdot (0.5 \cdot \frac{0}{2} + 0.5 \cdot \frac{3}{23})$

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)

Lösung zu 2)

1. Personalpronomen *he,she*, da diese oft Subjekte sind. Andere richtige Antworten sind natürlich auch möglich.

2. Satzlängen. Sind Anfangs- und Endtoken häufig, hat das Korpus im Durchschnitt kurze Sätze.

3. Mehrere Beispiele von long-distance dependencies möglich: Verbendstellungen im Deutschen bei Fragesätzen (*hast du den neuen Film von Tarantino gestern gesehen*) zum Beispiel oder lange Unterbrechungen durch Relativsätze.

4. Sonst müssten wir nicht nur die eh schon oft seltenen Sequenzhäufigkeiten zählen, sondern auch noch Wahrscheinlichkeiten schätzen wie $p_{Satz-n}(hastDu)$. Damit würde sich das Data Sparseness Problem noch weiter verschärfen. Manchmal trifft dies jedoch nicht zu. Beispiel 1: In gewissen Genres ist es gut möglich, dass genrespezifisch gewisse Wortsequenzen am Anfang oder Ende eines Textes bedeutend wahrscheinlicher sind. So wäre das Trigramm *Mit freundlichen Grüßen* sicher im Genre von Briefen eher am Briefende zu erwarten als am Anfang oder in der Mitte. Dies berücksichtigt ein n-gram Modell nicht. (Vorsicht, dies hat nichts mit Satzanfängen oder enden zu tun, sondern mit Textpositionen). Beispiel 2: gewisse Wortarten sind auch in den ersten Sätzen eines Textes unwahrscheinlicher. So sind Pronomen *he,she* im ersten Satz weniger häufig, weil man sich ja mit Pronomen auf etwas vergangenes beziehen muss, wozu man am Textanfang noch nicht viel Spielraum hat.

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)

Lösung zu 3)

- Manche Buchstabenkombinationen sind am Wortende oder Anfang häufiger. Beispiel im Deutschen "ch" am Satzanfang seltener.
- Sprachen haben unterschiedliche Wortlängen. Die Häufigkeit des Leerzeichens enkodiert implizit auch Wortlänge.

Die Antworten sind in gewisser Weise äquivalent zu den entsprechenden Fragen für Wort n-gram-Modelle und deswegen niedrigbepunktet. Die Studierenden sollen hauptsächlich diese Äquivalenz erkennen.