You can use the this slide to practice for the exam.
I added suggested answers for some questions.

For the questions that do not have answer, try to find out the answer by your self.

If you have difficulty then we can solve it together in the Tutorium.

Erklären Sie wie ein Sprachidentifizierer funktioniert und wofür die Backoff-Glättung angewendet wird.

Erklären Sie wie ein Sprachidentifizierer (mit Markov-Model) funktioniert und wofür die Backoff-Glättung angewendet wird.

Ein **Sprachidentifizierer** (Language Guesser) bestimmt die wahrscheinlichste Sprache $\hat{L}$ eines gegebenen Textes T. Er berechnet:

$$\hat{L} = \arg\max_{L} p(L|T)$$

p(L|T) wird durch mehrere Umformungen und Annahme zu $p_L(T) = \prod_{i=1}^{n+1} p_L(a_i|a_{i-k}...a_{i-1})$

Falls wir p(a_i | a_i-k...a_i-1) mit relative Häufigkeit schätzen, die Häufigkeit von unbekanntes N-Gramm wird Null sein und dadurch the das Produkt von p(a_i | a_i-k...a_i-1) wird Null.
Die Backoff-Glättung wird zur Berechnung von p(a_i | a_i-k...a_i-1) verwendet und null Wahrscheinlichkeit zu vermeiden.

-Umgang mit unbekannte N-Gramme

Herleiten und Erklären jedes Schritt. Welche Annahme wurde gemacht?

$$p(L|T)$$

$$p_L(T) = \prod_{i=1}^{n+1} p_L(a_i|a_{i-k}...a_{i-1})$$

Aufgabe 1) Wie ist bei einem Markowmodell zweiter Ordnung (= Trigramm-Modell) über Buchstaben die Wahrscheinlichkeit einer Buchstabenfolge $a_1, a_2, ..., a_n$ allgemein definiert?
Wie ist die Wahrscheinlichkeit der Buchstabenfolge 'abc' konkret definiert? Was ist am Anfang und Ende der Buchstabenfolge zu beachten? (4 Punkte)

p(abc) = p(a|<s>,<s>) * p(b|<s>, a) * p(c|a, b,) *p(<s>|b, c)

<s>,<s>, a, b, c , <s>
3-gram
k=2
<s>,<s>**, a,**
<s>, a, **b**
a, b, **c**
b, c , **<s>**

Aufgabe 4) Wozu dient eine **Parameterglättung**? Welches Problem soll gelöst werden? (1 Punkt)

Aufgabe 2) Beschreiben Sie, wie ein Sprachidentifizierer mit Hilfe von solchen Markowmodellen implementiert werden kann. (3 Punkte)

Prinzip von Parameterglättung?

Nullwahrscheinlichkeiten müssen vermieden werden, außer das entsprechende n-Gramm ist unmöglich.

Prinzip der Parameterglättung: Man nimmt den beobachteten n-Grammen etwas Wahrscheinlichkeit weg und verteilt sie an die nicht beobachteten n-Gramme.

Grundidee von Backoff- Parameterglättung?

Aufgabe 4) Was ist die Grundidee der Backoff-Glättung? Geben Sie die zugehörige Formel zur Berechnung der geglätteten Wahrscheinlichkeiten mit einem interpolierten Backoff-Modell an. (3 Punkte)

Grundidee von Backoff- Parameterglättung?

Statt die Wahrscheinlichkeitsmasse gleichmäßig über alle unbeobachteten Wörter zu verteilen, werden Sie gemäß einer Backoff-Verteilung mit kleinerem Kontext verteilt.

Aufgabe 1) Geben Sie an, wie bei einem Markowmodell vierter Ordnung (= 5Gramm-Modell) die Wahrscheinlichkeit einer Buchstabenfolge $a_1, ..., a_n$ definiert ist. (3 Punkte)

Wie wird die Wahrscheinlichkeit p(w | context) ohne Glättung berechnet (Relative Häufigkeit Schätzung)?
- Schreib die Formel detailiert und erklären einzelnen Teilen

Wie wird die Wahrscheinlichkeit p(w | context) ohne Glättung berechnet (Relative Häufigkeit Schätzung)?
- Schreib die Formel und erklären einzelnen Teilen

$$p(w_n|w_1 \ldots w_{n-1}) = \frac{f(w_1 \ldots w_n)}{f(w_1 \ldots w_{n-1})} = \frac{f(w_1 \ldots w_n)}{\sum_w f(w_1 \ldots w_{n-1} w)}$$

oben: N-Gramm-Häufigkeit
unten: Kontext-Häufigkeit

Aufgabe 3) Wie lautet die Formel für die Berechnung der Wahrscheinlichkeit $p(a_5|a_1, ..., a_4)$, wenn Sie eine Maximum-Likelihood-Schätzung (= relative Häufigkeiten) verwenden? Welches Problem ergibt sich bei dieser Art der Parameterschätzung? (3 Punkte)

deine Antwort

Aufgabe 3) Wie wird die (ungeglättete) bedingte Wahrscheinlichkeit $p(w_3|w_1, w_2)$ aus Häufigkeiten geschätzt? (1 Punkt)

deine Antwort

Aufgabe 2) Wie wird die bedingte Wahrscheinlichkeit $p(w|w', w'')$ (ohne Glättung) geschätzt? Geben Sie die Formel an und erklären Sie kurz, was die verwendeten Ausdrücke/Variablen bedeuten. (1 Punkt)

deine Antwort

Aufgabe 4) Wie wird die bedingte Wahrscheinlichkeit $\hat{p}(w_3|w_1, w_2)$ mit interpolierter Backoff-Glättung und absolute Discounting geschätzt? Wie wird der Discount berechnet? Wie wird der Backofffaktor berechnet? (3 Punkte)

deine Antwort

Aufgabe 3) Wie lautet die Formel für die Berechnung geglätteter bedingter Wahrscheinlichkeiten nach dem normalen interpolierten Backoff-Verfahren? (2 Punkte)

$$p(w_i|w_{i-k}^{i-1}) = \frac{\max(0, f(w_{i-k}^{i}) - \delta_k)}{f(w_{i-k}^{i-1})} + \alpha(w_{i-k}^{i-1})\, p(w_i|w_{i-k+1}^{i-1})$$

Was bedeutet die einzelnen Komponenten ?

$p(w_i|w_{i-k}^{i-1})$:

$\frac{\max(0, f(w_{i-k}^{i}) - \delta_k)}{f(w_{i-k}^{i-1})}$

$\alpha(w_{i-k}^{i-1})$

$p(w_i|w_{i-k+1}^{i-1})$

Wie werden folgendes berechnet? (mit Formel)

$$\frac{, f(w_{i-k}^{i})}{f(w_{i-k}^{i-1})}$$

Für ngram

$$\frac{, f(w_{i-k}^{i})}{f(w_{i-k}^{i-1})}$$

Für **n-1** gram(in Backoff step)

$$\delta_k)$$

Was bedeutet k? Welche Information brauchen wir um den Discount zu berechnen?

$$\alpha(w_{i-k}^{i-1})$$

Wie werden folgendes berechnet? (mit Formel)

$$\frac{, f(w_{i-k}^{i})}{f(w_{i-k}^{i-1})}$$

Für ngram

$$f(w_{i-k}^{i-1}) = \sum_w f(w_{i-k}^{i-1} w)$$

$$\frac{, f(w_{i-k}^{i})}{f(w_{i-k}^{i-1})}$$

$$p_{backoff}(w|C) = \frac{f^*(C,w)}{\sum_{w'} f^*(C,w')}$$

$$f(C,w) = \sum_{w'} f(w', C, w)$$

$$f^*(C,w) = \sum_{w'} \mathbf{1}_{f(w',C,w)>0}$$

Für **n-1** gram(in Backoff step)

$$\delta_k)$$

$$\delta = \frac{N_1}{N_1 + 2N_2}$$

k ist die Ordnung vom Markov Model. Z.b. k=2 ist gleich ein 3-Gramm Markov Model.
k ist auch die Anzahl der Kontexte von einem N-Gramm.
Wir brauchen die Häufigkeiten von (k+1)-Gramme, um den Discount zu berechnen.

$$\alpha(w_{i-k}^{i-1})$$

$$\alpha(C) = 1 - \sum_{w:f(C,w)>0} \frac{f(C,w) - \delta}{f(C)}$$

$$p(w_i|w_{i-k}^{i-1}) = \frac{\max(0, f(w_{i-k}^{i}) - \delta_k)}{f(w_{i-k}^{i-1})} + \alpha(w_{i-k}^{i-1})\, p(w_i|w_{i-k\underline{+1}}^{i-1})$$

Geben Sie ein konkretes Beispiel für ein 3-Gramm-Model

$$p(w_i|w_{i-k}^{i-1}) = \frac{\max(0, f(w_{i-k}^{i}) - \delta_k)}{f(w_{i-k}^{i-1})} + \alpha(w_{i-k}^{i-1})\, p(w_i|w_{i-k\underline{+1}}^{i-1})$$

Geben Sie ein konkretes Beispiel für ein 2-Gramm-Model

$$p(w_i|w_{i-k}^{i-1}) = \frac{\max(0, f(w_{i-k}^{i}) - \delta_k)}{f(w_{i-k}^{i-1})} + \alpha(w_{i-k}^{i-1})\ p(w_i|w_{i-k+1}^{i-1})$$

p(w|context) ist ein Parameter vom Model, das wir trainieren wollen.
Im Training schätzen wir dieses aus einem Training-Korpus.
Erklären Sie die Schritte für die Schätzung aller Parameters.
Anfang mit einem Korpus an, was machen wir als nächstes?
Was muss zuerst berechnet werden und so weiter?

$$p(w_i|w_{i-k}^{i-1}) = \frac{\max(0, f(w_{i-k}^{i}) - \delta_k)}{f(w_{i-k}^{i-1})} + \alpha(w_{i-k}^{i-1})\ p(w_i|w_{i-k+1}^{i-1})$$

***Training** z.B. 3-Gramm-Model
- Korpus lesen, 3-Gramme und ihre Häufigkeiten extrahieren
    - Für jedes 3-Gramm berechne die relative Häufigkeiten mit Discount
- Erstelle 2-Gramm-Häufigkeiten (Backoff Teil)
    - Für jedes 2-Gramm berechne die relative Häufigkeiten mit Discount
- Erstelle 1-Gramm-Häufigkeiten (Backoff Teil)
    - Für jedes 1-Gramm berechne die relative Häufigkeiten mit Discount

Diese relative Häufigkeiten (p*) werden in einer Dictionary gespeichert
p* = {“abc”: 0.002, “bcd”: 0.001, …}

-Für die Berechnung des Backoff-Faktors brauchen wir diese relative Häufigkeiten von alle N-Gramme
    - Backoff (context) = 1 - (p*(context, any character) + p*(context, any character) + ….)
-Wir haben alle nögtige Komponenten für die Berechnung. Wir machen das gleiche für alle Sprache

***Model Anwenden**
- Dann können wir das Model auf einen neuen Text verwenden.
- Wir zeuge 3-Gramme aus dem Input-Text und wollen alle p_smoothed(3-Grammar) multiplizieren
- Für p_smoothed(3-gram), nrauchen wir p*(3-gram) + backoff(context) * (p_smoothed(2-gram)) --- recursive
- Wir berechnen diese Score für alle Sprachen, dann nehmen wir die Sprache mit dem höchsten Score

```
# Trainingstext aus Datei einlesen
with open(sys.argv[2]) as file:
    text = file.read()

# N-Gramm-Häufigkeiten berechnen
ngramfreq = defaultdict(int)
for i in range(len(text) - L+1):
    ngram = text[i:i+L]
    ngramfreq[ngram] += 1
```

text = “Mr President, I believe that the precautionary principle needs to be quite radical in order to work, otherwise we shall always be confused by all the different interpretations possible. I will give an example: a fungicide has been proven to cause babies to be born blind. It is therefore a teratogenic substance.”

```
ngramfreq = {("das", "rote", "buch"): 5,
             ("dieses", "rote", "buch"): 2,
             ("gute", "rote", "buch"): 4,
             ("das", "gelbe", "buch"): 1,
             ("das", "rote", "kleid"): 2,
             ("dieses", "rote", "kleid"): 2,
             ("das", "rote", "haus"): 8, }
```

$$p(w_i | w_{i-k}^{i-1})$$

Nachdem all Parameter berechnet wurden, wie werden sie verwendet, um die Sprache von einem Text zu bestimmen?

Gegeben sind die Buchstaben-N-Gramm-Häufigkeiten f(a,a)=1, f(a,b)=2, f(b,a)=0 und f(b,b)=1.

Berechnen Sie den Discount für die Glättung der Bigramm-Wahrscheinlichkeiten.

(1 Punkt)

Gegeben sind die Buchstaben-N-Gramm-Häufigkeiten f(a,a)=1, f(a,b)=2, f(b,a)=0 und f(b,b)=1.

Berechnen Sie den Discount für die Glättung der Bigramm-Wahrscheinlichkeiten.

(1 Punkt)

Discount Berechnung

$N_1 = 2$

$N_2 = 1$

$\delta = 2 / (2 + 2*1) = 0.5$

Gegeben sind die Buchstaben-N-Gramm-Häufigkeiten f(a,a)=1, f(a,b)=2, f(b,a)=0 und f(b,b)=1.

Berechnen Sie dann die Unigramm-Häufigkeiten f(a) und f(b) einmal nach dem Standard-Backoff-Verfahren und einmal nach dem Kneser-Ney-Verfahren. (1 Punkt)

Gegeben sind die Buchstaben-N-Gramm-Häufigkeiten f(a,a)=1, f(a,b)=2, f(b,a)=0 und f(b,b)=1.

Berechnen Sie dann die Unigramm-Häufigkeiten f(a) und f(b) einmal nach dem Standard-Backoff-Verfahren und einmal nach dem Kneser-Ney-Verfahren. (1 Punkt)

Berechnung der Backoff-Wahrscheinlichkeitsverteilung

Standard-Backoff-Verfahren

$f(a) = f(a,a) + f(b,a) = 1$

$f(b) = f(a,b) + f(b,b) = 3$

Kneser-Ney-Verfahren

$f^*(a) = |(a, a)| = 1$

$f^*(b) = |(a, b), (b, b)| = 2$

Aufgabe 6) Wozu dient **Parameter-Glättung**? Wie funktioniert eine Backoff-Glättung (mit Formel)? Wie unterscheidet sich die Kneser-Ney-Glättung von einer normalen Backoff-Glättung? (4 Punkte)

Gegeben sind die Buchstaben-N-Gramm-Häufigkeiten f(a,a)=1, f(a,b)=2, f(b,a)=0 und f(b,b)=1.

Berechnen Sie aus den Kneser-Ney-Häufigkeiten die ungeglätteten Unigramm-Wahrscheinlichkeiten p(a) und p(b). (1 Punkt)

Wie sieht die Formel von der ungeglätteten Wahrscheinlichkeit aus?

Gegeben sind die Buchstaben-N-Gramm-Häufigkeiten f(a,a)=1, f(a,b)=2, f(b,a)=0 und f(b,b)=1.

Berechnen Sie aus den Kneser-Ney-Häufigkeiten die ungeglätteten Unigramm-Wahrscheinlichkeiten p(a) und p(b). (1 Punkt)

$$\mathsf{p(a)} = f^*(\mathsf{a}) \;/\; (f^*(\mathsf{a}) + f^*(\mathsf{b})) = 1/3$$
$$\mathsf{p(b)} = f^*(\mathsf{b}) \;/\; (f^*(\mathsf{a}) + f^*(\mathsf{b})) = 2/3$$

$$p_{backoff}(w|C) = \frac{f^*(C,w)}{\sum_{w'} f^*(C,w')}$$

Gegeben sind die Buchstaben-N-Gramm-Häufigkeiten f(a,a)=1, f(a,b)=2, f(b,a)=0 und f(b,b)=1.

Berechnen Sie die relativen Häufigkeiten mit Discount (Standard-Häufigkeiten) für alle möglichen Bigramme p*(w1|w2)

Wie sieht die Formel aus?

Discount Berechnung

$N_1 = 2$

$N_2 = 1$

$\delta = 2 / (2 + 2*1) = 0.5$

Gegeben sind die Buchstaben-N-Gramm-Häufigkeiten f(a,a)=1, f(a,b)=2, f(b,a)=0 und f(b,b)=1.

Berechnen Sie die relativen Häufigkeiten mit Discount (Standard-Häufigkeiten) für alle möglichen Bigramme p*(w1|w2)

$$\frac{, f(w_{i-k}^{i}) - \delta_k)}{f(w_{i-k}^{i-1})}$$

Berechnung der relativen Häufigkeiten mit Discount (Standard)

r(a|a) = (f(a,a)-δ)/(f(a,a)+f(a,b)) = max(0,(1-0.5))/(1+2) = 1/6
r(b|a) = (f(a,b)-δ)/(f(a,a)+f(a,b)) = max(0,(2-0.5))/(1+2) = 3/6

r(a|b) = (f(b,a)-δ)/(f(b,a)+f(b,b)) = max(0,(0-0.5))/(0+1) = 0
r(b|b) = (f(b,b)-δ)/(f(b,a)+f(b,b)) = max(0,(1-0.5))/(0+1) = 0.5

Aufgabe 4) Eine geglättete Wahrscheinlichkeitsverteilung sei wie folgt definiert:

$$p(w|w') = p^*(w|w') + \alpha(w')p(w) \quad \text{mit } p^*(w|w') = \frac{f(w', w) - \delta}{\sum_{w'} f(w', w')}$$

Leiten Sie die Formel für die Berechnung des Backoff-Faktors $\alpha(w')$ her. (3 Punkte)

Backoff-Faktor für interpolierte Backoff-Glättung

Der Backoff-Faktor $\alpha(C)$ stellt sicher, dass die Wahrsch. zu 1 summieren:

$$\sum_w p(w|C) = \sum_w \frac{max(0, f(C,w) - \delta)}{f(C)} + \alpha(C)p(w|C') = 1$$

Wenn $C = w_1^{n-1}$ dann $C' = w_2^{n-1}$

Durch Umformen erhalten wir:

$$\sum_w \alpha(C)p(w|C') = 1 - \sum_w \frac{max(0, f(C,w) - \delta)}{f(C)}$$

Ausklammern von $\alpha(C)$ liefert:

$$\alpha(C) \underbrace{\sum_w p(w|C')}_{=1} = 1 - \sum_w \frac{max(0, f(C,w) - \delta)}{f(C)}$$

Dies ist äquivalent zu:

$$\alpha(C) = 1 - \sum_{w:f(C,w)>0} \frac{f(C,w) - \delta}{f(C)}$$

Aufgabe 3) Wie lautet die Formel zur Berechnung der Wahrscheinlichkeit $p(t_3|t_1, t_2)$, wenn Sie eine interpolierte Backoff-Schätzung mit Absolute Discounting verwenden?

(3 Punkte)

Aufgabe 6) Wie funktioniert die **Addiere-1**-Glättung (mit Formel) und warum ist sie nicht geeignet, um Wortwahrscheinlichkeiten zu glätten? Bei welchen Voraussetzungen ist die Addiere-1-Glättung optimal? (2 Punkte)

Aufgabe 2) In der Vorlesung wurde argumentiert, dass für einen Text T und eine Sprache L gilt:

$$\begin{aligned} \arg\max_L p(L|T) &= \arg\max_L \frac{p(T|L)p(L)}{p(T)} && (1) \\ &= \arg\max_L p(T|L)p(L) && (2) \\ &= \arg\max_L p(T|L) && (3) \end{aligned}$$

Geben Sie die Begründung für jeden der drei Schritte an. Bei welchem Schritt ist eine zusätzliche Annahme notwendig. Wie lautet sie? (3 Punkte)

Aufgabe 1) Leiten Sie die Formel für **Hidden-Markowmodelle** her, d.h. zeigen Sie, wie Sie von $\arg\max_{t_1^n} p(t_1^n|w_1^n)$ zu $\arg\max_{t_1^n} \prod_{i=1}^{n+1} p(t_i|t_{i-1})p(w_i|t_i)$ kommen.

Geben Sie an, welche vereinfachenden Annahmen Sie dabei machen.

Erklären Sie, warum der Index i in der Produktformel bis $n+1$ läuft. (5 Punkte)

Aufgabe 5) Wie ist bei einem Trigramm-Sprachmodell die Wahrscheinlichkeit der Wortfolge $w_1, ..., w_n$ definiert? Erklären Sie welche Besonderheiten am Satzanfang und Satzende zu beachten sind. (3 Punkte)