DATE:

NLP Assignment

1) Differentiate between interpolation and Backoff

Backoff

- Backoff N-gram modeling is a non-linear method
- We build on a N-gram model based upon (N-1) gram model
- Difference is that backoff, if we have non-zero trigram counts, we solely rely on trigram counts & dont interpolate the bigrams & unigram counts.
- Backoff model in trigram format

$$P(w_i \mid w_{i-2} w_{i-1}) = \begin{cases} \tilde{P}(w_i \mid w_{i-2} w_{i-1}) & \text{if } c(w_{i-2} w_{i-1} w_i) > 0 \\ \alpha(w_{n-2}^{n-1}) \tilde{P}(w_i \mid w_{i-1}) & \text{if } c(w_{i-2} w_{i-1} w_i) = 0 \\ \alpha(w_{n-1}) \cdot \tilde{P}(w_i) & \text{and } c(w_{i-2} w_{i-1} w_i) > 0 \end{cases}$$

- Does not yield valid probability distributions
- Works well for huge datasets

Interpolation

- This linear method combines different N grams by linear interpolating all 3 models wherever we are computing any trigram
- Here we dont train 3 λs as trigram grammar. Instead we make each λ a function of the content.
- λ terms are used to decide how much to smooth
- $\sum_i \lambda_i = 1$

Mathematically,

$$\tilde{P}(w_0 \mid w_2 w_{-1}) = \lambda_3 \cdot P(w_0 \mid w_{-2} w_{-1}) + \lambda_2 P(w_0 \mid w_{-1}) + \lambda_1 \cdot P(w_0)$$

- Can interpolate 'customized' models with 'general' models

Q.2 Viterbi's Algorithm is a variation of the forward algorithm which considers all words simultaneously in order to compute the most likely path

Algorithm

→ Input: Observation of length T, state-graph of length N

→ Output: best-path

for-each state s from 1 to N to

$q[1, s] \leftarrow P(s|s_0) \cdot P(o_1|s)$

backpointers $[1, s] \leftarrow 0$

for each timestep t from 2 to T do

for each state s from 1 to N do

$q[t|s] \leftarrow \max_{s'=1}^{N} (q[t-1, s'] \cdot P(s|s') \cdot P(o_t|s))$

backpointers $[t, s] \leftarrow \operatorname{argmax}_{s'=1}^{N} q[t-1, s'] \cdot P(s|s')$

$s \leftarrow \operatorname{argmax}_{s'=1}^{N} q[T, s]$

return backtrace path from backpointers $[T, s]$

Example

Consider a two word language: "fish" & "sleep"

where,

- "fish" appears 8 times as noun and 5 as verb
- "sleep" : appears twice as noun & 5 times as verb

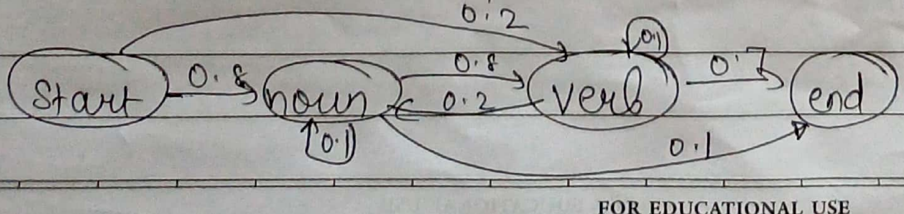
Emission Probabilities

Noun - $P(\text{fish}|\text{noun}) = 0.8$

- $P(\text{sleep}|\text{noun}) = 0.2$

Verb $P(\text{fish}|\text{verb}) = 0.5$

$P(\text{sleep}|\text{verb}) = 0.5$



Token 1 'fish'

	0	1	2	3
start	1	0		
verb	0	0.2*0.5		
noun	0	0.8*0.8		
end	0	0		

Token 2 sleep (if fish is verb)

	0	1	2	3
start	1	0	0	
verb	0	0.1	(0.64×0×0.5, 0.1×0.1×0.5) max	
noun	0	0.64	(0.64×0.1×0.2, 0.1×0.2×0.2) max	
end	0	0		

Token 3 end

	0	1	2	3
start	1	0	0	
verb	0	0.1	0.256	
noun	0	0.64	0.128	
end	0	0		

(0.256×0.7, 0.128×0.1) max

Now we can backtrace through most likely path

Q.3 Corpus

<s> I am from DJ </s>
<s> I am a teacher </s>
<s> All students are good & intelligent </s>
<s> students from DJ score high marks </s>

Test data

<s> students are from DJ </s>

Unigram	<s>	students	are	from	DJ	</s>
	4	2	1	2	2	4

Bigram
First we find occurrence count

	<s>	students	are	from	DJ	</s>
<s>	0	1	0	0	0	0
students	0	0	1	1	0	0
are	0	0	0	0	0	0
from	0	0	0	0	2	0
DJ	0	0	0	0	0	1
</s>	0	0	0	0	0	0

Bigram

	<s>	students	are	from	DJ	</s>
<s>	0	1/4	0	0	0	0
student	0	0	1/2	1/2	0	0
are	0	0	0	0	0	0
from	0	0	0	0	2/2 = 1	0
DJ	0	0	0	0	0	1/2
</s>	0	0	0	0	0	0

Using MLE to estimate probability

$P = P(\text{students} \mid \langle s\rangle) * P(\text{are} \mid \text{students})\ P(\text{from} \mid \text{are}) * P(\text{DJ} \mid \text{are}) * P(\langle /s\rangle \mid \text{DJ})$

$= 1/4 * 1/2 * 0 * 1 * 1/2 = 0$ ← apply laplace smoothening

Before applying Laplace smoothening we

V = count of unique vocab

= Count ({<s>, </s>, I, am, from, DJ, a, teacher, all, students, are, good, and, intelligent, score, high, marks}) = 17

$\therefore P = \left(\frac{1+1}{4+17}\right) * \left(\frac{1+1}{2+17}\right) * \left(\frac{0+1}{1+17}\right) * \left(\frac{2+1}{2+17}\right) + \left(\frac{1+1}{2+17}\right)$

$$= \frac{2}{21} \times \frac{1}{19} \times \frac{1}{18} \times \frac{3}{19} \times \frac{2}{19}$$

$$= 9.257 \times 10^{-6}$$

<s> I am Sam </e>

<s> Sam I am </e>

<s> I donot like green eggs & Ham</e>

Calculate bigram probability for

1. P(am | sam)
2. P (do | I)
3. P(am | I)

$P(w|w_{n-1})$

$P(\langle s\rangle, I),\ P(I, am),\ P(am, sam),\ P(sam, a)$

$$\frac{C(w_{n-1}, w_n)}{C(w_{n-1})}$$

(1) $$\frac{C(am, sam)}{C(sam)} = P(am|sam)$$

$$= \cancel{\frac{1}{2}} = 0.5$$

$$\boxed{P(am|sam) = 0.5}$$

(2) $$P(do|I) = \frac{C(I, do)}{C(I)} = \frac{1}{3}$$

(3) $$P(am|I) = \frac{C(I, am)}{C(I)}$$

$$= \frac{2}{3}$$

Q. Calculate the trigram Probability of I am Sam

$P(\text{sam} \mid \text{I am}) = \dfrac{C(\text{I am sam})}{C(\text{I am})}$

$= \dfrac{1}{2}$

Q. Calculate the MLE for "I am Sam" using bigram

MLE → (<s> I), (I, am), (am, sam), (sam, </e>)

$P(I \mid \langle s \rangle) * P(am \mid I) \times P(sam \mid am) * P(\langle e \rangle \mid sam)$

$\frac{2}{3} \times \frac{2}{3} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{9}$

Q. <s> John read Moby Dick <e>

<s> Mary Read a different book <e>

<s> She read a book by Cher <e>

a) Calculate MLE for John read a book using bigram

b) " " " Cher read a book

a) $\frac{1}{3} \times \frac{1}{1} \times \frac{2}{3} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{18}$

$P(John \mid \langle s \rangle) \times P(read \mid John) \times P(a \mid read) \times P(book \mid a) \times P(\langle e \rangle \mid book)$

b) $P(Cher \mid \langle s \rangle) \times P(read \mid Cher) \times P(a \mid read) \times P(book \mid a) \times P(\langle e \rangle \mid book)$

Using Add-On Smoothing

P(cher read a book

$= 0/3 \times 0/1 + 2/3 \times \frac{1}{2} \times 1/2$

$$= \frac{0+1}{11+3} \times \frac{0+1}{11+1} \times \frac{2+1}{11+3} \times \frac{1+1}{11+2} \times \frac{1+1}{11+2}$$

$= 0.00003$