

CSE556: NLPQuiz - 2 - Solutions

Answers are in red color.

1. "Smoothing always incurs some information loss to the original counts". Justify or refute. [3]

Smoothing is a mechanism to prevent assigning zero probabilities to unknown or out-of-vocabulary words. We borrow some probability mass from observed/seen words (bigrams/trigrams, etc.) and distribute it among OOVs. Since probability value is directly proportional to the word count, reducing the probability will reduce the counts as well.

2. Let $V = \{a, b, w\}$, Corpus: $\{[a\ b\ b\ a\ a\ w\ a\ b\ w\ b], [w\ a\ a\ a\ b\ a]\}$. Utilizing the trigram LM, find the next predicted word in the sequence "w b a ____". Use **Interpolation** as the smoothing technique, in case you're getting zero co-occurrence counts. Take appropriate alpha values. **Show computation.** [15]

Q2] Given, $V = \{a, b, w\}$
 Corpus = $\{[a\ b\ b\ a\ a\ w\ a\ b\ w\ b], [w\ a\ a\ a\ b\ a]\}$

We have to predict the next word in "w b a ____"

So using trigram

$\max(P(x/ba))$ where $x = \{a, b, w\}$
 will give us the next probable word.

$P(x/ba) = \frac{\text{count}(bax)}{\text{count}(ba)}$

trigram
 Co-occurrence (Bigram)

	w	a	b
ab	1	1	1
ba	0	1	0
aw	0	1	0
wa	0	1	1
bw	0	0	1
wb	0	0	0

since we get 0 for some columns of ba we will get 0 probability \therefore we have to use interpolation smoothing.

So for that

Co-occurrence Bigram				Unigram Co-occurrence			
	w	a	b		w	a	b
w	0	2	1		3	8	5
a	1	3	3				
b	1	2	1				

Total token count in corpus = 16

$P(w) = \frac{3}{16}$ $P(a) = \frac{8}{16}$

$P(b) = \frac{5}{16}$

Considering $x = "a"$ and $\alpha_1 = 0.5, \alpha_2 = 0.25, \alpha_3 = 0.25$

$P(a/ba) = \alpha_1 P(a/ba) + \alpha_2 P(a/a) + \alpha_3 P(a)$

$= 0.5 \left(\frac{1}{2}\right) + 0.25 \left(\frac{5}{8}\right) + 0.25 \left(\frac{8}{16}\right)$

$= 0.25 + 0.093 + 0.125$

$= 0.468$

Now $x = "b"$

$P(b/ba) = \alpha_1 P(b/ba) + \alpha_2 P(b/a) + \alpha_3 P(b)$

$= 0.5 (0) + 0.25 \left(\frac{3}{8}\right) + 0.25 \left(\frac{5}{16}\right)$

$= 0 + 0.093 + 0.0781$

$= 0.1711$

Now $x = "w"$

$P(w/ba) = \alpha_1 P(w/ba) + \alpha_2 P(w/a) + \alpha_3 P(w)$

$= 0.5 (0) + 0.25 \left(\frac{1}{8}\right) + 0.25 \left(\frac{3}{16}\right)$

$= 0 + 0.03125 + 0.0468$

$= 0.07805$

$\max(P(a/ba), P(b/ba), P(w/ba))$

$\max(0.468, 0.1711, 0.07805)$

hence the most probable next word is "a"

3. Write shorthand orthographic rules (similar to the equation below) for Y replacement -- "y changes to -ie before -s, -i before -ed". [4]

$$FST_i: \varepsilon \rightarrow e / \left\{ \begin{matrix} x \\ s \\ z \end{matrix} \right\} \wedge \text{---} s\#$$

[No partial marking.]

$$FST: y \rightarrow ie/\{e\}^{\wedge} - s\#$$

$$FST: y \rightarrow i/\{e\}^{\wedge} - ed\#$$

4. Write the **lexical** and **intermediate** forms for the surface forms **Boys**, **Cities**, and **Fish**. Use conventional notations. [8]

Surface \Rightarrow Intermediate (4 marks) \Rightarrow Lexical (4 marks)

Boys \Rightarrow Boy $^s \# \Rightarrow$ Boy +N +pl

Cities \Rightarrow City $^s \# \Rightarrow$ City +N +pl

Fish \Rightarrow Fish $\# \Rightarrow$ Fish +N +sg

Fish \Rightarrow Fish $\# \Rightarrow$ Fish +N +pl