Chapter 3 : N-gram Language Models

Exercises

(§ 3.1) $P(w_n \mid w_{n-2} w_{n-1}) = \dfrac{C(w_{n-2} w_{n-1} w_n)}{C(w_{n-2} w_{n-1})}$

<s><s> I am Sam </s>
<s><s> Sam I am </s>
<s><s> I do not like green eggs and ham </s>

$P(\text{I} \mid \text{<s><s>}) = 2/3$
$P(\text{am} \mid \text{<s> I}) = 1/2$

(§ 3.2) $P(\text{i want chinese food}) = P(\text{i} \mid \text{<s>})\, P(\text{want} \mid \text{i})\, P(\text{chinese} \mid \text{want})\, P(\text{food} \mid \text{chinese})\, P(\text{</s>} \mid \text{food})$

$= 0.25 \times .33 \times 0.0065 \times .52 \times 0.68$

$= 0.0001896$

$P(\text{i want chinese food}) = P(\text{i} \mid \text{<s>})\, P(\text{want} \mid \text{i})\, P(\text{chinese} \mid \text{want})\, P(\text{food} \mid \text{chinese})\, P(\text{</s>} \mid \text{food})$

$= .19 \times 0.21 \times 0.0029 \times 0.052 \times .4$

$= 0.000002406$

(§ 3.3) The unsmoothed probability is higher because the bigrams used in the sentences are very common and has probabilities. However, in the smoothed case, their probabilities are distributed among not-so-common bigrams which are not used in our test statement.

(Q3.4)

	<s>	I	am	Sam	do	not	like	green	eggs	and	</s>
<s>	1	4	1	2	1	1	1	1	1	1	1
I	1	1	4	1	2	1	1	1	1	1	1
am	1	1	1	3	1	1	1	1	1	1	2
Sam	1	2	1	1	1	1	1	1	1	1	4
do	"	1	"	"	"	2	"	"	"	"	1
not	"	"	"	"	"	1	2	"	"	"	"
like	"	"	"	"	"	"	1	2	"	"	"
green	"	"	"	"	"	"	"	1	2	"	"
eggs	"	"	"	"	"	"	"	"	1	2	"
and	"	"	"	2	"	"	"	"	"	1	"
</s>	"	"	"	1	"	"	"	"	"	"	"

$$P(\text{Sam} \mid \text{am}) = \frac{C(\text{am Sam})}{C(\text{am})} = \frac{3}{14} = .214$$

(Q3.5)

	<s>	a	b
<s>	0	2	2
a	0	1	1
b	0	1	1

- $P(aa) = P(a \mid <s>)\, P(a \mid a) = 0.5 \times 0.5 = 0.25$
  $P(ab) = P(a \mid <s>)\, P(b \mid a) = 0.5 \times 0.5 = 0.25$
  $P(bb) = P(b \mid <s>)\, P(b \mid b) = 0.5 \times 0.5 = 0.25$
  $P(ba) = P(b \mid <s>)\, P(a \mid b) = 0.5 \times 0.5 = 0.25$

$$P(s \in \{a,b\}^2) = 1$$

(Q 3.6) $P(w_3 | w_1 w_2) = \dfrac{C(w_1 w_2 w_3) + 1}{\sum_{\omega} (C(w_1 w_2 \omega) + 1)}$

$= \dfrac{C(w_1 w_2 w_3) + 1}{C(w_1 w_2) + v}$

(Q 3.7) $= P(\text{sam} | \text{am}) = \lambda_1 P(\text{sam}) + \lambda_2 (\text{sam} | \text{am})$

$= \frac{1}{2} \times \frac{42}{25} + \frac{1}{2} \times \frac{2}{3}$

$= \frac{2}{25} + \frac{1}{3} = 0.41$

(Q 3.12) $PP(w) = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_i)}}$

$P(0) = \frac{91}{100}$

$P(1) = P(2) \ldots P(9) = \frac{1}{100}$

$= \sqrt[10]{\frac{(100)^{10}}{(91)^{9}}}$

$= 1.726$