Einführung in die Computerlinguistik
Hausaufgabe Language Models, Abgabe 18.05.2020

Laura Kallmeyer

SoSe 2020, Heinrich-Heine-Universität Düsseldorf

Aufgabe 1 *Consider a toy example where the sentences are built over* $\{a, b\}$ *and our training data consists of the following sentences:*

<s> *a a b a a b b* </s>
<s> *a b a b b a a b a a* </s>
<s> *b b a a b a b a b a* </s>
<s> *b b a b a b* </s>

1. *Give the following bigram probabilities and their* $\log_2$ *values:*
   $P(b|\texttt{<s>})$ $P(b|b)$ $P(\texttt{</s>}|b)$
2. *Using these values, the probability of a sequence* <s>$w_1 w_2 \dots w_n$ *amounts to*

$$P(\texttt{<s>}w_1 w_2 \dots w_n) \quad = \quad 2^{\sum_{k=1}^{n} \log_2 P(w_k|w_{k-1})}$$

   while the perplexity value of this sequence is

$$PP(\texttt{<s>}w_1 w_2 \dots w_n) \quad = \quad \sqrt[n]{\frac{1}{2^{\sum_{k=1}^{n} \log_2 P(w_k|w_{k-1})}}} = \sqrt[n]{2^{\sum_{k=1}^{n} |\log_2 P(w_k|w_{k-1})|}}$$

   Compute both, probability and perplexity, of the following sequences:

   (1) <s> *b* </s>

   (2) <s> *b b b* </s>

Solution:

1. $P(b|\texttt{<s>}) = \frac{1}{2}$, $\log_2 \frac{1}{2} = -1$ $P(b|b) = \frac{4}{16}$, $\log_2 \frac{1}{4} = -2$ $P(\texttt{</s>}|b) = \frac{1}{8}$, $\log_2 \frac{1}{8} = -3$
2. (1): probability $2^{-1-3} = 2^{-4} = \frac{1}{2^4} = \frac{1}{16}$, perplexity $\sqrt[2]{2^4} = 4$
   (2): probability $2^{-1-2-2-3} = 2^{-8} = \frac{1}{2^8} = \frac{1}{612}$, perplexity $\sqrt[4]{2^8} = 2^2 = 4$

Aufgabe 2 *Assume that we want to obtain a bigram language model with Laplace Smoothing. Our training data are the following two sentences:*

<s> *b a a b a* </s>
<s> *a b a b a b a b b* </s>

(The size of the vocabulary is 3 since we have three possible second elements of our bigrams, a, b, and </s>*.)*

This time, we compute using the probabilities, not their log values.

1. *What are the following probabilities according to this model?*
   (a) $P(a|\texttt{<s>})$ *(b)* $P(b|a)$ *(b)* $P(a|a)$ *(b)* $P(\texttt{</s>}|b)$
2. *Give probability and perplexity of the following input sentence:*

<s> *a a b* </s>

Solution:

1. (a) $P(a|\texttt{<s>}) = \frac{2}{5}$ (b) $P(b|a) = \frac{6}{10} = \frac{3}{5}$ (c) $P(a|a) = \frac{2}{10} = \frac{1}{5}$
   (d) $P(\texttt{</s>}|b) = \frac{2}{10} = \frac{1}{5}$
2. Probability: $\frac{2}{5} \cdot \frac{1}{5} \cdot \frac{3}{5} \cdot \frac{1}{5} = \frac{6}{625}$

   Perplexity: $\sqrt[4]{\frac{625}{6}}$

Aufgabe 3 *Consider again the same training data as in the preceding question.*

1. *Modify the training data adopting a treatment of unknown words that changes the first occurrence of every word into* <UNK>.

   Give the following bigram probabilities estimated by this model:

   $P(a|\texttt{<UNK>})$ $P(\texttt{<UNK>}|a)$ $P(\texttt{<UNK>}|\texttt{<s>})$ $P(a|\texttt{<s>})$
2. *Calculate the probability of the two sequences*

   (3) *a.* <s> *a c*
   b. <s> *c a*

   according to this model.
3. *Now add Lapace Smoothing. What are the new values of our four bigram probabilities and the probabilities of (3-a) and (3-b) in this new model?*

Solution:

1. Modified training data:

   <s> <UNK> <UNK> a b a </s>
   <s> a b a b a b a b b </s>

   $P(a|\texttt{<UNK>}) = \frac{1}{2}$ $P(\texttt{<UNK>}|a) = 0$ $P(\texttt{<UNK>}|\texttt{<s>}) = \frac{1}{2}$ $P(a|\texttt{<s>}) = \frac{1}{2}$
2. $P(\texttt{<s>}ac) = \frac{1}{2} \cdot 0 = 0$ $P(\texttt{<s>}ca) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$
3. With additional Laplace Smoothing:

   $P(a|\texttt{<UNK>}) = \frac{2}{6}$ $P(\texttt{<UNK>}|a) = \frac{1}{10}$ $P(\texttt{<UNK>}|\texttt{<s>}) = \frac{2}{6}$ $P(a|\texttt{<s>}) = \frac{2}{6}$

   $P(\texttt{<s>}ac) = \frac{1}{3} \cdot \frac{1}{10} = \frac{1}{30}$ $P(\texttt{<s>}ca) = \frac{1}{3} \cdot \frac{1}{3} = \frac{1}{9}$