Einführung in die Computerlinguistik
Hausaufgabe Language Models, Abgabe 07.05.2019

Laura Kallmeyer

SoSe 2019, Heinrich-Heine-Universität Düsseldorf

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

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

1. *Give the following bigram probabilities and their* $\log_2$ *values:*
   $P(a|\texttt{<s>})$ $P(a|a)$ $P(\texttt{</s>}|a)$
2. *Using these values, the probability of a sequence* $\texttt{<s>}w_1w_2\dots w_n$ *amounts to*

$$P(\texttt{<s>}w_1w_2\dots w_n) \quad = \quad 2^{\sum_{k=1}^{n}\log_2 P(w_k|w_{k-1})}$$

   while the perplexity value of this sequence is

$$PP(\texttt{<s>}w_1w_2\dots w_n) \quad = \quad \sqrt[n]{\frac{1}{2^{\sum_{k=1}^{n}\log_2 P(w_k|w_{k-1})}}} = \sqrt[n]{2^{\sum_{k=1}^{n}|\log_2 P(w_k|w_{k-1})|}}$$

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

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

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

3. *Explain why the perplexity value is better for measuring the quality of (1) and (2).*

Solution:

1. $P(a|\texttt{<s>}) = \frac{1}{2}$, $\log_2\frac{1}{2} = -1$ $P(a|a) = \frac{1}{4}$, $\log_2\frac{1}{4} = -2$ $P(\texttt{</s>}|a) = \frac{1}{8}$, $\log_2\frac{1}{8} = -3$
2. (1): probability $2^{-1-3} = 2^{-4} = \frac{1}{2^4} = \frac{1}{16}$, perplexity $\sqrt[2]{2^4} = 4$
   (2): probability $2^{-1-2-2-2-3} = 2^{-10} = \frac{1}{2^{10}} = \frac{1}{1024}$, perplexity $\sqrt[5]{2^{10}} = 2^2 = 4$
3. The probability of a sequence is not a good candidate for measuring its quality since the longer a sequence the lower its probability, i.e., shorter sequences are preferred. This is why (1) has a much higher probability than (2).

   The perplexity value does not have this problem since for a sentence of length n, we have an exponent $\frac{1}{n}$, which balances the multiplication of n probabilities ($\sqrt[n]{x^n} = x$ for all n).

Aufgabe 2 *Consider again the same training data.*

1. *Modify the training data adopting a treatment of unknown words that changes the first occurrence of every word (here first occurrence of a, b, c) into* <UNK>.

   Give the following bigram probabilities estimated by this model:
   $P(a|a)$ $P(b|a)$ $P(c|a)$ $P(\texttt{<UNK>}|a)$ $P(a|\texttt{<s>})$

2. *Calculate the probability of the sequence*

   (3) `<s>` *a a d*

   according to this model.

Solution:

1. Modified training data:

   `<s>` `<UNK>` a `<UNK>` `<UNK>` a b c `</s>`
   `<s>` a b a b b a a b b a a `</s>`
   `<s>` b b a a b a b a b a `</s>`
   `<s>` b b c a b a b `</s>`

   $P(a|a) = \frac{3}{15} = \frac{1}{5}$ $P(b|a) = \frac{9}{15} = \frac{3}{5}$ $P(c|a) = \frac{0}{15} = 0$ $P(\texttt{<UNK>}|a) = \frac{1}{15}$ $P(a|\texttt{<s>}) = \frac{1}{4}$

2. $P(\texttt{<s>}aad) = \frac{1}{4} \cdot \frac{1}{5} \cdot \frac{1}{15}$

Aufgabe 3 *Consider again the training data from the first exercise.*

This time, we assume that treating unknown words is not necessary. But we do Laplace smoothing in order to cope with unseen n-grams.

Note that our vocabulary of words that can be the second element of a bigram includes the end of sentence marker `</s>`*, i.e.,* $V = \{a, b, c, \texttt{</s>}\}$.

1. *Calculate the new values we obtain for*

   $P(a|\texttt{<s>})$ $P(a|a)$ $P(c|a)$ $P(a|c)$ $P(\texttt{</s>}|a)$

2. *Calculate the probabilities of the following sequences using this new model:*

   (4) `<s>` *a c a* `</s>`

   (5) `<s>` *a a c a c a* `</s>`

Solution:

1. $P(a|\texttt{<s>}) = \frac{3}{8}$ $P(a|a) = \frac{5}{20} = \frac{1}{4}$ $P(c|a) = \frac{1}{20}$ $P(a|c) = \frac{3}{7}$ $P(\texttt{</s>}|a) = \frac{3}{20}$
2. $P((4)) = \frac{3}{8} \cdot \frac{1}{20} \cdot \frac{3}{7} \cdot \frac{3}{20}$
   $P((5)) = \frac{3}{8} \cdot \frac{1}{4} \cdot \frac{1}{20} \cdot \frac{3}{7} \cdot \frac{1}{20} \cdot \frac{3}{7} \cdot \frac{3}{20}$