Einführung in die Computerlinguistik
Hausaufgabe Language Models, Abgabe 10.05.2022 vor der Vorlesung

Laura Kallmeyer

SoSe 2022, Heinrich-Heine-Universität Düsseldorf

Aufgabe 1 *Consider the following toy example (similar to the one from Jurafsky and Martin):*

Training data:

```
<s> Sam I like </s>
<s> I do like Sam </s>
<s> Sam does like Sam </s>
<s> I like Sam </s>
```

Assume that we use a bigram language model based on the above training data.

1. *Give all the bigram probabilities one obtains from that data.*
2. *What is the most probable next word predicted by the model for the following word sequences (end of sentence* `</s>` *is also a possibility)?*

   (1) `<s>` *I like ...*

   (2) `<s>` *Sam does ...*

3. *Compute the probabilities for the following sentences according to the bigram language model. Furthermore, compute the perplexity, if possible. In case the perplexity value is not defined, explain why this is so.*

   (3) `<s>` *Sam I do like* `</s>`

   (4) `<s>` *does Sam like Sam* `</s>`

Solution:

1. Bigram probabilities:
   $P(\texttt{Sam}|\texttt{<s>}) = \frac{2}{4} = \frac{1}{2}$ $P(\texttt{I}|\texttt{<s>}) = \frac{1}{2}$
   $P(\texttt{like}|\texttt{I}) = \frac{2}{3}$ $P(\texttt{do}|\texttt{I}) = \frac{1}{3}$
   $P(\texttt{I}|\texttt{Sam}) = \frac{1}{5}$ $P(\texttt{</s>}|\texttt{Sam}) = \frac{3}{5}$ $P(\texttt{does}|\texttt{Sam}) = \frac{1}{5}$
   $P(\texttt{</s>}|\texttt{like}) = \frac{1}{4}$ $P(\texttt{Sam}|\texttt{like}) = \frac{3}{4}$
   $P(\texttt{like}|\texttt{do}) = \frac{1}{1} = 1$
   $P(\texttt{like}|\texttt{does}) = \frac{1}{1} = 1$
   All other bigram probabilities are 0.
2. (1): `Sam`
   (2): `like`
3. Probabilities:
   (3): `<s>` Sam I do like `</s>`
   Probability: $\frac{1}{2} \cdot \frac{1}{5} \cdot \frac{1}{3} \cdot 1 \cdot \frac{1}{4} = \frac{1}{120}$
   Perplexity: $\sqrt[5]{120}$

   (4): `<s>` does Sam like Sam `</s>`
   0 (since $P(\texttt{does}|\texttt{<s>}) = P(\texttt{Sam}|\texttt{does}) = P(\texttt{like}|\texttt{Sam}) = 0$) (unseen bigram)
   Perplexity is not defined, since we cannot divide by 0.

Aufgabe 2 *Consider again the training data from the preceding question:*

<s> Sam I like </s>
<s> I do like Sam </s>
<s> Sam does like Sam </s>
<s> I like Sam </s>

Assume that we use a bigram language model with
1. an unknown word treatment that consists of replacing any occurrences of the two low frequency tokens do *and* does *with* <UNK>*, and*
2. Laplace smoothing.

1. *Compute the following bigram probabilities in this new model:*
   $P(\texttt{I}|\texttt{<s>})$ $P(\texttt{like}|\texttt{I})$ $P(\texttt{<UNK>}|\texttt{I})$ $P(\texttt{like}|\texttt{<UNK>})$ $P(\texttt{</s>}|\texttt{like})$
2. *Compute probablility and perplexity of the following sentences with this new model (*don't *is taken to be a single token):*

   (5) <s> I like </s>

   (6) <s> I do like </s>

   (7) <s> I don't like </s>

Solution:

Manipulated training data:

<s> Sam I like </s> <s> I <UNK> like Sam </s> <s> Sam <UNK> like Sam </s> <s> I like Sam </s>

1. Bigram probabilities (vocabulary size 5 since we are including </s>):
   $P(\texttt{I}|\texttt{<s>}) = \frac{2+1}{5+4} = \frac{1}{3}$ $P(\texttt{like}|\texttt{I}) = \frac{2+1}{5+3} = \frac{3}{8}$ $P(\texttt{<UNK>}|\texttt{I}) = \frac{1+1}{5+3} = \frac{1}{4}$
   $P(\texttt{like}|\texttt{<UNK>}) = \frac{2+1}{5+2} = \frac{3}{7}$ $P(\texttt{</s>}|\texttt{like}) = \frac{1+1}{5+4} = \frac{2}{9}$
2. Probabilities:

   (5): <s> I like </s>
   Probability: $\frac{1}{3} \cdot \frac{3}{8} \cdot \frac{2}{9} = \frac{1}{36}$
   Perplexity: $\sqrt[3]{36} \approx 3,30$

   (6): <s> I do like </s>
   Probability: $\frac{1}{3} \cdot \frac{1}{4} \cdot \frac{3}{7} \cdot \frac{2}{9} = \frac{1}{126}$
   Perplexity: $\sqrt[4]{126} \approx 3,35$

   (7): <s> I don't like </s>
   Probability and perplexity as for (6).