Einführung in die Computerlinguistik
Hausaufgabe 4, Abgabe 14.05.2018

Laura Kallmeyer

SoSe 2018, Heinrich-Heine-Universität Düsseldorf

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

Training data:

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

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

1. *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>` *Sam ...*

   (2) `<s>` *Sam I do ...*

   (3) `<s>` *I am Sam I ...*

   (4) `<s>` *do I like ...*

2. *Compute the probabilities for the following sentences according to the bigram language model.*

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

   (6) `<s>` *Sam I am* `</s>`

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

Solution:

Bigram probabilities:

$P(\texttt{Sam}|\texttt{<s>}) = \frac{2}{5}$ $P(\texttt{I}|\texttt{<s>}) = \frac{2}{5}$ $P(\texttt{do}|\texttt{<s>}) = \frac{1}{5}$
$P(\texttt{am}|\texttt{I}) = \frac{1}{5}$ $P(\texttt{like}|\texttt{I}) = \frac{2}{5}$ $P(\texttt{do}|\texttt{I}) = \frac{2}{5}$
$P(\texttt{I}|\texttt{Sam}) = \frac{2}{5}$ $P(\texttt{</s>}|\texttt{Sam}) = \frac{3}{5}$
$P(\texttt{Sam}|\texttt{am}) = 1$
$P(\texttt{</s>}|\texttt{like}) = \frac{1}{2}$ $P(\texttt{Sam}|\texttt{like}) = \frac{1}{2}$
$P(\texttt{like}|\texttt{do}) = \frac{2}{3}$ $P(\texttt{I}|\texttt{do}) = \frac{1}{3}$

All other bigram probabilities are 0.

1. (1): `</s>`
   (2): `like`
   (3): `do` and `like` are equally probable
   (4): `</s>` and `Sam` are equally probable
2. Probabilities:
   (5): $\frac{2}{5} \cdot \frac{2}{5} \cdot \frac{2}{5} \cdot \frac{1}{3} \cdot \frac{2}{5} \cdot \frac{1}{2}$

(6): $\frac{2}{5} \cdot \frac{2}{5} \cdot \frac{1}{5} \cdot 0 = 0$
(7): $\frac{2}{5} \cdot \frac{2}{5} \cdot \frac{2}{3} \cdot \frac{1}{2} \cdot \frac{2}{5} \cdot \frac{2}{5} \cdot \frac{1}{2}$

Aufgabe 2 *Now consider a toy example where the sentences are built over $\{a, b\}$ and we have only the following two sentences in our training data:*

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

1. *Compute the probabilities for the following sentences according to the bigram language model that results from these training data.*

   (8) `<s>` *a* `</s>`

   (9) `<s>` *a a* `</s>`

   (10) `<s>` *a a b b a b b a* `</s>`

2. *Which is in general (not only among (8)–(10)) the most probable sentence among all possible sequences of as and bs according to the model? Explain your answer.*

3. *Comment on the differences between the probabilities in 1. Do you see shortcomings of this model?*

Solution:

Bigram probabilities (only those $\neq 0$ are listed):

$P(a\|\texttt{<s>}) = \frac{2}{2} = 1$	$P(b\|\texttt{<s>}) = \frac{0}{2} = 0$	
$P(a\|a) = \frac{4}{11}$	$P(b\|a) = \frac{5}{11}$	$P(\texttt{</s>}\|a) = \frac{2}{11}$
$P(a\|b) = \frac{5}{9}$	$P(b\|b) = \frac{4}{9}$	$P(\texttt{</s>}\|b) = \frac{0}{9} = 0$

1. (8): $1 \cdot \frac{2}{11} = \frac{2}{11} \approx 0.18$
   (9): $1 \cdot \frac{4}{11} \cdot \frac{2}{11} = \frac{8}{11^2} \approx 0.066$
   (10): $1 \cdot \frac{4}{11} \cdot \frac{5}{11} \cdot \frac{4}{9} \cdot \frac{5}{9} \cdot \frac{5}{11} \cdot \frac{4}{9} \cdot \frac{5}{9} \cdot \frac{2}{11} = \frac{80.000}{11^4 \cdot 9^4} \approx 0.00083$

2. Sentence (8) is the most probable sentence in general since (i) any sentence has to start with a, otherwise the probability becomes 0, (ii) any sentence has to end with a as well, otherwise the probability is 0, and (iii) the longer the sentence, the lower the probability since every further symbol adds a factor p with $0 \leq p < 1$.

3. Sentence (10) is the only sentence we have seen during training. However, it receives a very low probability. Shortcomings of this model are

   (i) that shorter sentences are preferred no matter what the average sentence length in the training corpus is;

   (ii) the fact that a sentence belongs to the training corpus and is therefore probably better than other sentences of a similar length that are not in the training set is not taken into account;

   (iii) the model makes strong independence assumptions by using only bigrams, more complex structures are not taken into consideration, for instance the pattern $a^n b^n w w^R$ ($w^R = w$ in reverse order) that we can see in the two training sentences.