Example of bigram language model evaluation

Let’s have a short trainset of sentences.

Ala ma kota.
Ala chce kota.
Asia woli psa.
Ala chce psa.
Asia woli kota.
Asia chce psa.

Now let’s count words occurrences. Punctuation marks are treated as words.

6 .
3 psa
3 kota
3 chce
3 Asia
3 Ala
2 woli
1 ma

Words “ma” and “woli” occurred less than 3 times so they are substituted with <UNK> token. The signs <s> and </s> are added. In the end the formatted train file looks like:

<s> Ala <UNK> kota . </s>
<s> Ala chce kota . </s>
<s> Asia <UNK> psa . </s>
<s> Ala chce psa . </s>
<s> Asia <UNK> kota . </s>
<s> Asia chce psa . </s>

The vocabulary looks like following:

.
psa
kota
chce
Asia
Ala
<UNK>
<s>
</s>

There are 3 <UNK> tokens and there are 24 words altogether. To calculate OOV rate we also count six ending marks </s> so in total the OOV rate of training set equals 3/30 = 10%.

Now let's count probabilities for bigram model.

	.	<UNK>	Ala	Asia	chce	kota	psa	</s>
.	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	1.00000
<UNK>	0.00000	0.00000	0.00000	0.00000	0.00000	0.66667	0.33333	0.00000
Ala	0.00000	0.33333	0.00000	0.00000	0.66667	0.00000	0.00000	0.00000
Asia	0.00000	0.66667	0.00000	0.00000	0.33333	0.00000	0.00000	0.00000
chce	0.00000	0.00000	0.00000	0.00000	0.00000	0.33333	0.66667	0.00000
kota	1.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
psa	1.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
UNK	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
<s>	0.00000	0.00000	0.50000	0.50000	0.00000	0.00000	0.00000	0.00000

So here for instance after beginning of sequence <s> three times there is a word "Ala" and three times word "Asia", hence probabilities P(Ala | <s>) = 0.5 and P(Asia | <s>) = 0.5

After Laplace smoothing:

	.	<UNK>	Ala	Asia	chce	kota	psa	</s>
.	0.06667	0.06667	0.06667	0.06667	0.06667	0.06667	0.06667	0.46667
<UNK>	0.08333	0.08333	0.08333	0.08333	0.08333	0.25000	0.16667	0.08333
Ala	0.08333	0.16667	0.08333	0.08333	0.25000	0.08333	0.08333	0.08333
Asia	0.08333	0.25000	0.08333	0.08333	0.16667	0.08333	0.08333	0.08333
chce	0.08333	0.08333	0.08333	0.08333	0.08333	0.16667	0.25000	0.08333
kota	0.33333	0.08333	0.08333	0.08333	0.08333	0.08333	0.08333	0.08333
psa	0.33333	0.08333	0.08333	0.08333	0.08333	0.08333	0.08333	0.08333
<s>	0.06667	0.06667	0.26667	0.26667	0.06667	0.06667	0.06667	0.06667

Let’s define test corpus, on which the perplexity will be measured:

Asia woli kota.
Ania chce kota.

Word “Ania” didn’t occur in the trainset at all so it is mapped to <UNK>. The word “woli” occurred twice in trainset so it is also mapped to <UNK>. After formatting it looks like following:

<s> Asia <UNK> kota . </s>
<s> <UNK> chce kota . </s>

The total number of words in the testset are 8, but we calculate also two endings of the sentence, which gives us altogether 10 words. There are two <UNK> words so the OOV rate is 1/5.

The perplexity of the testset equals

$$\sqrt[10]{\frac{1}{P(Asia|<s>) * P(<UNK>|Asia) * P(kota|<UNK>) * P(.|kota) * P(</s>|.) * P(<UNK>|<s>) * P(chce|<UNK>) * P(kota|chce) * P(.|kota) * P(</s>|.)}}$$

which after putting values from the table equals 4.393194