An NLP Tip

javad.pourmostafa

October 2019

1 Perplexity Example

Task: Apply add-one smoothing to the trigram language model trained on the sentence:

"This is the cat that killed the rat that ate the malt that lay in the house that Jack built."

Find the perplexity of this smoothed model on the test sentence:

"This is the house that Jack built."

Solution:

Tip1: We have n=3, so we will add two start tokens $< s1 >$, $< s2 >$ and one end token $< end >$.

Tip2: Note, that we add (n-1) start tokens, since the start tokens are needed to condition the probability of the first word on them. The role of the end token is different and we always add just one end token. It's needed to be able to finish the sentence in the generative process at some point.
So, what we have is:

train: $< s1 >< s2 >$ This is the cat that killed the rat that ate the malt that lay in the house that Jack built $< end >$
test: $< s1 >< s2 >$ This is the house that Jack built $< end >$

Number of unique unigrams in train is 14, so V = 14 + 1 = 15.
Number of words in the test sentence is 7, so N = 7.
$Perplexity = p(w_{test})^{-\frac{1}{N}}$,
where
$p(w_{test}) = \prod_{i=1}^{8} p(w_i|w_{i-2}w_{i-1}) = \prod_{i=1}^{8} \frac{c(w_{i-2}w_{i-1}w_i)+1}{c(w_{i-2}w_{i-1}w_i)+15}$

All right, now we need to compute 8 conditional probabilities. We can do it straightforwardly or notice a few things to make our life easier.

First, note that all bigrams from the test sentence occur in the train sentence exactly once, which means we have (1 + 15) in all denominators.

Also note, that "is the house" is the only trigram from the test sentence that is not present in the train sentence. The corresponding probability is

$p(house|isthe) = (0 + 1)/(1 + 15) = 0.0625.$

All other trigrams from the test sentence occur in the train sentence exactly once. So their conditional probabilities will be equal to

$(1 + 1)/(1 + 15) = 0.125.$

In this way, perplexity is $(0.0625 * 0.125\ ^{7})^{(} - 1/7) = 11.89.$