An NLP Tip

javad.pourmostafa

October 2019

1 end-symbol magic

Without loss of generality, let's consider a bigram model (looking at two words at a time), without a beginning or end marker. Let's also assume our language has at least one sentence of length two, at least one sentence of length three, and at least one sentence of length four.

What's the probability of seeing the sentence "I am", with nothing before or after it? In our model, it would be $P(\text{"}I\text{"}) \times P(\text{"}am\text{"}|\text{"}I\text{"})$: the odds of seeing "I", times the odds of seeing "am" when you've already seen "I".

Now, what's the probability of seeing any sentence of length two in this model? We can see it should be $\sum_v \sum_w P(v) \times P(w|v)$: same as above, but summing over all possible words instead of just "I" and "am".

But this is the same as $\sum_v P(v) \sum_w P(w|v) = \sum_v P(v) \times 1$, by laws of probability. And we know that $\sum_v P(v) = 1$. Therefore the probability of having a sentence of length two is 1.

And this is a problem. Because we know that there are also sentences that are not length two. You can repeat this proof with length three, and find that the probability of having a sentence of length three is also 1. And same for four, and so on...

But we want the probability of seeing a sentence to be 1, because that's how probability works, which means the probability of seeing a sentence of length two (and the probability of seeing a sentence with length three, and...) has to be less than 1.

It turns out, adding a sentence-end marker fixes this. Because now when we're taking that sum over v and w, we can specify that $w = \text{" } < /s > \text{"}$, and the sum becomes $\sum_v P(v, \text{" } < /s > \text{"})$, which is significantly less than 1. I won't prove it here, but you can go through some tedious probability math to show that adding beginning and ending markers makes the probability distribution valid (the probability of seeing a sentence at all is now 1).1