Language model evaluation

- When dealing with BOW features, our input dimensions were fixed $(M \times N)$, giving us a direct means to evaluate our model on any given example i using the NLL:

$$NLL(i^{th}\text{ example}) = -\log P(y^{(i)} | \mathbf{x}^{(i)})$$

- In language modeling, $y^{(i)}$ represents the last word in a length T sequence, $\mathbf{x}^{(T)_i}$, while $\mathbf{x}^{(i)}$ is replaced by the prefix of that word: $\{\mathbf{x}^{(1)_i} \ldots \mathbf{x}^{(T-1)_i}\}$. In this case our interpretation of log likelihood is muddled by the fact that, all else being equal, longer sequences yield lower NLL:

$$NLL(i^{th}\text{ example}) = -\sum_{t=1}^{T} \log P(\mathbf{x}^{(t)_i} | \mathbf{x}^{(t-n)_i}, \ldots, \mathbf{x}^{(t-1)_i})$$

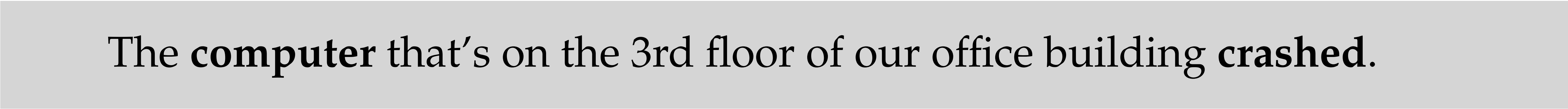
- *Perplexity offers a useful* evaluation metric for LMs (lower is better):

$$PPL(\mathbf{X}^{(i)}) = \exp\left\{ -\frac{1}{T} \sum_{t=1}^{T} \log P\left(\mathbf{x}^{(t)_i} | \mathbf{x}^{(t-n)_i} \ldots \mathbf{x}^{(t-1)_i} \right) \right\}$$

n-gram modeling has a fundamental limitation

- In lecture-03 we learned that word2vec was motivated by this idea of capturing the meaning of words within their contexts. We didn’t formalize the learning problem as one of language modeling, but in fact it is a language model that uses the skip-gram assumptions (Markov assumption + context word ordering doesn’t matter).
- But, are we really solving the problem? We’ve learned that *n* is practically limited to around 10 or so; even at n=10 we have a sparsity problem and are forced to introduce bias in order to even compute the probability of sequences. We can hardly say that we’re capturing context needed for understanding. For example, any reasonable model of human language should be able to capture the intended meaning of “*crashed”* in the following sentence:

Examples from Eisenstein (2018)



The **computer** that’s on the 3rd floor of our office building **crashed**.

- In practice, computing *n*-gram probabilities from occurrence frequency is adequate for some tasks, and far from adequate for others.