Finishing Information Theory

CSCI 384: Computational Linguistics

CSCI 384

October 13, 2017

$$
\begin{aligned}
\textit{perplexity} &= \left(\prod_{i=1}^{K} P(w_i \mid h)\right)^{\frac{-1}{N}} \\
&= \sqrt[N]{\frac{1}{\prod_{i=1}^{K} P(w_i \mid h)}} \\
&\approx \sqrt[n]{\frac{1}{q(x_{1\ldots n})}} && \text{adjusting notation} \\
&= \left(\frac{1}{q(x_{1\ldots n})}\right)^{\frac{1}{n}} \\
&= 2^{\lg\left(\left(\frac{1}{q(x_{1\ldots n})}\right)^{\frac{1}{n}}\right)} \\
&= 2^{\frac{1}{n}\cdot \lg \frac{1}{q(x_{1\ldots n})}} \\
&= 2^{H(X,Y)}
\end{aligned}
$$

Perplexity and cross entropy

We suspect that speech recognition people prefer to report on the larger non-logarithmic numbers given by perplexity mainly because it is much easier to impress funding bodies by saying that "we've managed to reduce preplexity from 950 to only 540" than by saying that "we've reduced cross entropy from 9.9 to 9.1 bits." However, perplexity does also have an intuitive reading: a perplexity of k means that you are as surprised on average as you would have been if you had had to guess between k equiprobable choices at each step.
Manning and Schütze, *Foundations of Statistical Natural Language Processing*, pg 78.

Wheel of Fortune

smscs.com

Lacunae in ancient manuscripts

MS 2650
Bible: Matthew. Egypt, 1st half of 4th c.
Unique text of the Gospel. 8 chapters are the earliest known of this part of the Bible

www.hds.harvard.edu

Comments from Brown et al.

> *Our bound is higher than previous entropy estimates, but it is statistically more reliable since it is based on a much larger test sample. Previous estimates were necessarily based on very small samples since they relied on human subjects to predict characters. Quite apart from any issue of statistical significance, however, it is probable that people predict English text better than the simple model that we have employed here.*

Brown et al, "Estimating the Entropy of English"

Comments from Brown et al.

> *We can also think of our cross-entropy as a measure of the compressibililty of the data in the Brown Corpus. The ASCII code for the characters in the Brown Corpus . . . [can be] reduced to 7 bits per character. With a simple Huffman code, which allots bits so that common characters get short bit strings at the expense of rare characters, we can reach 4.46 bits per character.*

Brown et al, "Estimating the Entropy of English"

Comments from Brown et al.

> *From a loftier perspective, we cannot help but notice that linguistically the trigram concept, which is the workhorse of our language model, seems almost moronic. It captures local tactic constraints by sheer force of numbers but the more well protected bastions of semantic, pragmatic, and discourse constraint and even morphological and global syntactic constraint remain unscathed, in fact unnoticed. Surely the extensive work on these topics in recent years can be harnessed to predict English better than we have yet predicted it. We see this paper as a gauntlet thrown down before the computational linguistics community. . . . We hope by proposing this standard task to unleash a fury of competitive energy that will gradually corral the wild and unruly thing that we know the English language to be.*

Brown et al, "Estimating the Entropy of English"