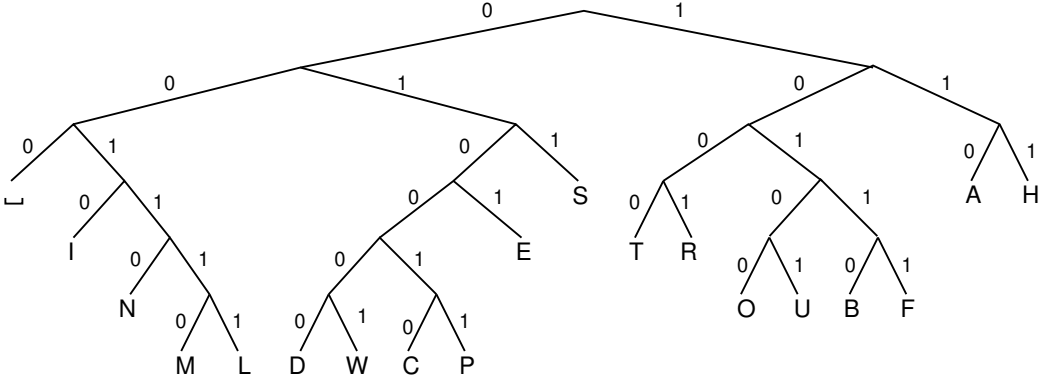
Entish

'Hoom, hmm! Come now! Not so hasty! You call yourselves hobbits? But you should not go telling just anybody. You'll be letting out your own right names if you're not careful.'
'We aren't careful about that,' said Merry. 'As a matter of fact I'm a Brandybuck, Meriadoc Brandybuck, though most people call me just Merry.'
'And I'm a Took, Peregrin Took, but I'm generally called Pippin, or even Pip.'
'Hm, but you are hasty folk, I see,' said Treebeard. 'I am honoured by your confidence; but you should not be too free all at once. There are Ents and Ents, you know; or there are Ents and things that look like Ents but ain't, as you might say. I'll call you Merry and Pippin if you please— nice names. For I am not going to tell you my name, not yet at any rate.' A queer half-knowing, half-humorous look came with a green flicker into his eyes. 'For one thing it would take a long while: my name is growing all the time, and I've lived a very long, long time; so my name is like a story. Real names tell you the story of the things they belong to in my language, in the Old Entish as you might say. It is a lovely language, but it takes a very long time to say anything in it, because we do not say anything in it, unless it is worth taking a long time to say, and to listen to. Tolkien, *TLotR III.4*

English/Spanish instructions

␣	000	E	0101	M	001110	S	011
A	110	F	10111	N	00110	T	1001
B	10110	H	111	O	10100	U	10101
C	010010	I	0010	P	010011	W	010000
D	010001	L	001111	R	1000		



Example

From Thomas Cover and Thomas Joy, *Elements of Information Theory*, John Wiley and Sons Inc, 1991. Pg 44.

The World Series is a seven-game series that terminates as soon as either team wins four games. Let X be the random variable that represents the outcome of a World Series between teams A and B; possible values of X are $AAAA$, $BABABAB$, $BBBAAAA$. Let Y be the number of games played, which ranges from 4 to 7. Assuming that A and B are equally matched and that the games are independent, calculate $H(X)$, $H(Y)$, $H(X|Y)$, and $H(Y|X)$.

The meaning of entropy

The word *entropy* had of course been used before Shannon. In 1864 Rudolf Clausius introduced the term. . . to represent a "transformation" that always accompanies a conversion between thermal and mechanical energy. . . .
[One of the authors] asked Shannon what he had thought about when he had finally confirmed his famous measure. Shannon replied: "My greatest concern was what to call it. I thought of calling it 'information,' but that word was overly used, so I decided to call it 'uncertainty.' When I discussed it with John von Neumann, he had a better idea. Von Neumann told me, 'You should call it *entropy*, for two reasons. In first place your uncertainty function has been used in statistical mechanics under that name, so it already has a name. In the second place, and more important, no one knows what entropy is, so in a debate you will always have the advantage.' "
Tribus and McIrvine, "Energy and Information", *Scientific American* # 224, Sept 1971, pg 178–184

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

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"