Entropy of Spoken English - Shannon

Autor:
Data de publicació: 29-08-2021

Entropy of Spoken English

Information, what is it really? Claude Shannon, the creator of the field of Information theory, described information as a reduction in uncertainty.

Is language, both spoken and written, information? Yes.

How do we measure information? Entropy. Entropy is our surprise about information.

Claude Shannon brought entropy to Information theory. He loved Entropy so much. In fact, he called his house the Entropy House.

Language can be described using entropy. Some things we say more often, and they don't capture as much surprise.

Low entropy: "The", "And", "There", "Password"

High entropy: "supercalifragilisticexpialidocious", "xylophone"

For example, you probably wouldn't want to use a low entropy word to be your password…

Claude Shannon saw that language was information, in fact, he saw that all language was statistical in nature, saying: "anyone speaking a language possesses, implicitly, an enormous knowledge of the statistics of the language. Familiarity with the words, idioms, cliches, and grammar enables him to fill in missing or incorrect letters in proof-reading, or to complete an unfinished phrase in conversation."

See how well you can use your statistical understanding of english by filling in the blanks of this phrase:

_____ Shannon was the invento_ of entrop_, he is my her_, I want to study informati_ theory for the ____ of my life.

Claude Shannon also did this in his foundational paper: Prediction and Entropy of Printed English.

In his paper, Claude outlined it as below. The first line is the original text; the second line contains a dash for each letter that was correctly guessed using the statistical nature of english.

Shannon even built his own communication system for encoding (reducing) and decoding (predicting) messages using statistics.

So, how did we arrive at English, or any language at all, as a species?

We arrive at English by making restrictive rules. There is a a restricted vocabulary and grammar to our language. We

need to be predictable, and we need to be less surprising (lower entropy).

Shannon once said, “Why doesn’t anyone say XFOML RXKHRJFFJUJ?”

What is “free” communication in English? And are we ever not restricted? How random and surprising (high entropy) can your English be?

Claude Shannon was a code breaker during World War 2, so he knew better than anyone, that you can break languages into pieces, and look at the frequencies (inverse of surprise) for words, pairs of letters, letters, or any other reduction of language.

But how do you bring all words/letters into one common scale?

After you capture the entropy, you can convert it into bits.

What is a bit? It can either be 1 or 0. Heads or Tails.

All information in the universe can be reduced to bits. Silicon Valley made its billions from bits. And Shannon invented the bit (with some help from his friends at Bell Labs).

At the heart of Information theory, is the mathematical communication of information.

Here is Shannon’s famous diagram:

Before Claude Shannon, “everyone thought that communication was involved in trying to find ways of communication written language, spoken language, pictures, video, and all of these different things— that all of these would require different ways of communicating. Claude said no, you can turn all of them into binary digits. And then you can find ways of communicating the binary digits.” -Shannon’s colleague and friend Robert Gallager