Formal Models of Language

Paula Buttery

Dept of Computer Science & Technology, University of Cambridge

Languages transmit **information**

In previous lectures we have thought about language in terms of **computation**.

Today we are going to discuss language in terms of the **information** it conveys...

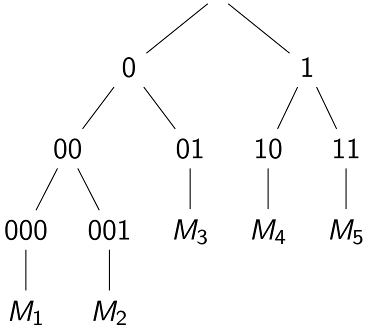
Entropy is a measure of information

- Information **sources** produce **information** as **events** or **messages**.
- Represented by a random variable X over a discrete set of symbols (or alphabet) $\mathcal{X}$.
- e.g. for a dice roll $\mathcal{X} = \{1, 2, 3, 4, 5, 6\}$ for a source that produces characters of written English $\mathcal{X} = \{a...z, \}$
- **Entropy** (or **self-information**) may be thought of as:
  - the average amount of information produced by a source
  - the average amount of uncertainty of a random variable
  - the average amount of information we gain when receiving a message from a source
  - the average amount of information we lack before receiving the message
  - the average amount of uncertainty we have in a message we are about to receive

Entropy is a measure of information

- Entropy, H, is measured in **bits**.
- If X has M equally likely events: $H(X) = \log_2 M$
- Entropy gives us a **lower limit** on:
  - the number of bits we need to represent an event space.
  - the average number of bits you need per message code.

$$\begin{aligned} avg\_length &= \frac{(3*2)+(2*3)}{5} = 2.4 \\ &> H(5) = \log_2 5 = 2.32 \end{aligned}$$



Surprisal is also measured in bits

- Let $p(x)$ be the probability mass function of a random variable, X over a discrete set of symbols $\mathcal{X}$.
- The **surprisal** of x is $s(x) = \log_2 \left(\frac{1}{p(x)}\right) = -\log_2 p(x)$
- Surprisal is also measured in **bits**
- Surprisal gives us a measure of information that is inversely proportional to the probability of an event/message occurring
- i.e probable events convey a small amount of information and improbable events a large amount of information
- The average information (entropy) produced by X is the weighted sum of the surprisal (the average surprise): $H(X) = -\sum_{x \in \mathcal{X}} p(x) \log_2 p(x)$
- Note, that when all M items in $\mathcal{X}$ are equally likely (i.e. $p(x) = \frac{1}{M}$) then $H(X) = -\log_2 p(x) = \log_2 M$

The surprisal of the alphabet in *Alice in Wonderland*

x	$f(x)$	$p(x)$	$s(x)$
	26378	0.197	2.33
e	13568	0.101	3.30
t	10686	0.080	3.65
a	8787	0.066	3.93
o	8142	0.056	4.04
i	7508	0.055	4.16
…			
v	845	0.006	7.31
q	209	0.002	9.32
x	148	0.001	9.83
j	146	0.001	9.84
z	78	0.001	10.75

- If uniformly distributed: $H(X) = \log_2 27 = 4.75$
- As distributed in *Alice*: $H(X) = 4.05$
- Re. example 1:
- Average surprisal of a vowel = 4.16 bits (3.86 without u)
- Average surprisal of a consonant = 6.03 bits

Example 1

Last consonant removed:

Jus the he hea struc agains te roo o te hal: i fac se wa no rathe moe tha nie fee hig.

average missing information: 4.59 bits

Last vowel removed:

Jst thn hr hed strck aganst th rof f th hll: n fct sh ws nw rathr mor thn nin fet hgh.

average missing information: 3.85 bits

Original sentence:

Just then her head struck against the roof of the hall: in fact she was now rather more than nine feet high.

The surprisal of words in *Alice in Wonderland*

x	$f(x)$	$p(x)$	$s(x)$
the	1643	0.062	4.02
and	872	0.033	4.94
to	729	0.027	5.19
a	632	0.024	5.40
she	541	0.020	5.62
it	530	0.020	5.65
of	514	0.019	5.70
said	462	0.017	5.85
i	410	0.015	6.02
alice	386	0.014	6.11
...			
<any>	3	0.000	13.2
<any>	2	0.000	13.7
<any>	1	0.000	14.7

Example 2

She stretched herself up on tiptoe, and peeped over the edge of the mushroom, and her eyes immediately met those of a large blue caterpillar, that was sitting on the top with its arms folded, quietly smoking a long hookah, and taking not the smallest notice of her or of anything else.

Average information of *of* $= 5.7$ bits

Average information of low frequency compulsory content words $=$
14.7 bits (freq $= 1$), 13.7 bits (freq $= 2$), 13.2 bits (freq $= 3$)

Aside: Is written English a good code?

Highly efficient codes make use of regularities in the messages from the source using shorter codes for more probable messages.

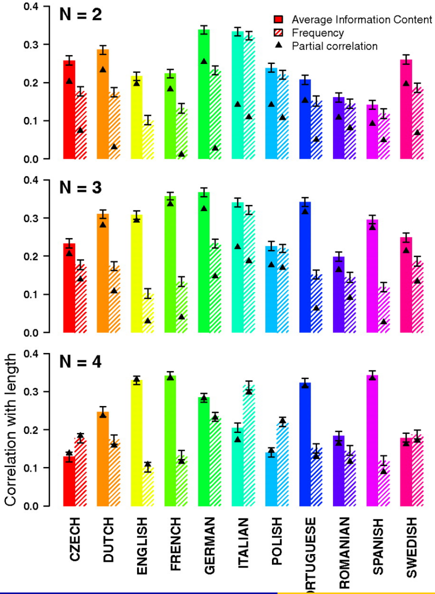
- From an encoding point of view, surprisal gives an indication of the number of bits we would want to assign a message symbol.
- It is efficient to give probable items (with low surprisal) a small bit code because we have to transmit them often.
- So, is English efficiently encoded?
- Can we predict the information provided by a word from its length?

Aside: Is written English a good code?

Piantadosi et al. investigated whether the surprisal of a word correlates with the word length.

- They calculated the average surprisal (average information) of a word w given its context c
- That is, $-\frac{1}{C}\sum_{i=1}^{C}\log_2 p(w|c_i)$
- Context is approximated by the n previous words.

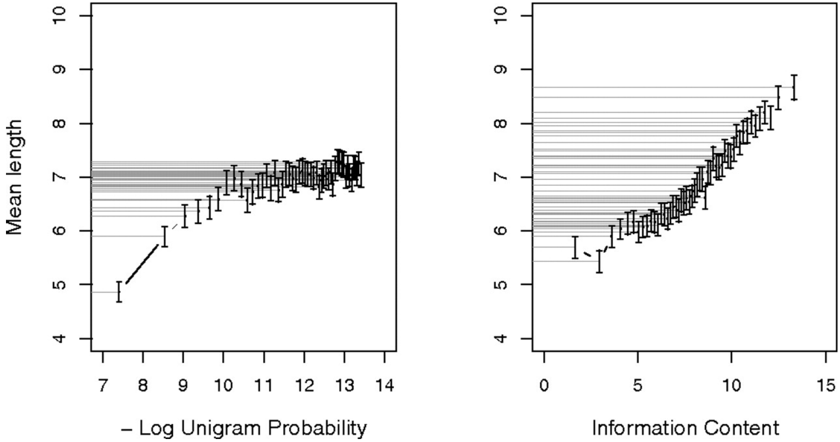
Aside: Is written English a good code?



- Piantadosi et al. results for Google n-gram corpus.
- Spearman's rank on y-axis (0=no correlation, 1=monotonically related)
- Context approximated in terms of 2, 3 or 4-grams (i.e. 1, 2, or 3 previous words)
- Average information is a better predictor than frequency most of the time.

Aside: Is written English a good code?

Piantadosi et al: Relationship between frequency (negative log unigram probability) and length, and information content and length.



In language, events depend on context

Examples from *Alice in Wonderland*:

- Generated using $p(x)$ for $x \in \{a\text{-}z, \_\}$:

  dgnt_a_hi_tio__iui_shsnghihp_tceboi_c_ietl_ntwe_c_a_ad__ne_saa
  __hhpr___bre_c_ige_duvtnltueyi_tt__doe

- Generated using $p(x|y)$ for $x, y \in \{a\text{-}z, \_\}$:

  s_ilo_user_wa_le_anembe_t_anceasoke_ghed_mino_fftheak_ise_linld_met
  _thi_wallay_f_belle_y belde_se_ce

In language, events depend on context

Examples from *Alice in Wonderland*:

- Generated using $p(x)$ for $x \in \{\textit{words in Alice}\}$:

  didnt and and hatter out no read leading the time it two down to just this must goes getting poor understand all came them think that fancying them before this

- Generated using $p(x|y)$ for $x, y \in \{\textit{words in Alice}\}$:

  murder to sea i dont be on spreading out of little animals that they saw mine doesnt like being broken glass there was in which and giving it after that

In language, events depend on context

- **Joint entropy** is the amount of information needed on average to specify two discrete random variables:
  $$H(X, Y) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x, y)\log_2 p(x, y)$$
- **Conditional entropy** is the amount of extra information needed to communicate Y, given that X is already known:
  $$H(Y|X) = \sum_{x\in\mathcal{X}} p(x)H(Y|X = x) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x, y)\log_2 p(y|x)$$
- **Chain rule** connects joint and conditional entropy:
  $$H(X, Y) = H(X) + H(Y|X)$$
  $$H(X_1 ... X_n) = H(X_1) + H(X_2|X_1) + ... + H(X_n|X_1 ... X_{n-1})$$

Example 3

'Twas brillig, and the slithy toves
Did gyre and gimble in the wabe:
All mimsy were the borogoves,
And the mome raths outgrabe.

"Beware the Jabberwock, my son!
The jaws that bite, the claws that catch!
Beware the Jubjub bird, and shun
The frumious Bandersnatch!"

Information in transitions of *Bandersnatch*:

- Surprisal of n given a = 2.45 bits
- Surprisal of d given n = 2.47 bits

Remember average surprisal of a character, $H(X)$, was 4.05 bits.
$H(X|Y)$ turns out to be about 2.8 bits.

What about Example 4?

Thank you, it's a very interesting dance to watch,' said Alice, feeling very glad that it was over at last.

To make predictions about when we insert *that* we need to think about **entropy rate**.

Entropy of a language is the **entropy rate**

- Language is a stochastic process generating a sequence of word tokens
- The entropy of the language is the entropy rate for the stochastic process:

$$H_{rate}(L) = \lim_{n \to \infty} \frac{1}{n} H(X_1 \ldots X_n)$$

- The entropy rate of language is the limit of the entropy rate of a sample of the language, as the sample gets longer and longer.

Hypothesis: **constant** rates of information are preferred

- The **capacity** of a communication **channel** is the number of bits on average that it can transmit
- Capacity defined by the noise in the channel—mutual information of the channel input and output (more next week)

- Assumption: language users want to maximize information transmission while minimizing comprehender difficulty.
- Hypothesis: language users prefer to distribute information uniformly throughout a message
- Entropy Rate Constancy Principle (Genzel & Charniak), Smooth Signal Redundancy Hypothesis (Aylett & Turk), Uniform Information Density (Jaeger)

Hypothesis: **constant** rates of information are preferred

Could apply the hypothesis at all levels of language use:

- In speech we can modulate the duration and energy of our vocalisations
- For vocabulary we can choose longer and shorter forms

  maths vs. *mathematics*, *don't* vs. *do not*
- At sentence level, we may make syntactic reductions:

  The rabbit (that was) chased by Alice.

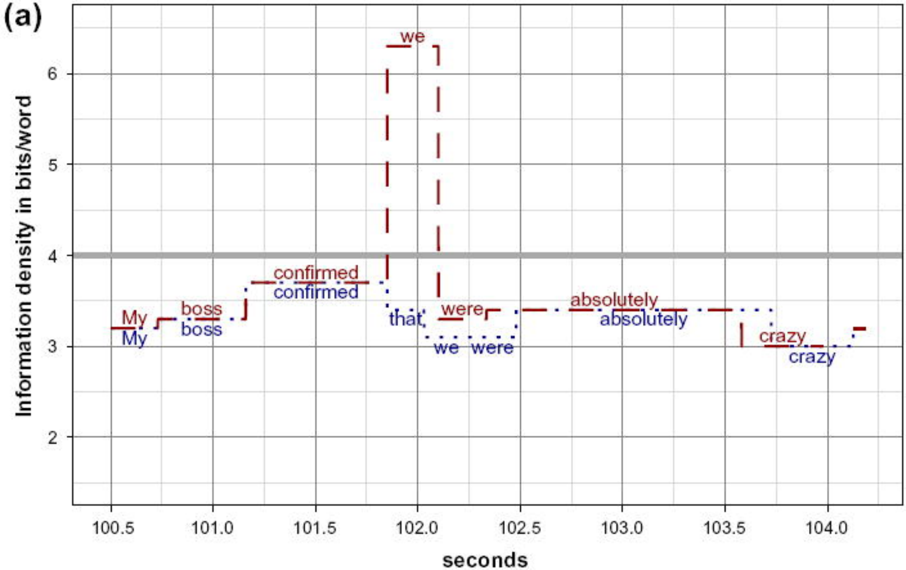
Hypothesis: **constant** rates of information are preferred

Uniform Information Density:

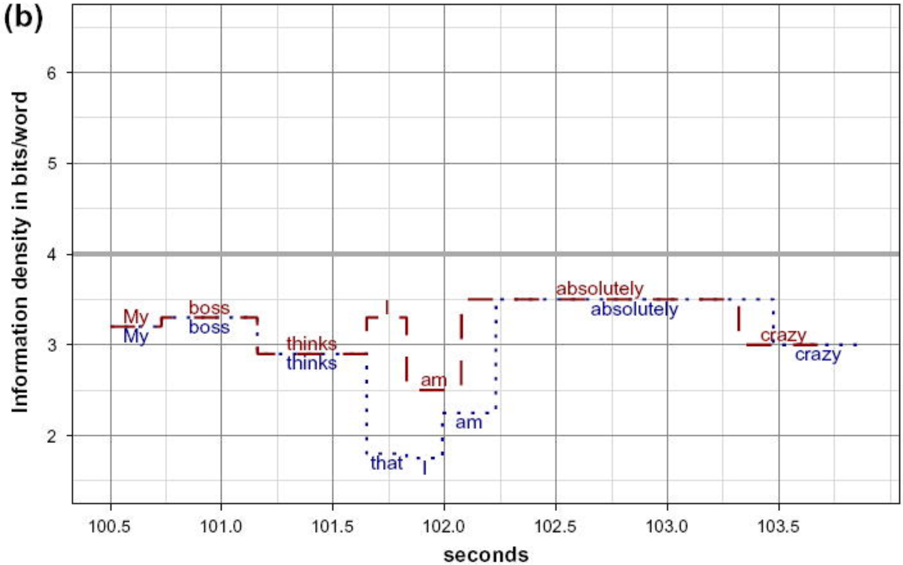
- Within the bounds defined by grammar, speakers prefer utterances that distribute information uniformly across the signal
- Where speakers have a choice between several variants to encode their message, they prefer the variant with more uniform information density

Evaluated on a large scale corpus study of complement clause structures in spontaneous speech (Switchboard Corpus of telephone dialogues)

Hypothesis: **constant** rates of information are preferred



Hypothesis: **constant** rates of information are preferred



- Notice that these information theoretic accounts are rarely explanatory (doesn't explicitly tell us what might be happening in the brain)
- An exception is Hale (2001) where we used surprisal to reason about parse trees and full parallelism
- Information theoretic accounts are unlikely to be the full story but they are predictive of certain phenomena