Entropy

- Entropy, definitions, illustrations
- Entropy measures task difficulty
- Conditional Entropy & Comprehensibility
- Information Gain
- Mutual Information
- Cross-Entropy

Entropy

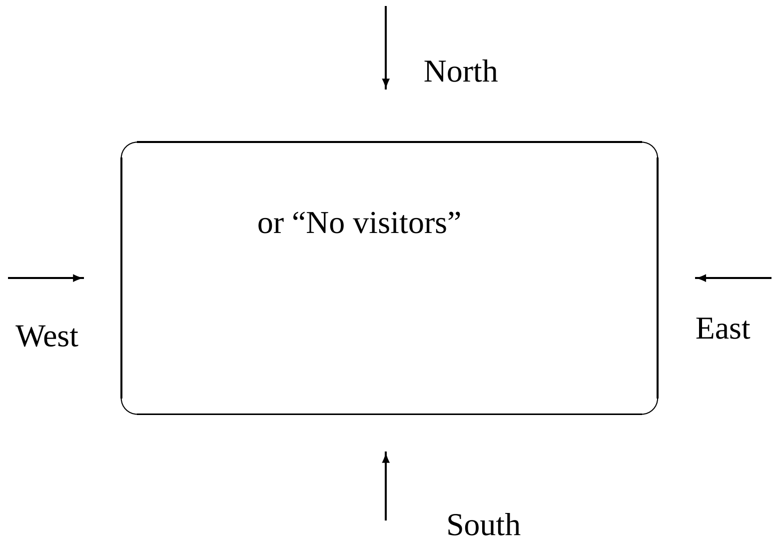
Entropy a.k.a. uncertainty a.k.a. impurity a.k.a. disorder

First in physics (disorder of gas), then in telcommunications.

Optimal coding uses the minimal length in bits.

Messages from Lookout

Consider situation where a lookout must report either no visitor or the direction from which a visitor is approachin, i.e. one of five messages:



Should we code 000, 001, 010, 011, 100? All codes three bits.

Entropy

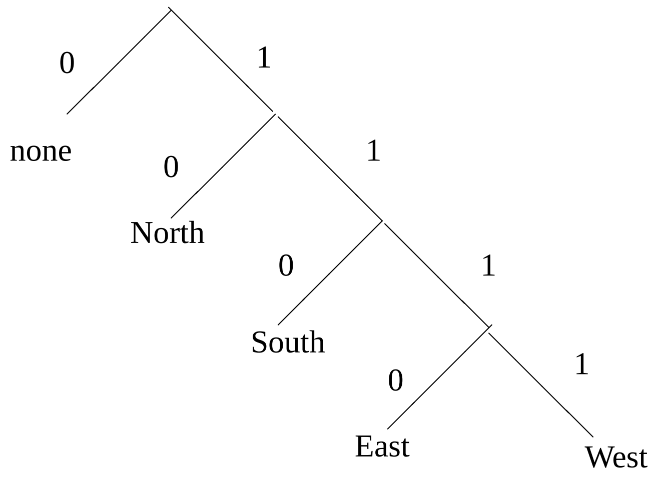
With no further information, we seem to need a code length of three:

$$\text{code length} = \lceil \log_2 |M| \rceil, \text{ where } M \text{ are the messages}$$

But suppose we know that some messages are more frequent than others.

message	rel. freq.
no visitor	99%
North	0.5%
South	0.25%
East, West	0.125%

A Code Tree



message	code
no visitor	0
North	10
South	110
East	1110
West	1111

Expected Code Length

We now calculate the expected code length:

message	code length	rel. freq.	expected bit length
no visitor	1	0.99	0.99
North	2	0.005	0.01
South	3	0.0025	0.0075
East	4	0.00125	0.005
West	4	0.00125	0.005
Total			1.0175

Compare to 3 bits,

code length $= \lceil \log_2 |M| \rceil$, where M are the messages

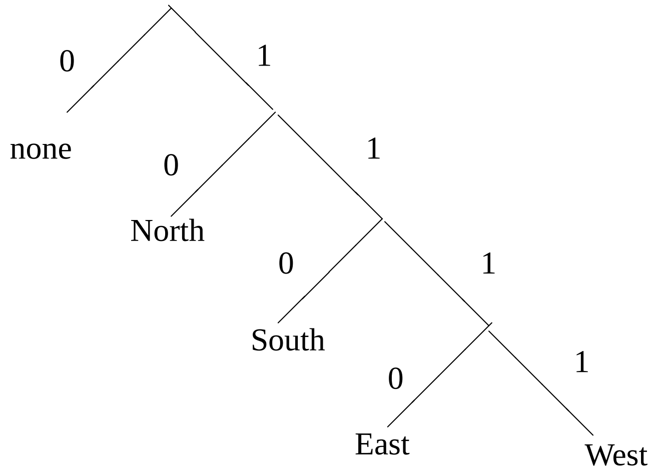
Moral: Bit-Length Should Reflect Likelihood

Let most likely messages be encoded in fewest bits.

Shannon: $-\log_2 p_i$ reflects “uncertainty” of message p_i

message	p_i	$-\log_2 p_i$
no visitor	0.99	0.004
North	0.005	2.3
South	0.0025	2.6
East	0.00125	2.9
West	0.00125	2.9

Communication $\propto$ Information



Binary coding is analogous to receiving yes-no information.

Think of entropy as the ”20 questions” game: You need to ask 0.021 yes/no questions on average to identify the message (information)

Decisions Expressed in Bits

In the entropy formula we sum over all the options, using p_i factor to gives us a weighted average:

$$H(S) = \sum_{i \in S} p_i(-\log_2 p_i)$$

The rest? $-\log_2 p_i$

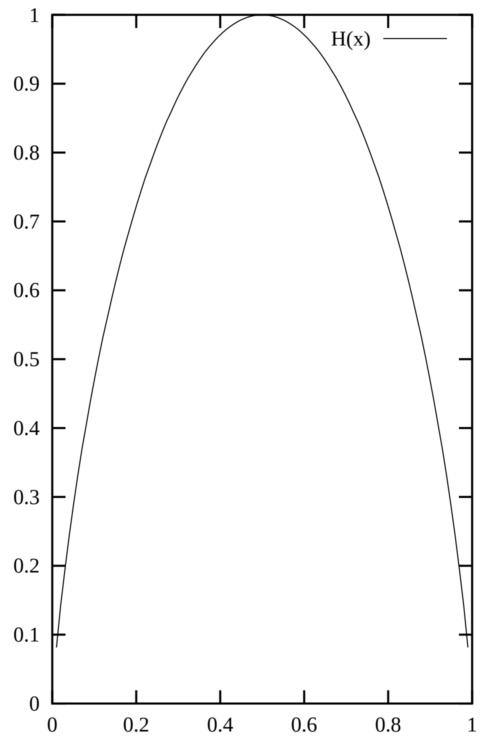
Entropy

Shannon: The optimal code cannot be compressed further than the **entropy** (informational uncertainty) of the dataset:

$$H(S) = -\sum_{i \in S} p_i \log_2 p_i$$

message	p_i	$-\log p_i$	$p_i \log p_i$
no visitor	0.99	0.004	0.0044
North	0.005	2.3	0.0115
South	0.0025	2.6	0.0065
East	0.00125	2.9	0.0036
West	0.00125	2.9	0.0036
Total			0.021

Entropy of Two-Way Choice



Taking Stock

- Entropy measures the amount of information in a random variable
- Directly applicable to categorical variables (see above)
- ..i.e. the degree of freedom in a given situation
- Great freedom of choice (phoneme, letter, words, etc) means few limitations and high entropy.

Measure of Task Difficulty

Example 1: Phonotactics Learning

- [fstrϵč] OK Russian, not English, Dutch, German
  How is this learned?
- Focus on monosyllables allows us to avoid segmentation issues
  Useful, not necessary simplification
- Perhaps psycholinguistically implausible—speech may be organized psychologically, for example, into syllables sequences
  —But sequence learning returns as problem at higher level

Data

- Data: **all** Dutch monosyllables
  - $6,205$ in CELEX
    $6,177$ unique orthographic strings,
    $5,684$ unique phonetic transcriptions
  - Withhold 10% for testing
  - Random strings to test discrimination
  - (Mostly) no negative data! (psychology)
  - Weighted by frequency (mostly)
  - Difficult set — lots of foreign words
    No filtering done to avoid biased selection
- Data: English child-directed speech from CHILDES (one experiment)
  - Described separately

How Difficult is the Task?

- Number of successors variable, $\diamond$ high
- Database entropy (# bits needed to decide which sound follows):

$$H(D) = -\sum_i p_i log_2 p_i$$

database entropy

as sound symbols	unweighted unigrams	4.3
	freq.-weighted unigrams	2.2

- (Baseline) accepting all words which contain only bigrams seen in training $\approx 87.9\%$

Difficulty as Predictor of Error

- Entropy, $H(p_i)$, at each step i of phoneme prediction should predict error
- Idea: a given position $i-1$ is difficult depending on the entropy of the distribution at position i.
- Applied to learning simulators, this correctly predicted onset-coda transition to be the location of the most errors (Stoianov, 2001, Groningen)
- Greater than nucleus-coda break!
- Difficulty of words sums over difficulty of each position

$$\sum_{i-1} H(p_i)$$

Information Gain (Entropy Reduction)

- By adding information, one reduces uncertainty. Information gain compares the entropies of the original system and the system after information is added.
- Suppose visitors never come on Mondays. Then adding information about the day of the week will reduce the entropy:

Day	P	Entropy
Mondays	0.143	0
Other	0.857	0.021
Total		0.018

- Information gain used in constructing decision trees (machine learning)

Linguistic Application of Information Gain

Example 2: Leonoor van der Beek *Topics in Corpus-Based Dutch Syntax*

Chap. 3 “Dative Alternations”

OBL OBJ1	Vervolgens gaf hij mij geel
OBJ1 OBL	Vervolgens gaf hij geel aan de speler
OBJ1 PP-O	Vervolgens gaf hij het mij
PP-0 OBJ1	Vervolgens gaf hij aan die speler een officiële waarschuwing

Cf English, where alternation involves order and category switch

Dative Alternation: Questions

Is alternation promoted by

- "heaviness" (length) of objects?
- informational status (definite vs. indefinite)?
- category of OBJ1 (full NP, *het*, pronoun)?
- verb lexeme?

Data

Some work with hand-corrected *Corpus Gesproken Nederlands* (1 Mil. wd.), *Alpino* corpus (140 K wd.)

- Twente News Corpus (75 Mil. wd.)
- automatically parsed (85.5% correct)
- selected examples manually checked
- excluding ex. with topicalization, clausal objects, passives, *er*-objects

Peeking at Data

Few categorical effects, e.g. even NP status (full, pro,*het*) non-categorical

Most frequent pronouns in double-object constructions

Shifted		Canonical	
542	het	372	dat
45	dat	83	dit
21	't	51	het
19	ze	28	die
7	dit	24	hem
. . .		. . .	

RuG

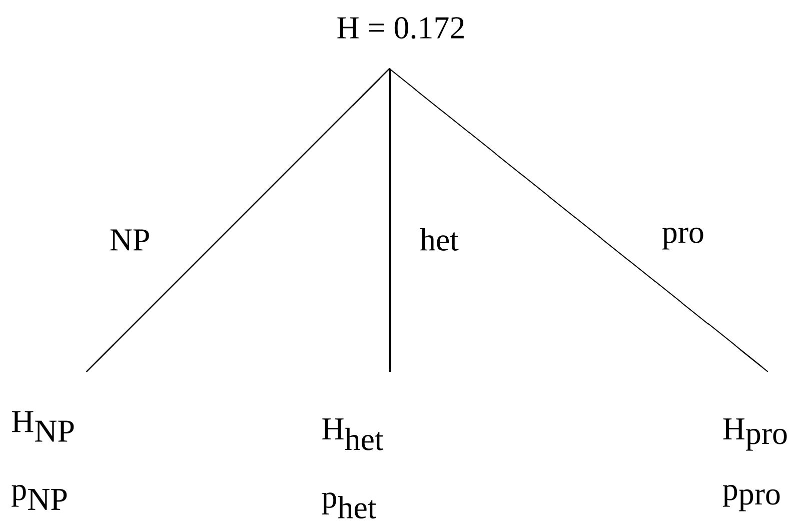
Strategy

1. Calculate entropy of canonical vs. shifted choice
2. For each putative determining factor, calculate entropy once factor is made constant
3. Take weighted ave. of entropies in (2) —remaining entropy
4. Compare original entroy with entropy resulting in (3)—this is INFORMATION GAIN.

Entropy of basic choice (no factors incorporated): 0.172

Canonical order dominates!

Information Gain



$$IG_f(S) = H(S) - \sum_{v \in \text{Values}(f)} \frac{|S_{f=v}|}{|S|} H(p_{f=v})$$

Effect on Order

Entropy of $\{OBJ1, OBL\}$ order = 0.172

1.	Cat of OBJ1 (NP,het,pro)	0.110	-36%
2.	Verb lexeme (give,send,...)	0.152	-12%
3.	OBJ1-Cat & Verb lexeme	0.094	-45%

Comments

1. category OBJ1 has a significant effect in reducing uncertainty of order
2. lexeme has surprisingly little, considering how many classes there are
3. (1) and (2) are largely orthogonal

Effect on Oblique Realization NP vs. PP

Entropy of $\{NP, PP\}$ realization = 0.578

1.	Cat of OBJ1 (NP,het,pro)	0.578	-0%
2.	Verb lexeme (give,send,...)	0.426	-26%
3.	OBJ1-Cat & Verb lexeme	0.094	-27%

Comments

1. category OBJ1 has a no effect in reducing uncertainty of category realization of OBL
2. lexeme has moderate effect
3. (1) and (2) seem orthogonal

Other Remarks

- Direct objects *heavier* in shifted construction, indirect objects *lighter*.
  —contrary to complexity idea (Behagel)
- Weight does not affect order in the *Mittelveld* (surprising), but it seems to promote object extraposition.
- Principle known (definite) elements early not strong:
  - 85% of OBL OBJ1 orders had indefinite OBJ1, only 45% of OBJ1 OBL orders (confirming), but
  - 32% of OBJ1 OBL orders had indef. OBJ1 & def. OBL

Joint entropy

- The joint entropy of a pair of random variables is the amount of info needed on average to specify both their values:

$$H(X,Y) = -\sum_{x \in X} \sum_{y \in Y} p(x,y) log_2 p(x,y)$$

Conditional entropy

- CE is always calculated in relation to other information
- CE relies on conditional probabilities
- CE of Y given X is the joint entropy of X and Y minus the entropy of X:

$$\begin{aligned} H(Y \mid X) &= H(X,Y) - H(X) \\ &= -\sum_{x \in X} \sum_{y \in Y} p(x,y) log_2 p(y \mid x) \end{aligned}$$

- As opposed to joint entropy, CE is not symmetrical:

$$H(Y \mid X) \neq H(X \mid Y)$$

Measuring Conditional Entropy

- $H(X \mid Y)$ is the uncertainty in X given knowledge of Y.
- CE measures how much entropy a random variable X has remaining if the value of a second random variable Y is known
- This means that in a linguistic context, CE can be used to measure the difficulty of predicting a unit which is dependent on another.

Application of Conditional Entropy: Comprehensibility

- Charlotte Gooskens & Jens Moberg are investigating Scandinavian “semicommunication”, Jens also working with me.
- Sandinavians hold conversations in which each speaks his own language
- They understand each other to varying degrees, e.g. Danes understand Swedes better than *vice versa*.
- Proposed explanations: linguistic differences, experience, attitudes
- Project focus: linguistic differences

The Relevant Mapping

- Idea: the mapping from one language to another may be more complicated in one direction than in reverse
- Perhaps Danes understand Swedes better than *vice versa* because the mapping is easier
- As an example we examine the mapping from Swedish to Danish
- Whose *comprehension* are we modeling?

Danish Comprehension of Swedish

- Whose *comprehension* are we modeling?
- The Dane hears a Swedish word and can understand it more easily *ceteribus paribus* if he can map it to Danish.
- Prediction: $\text{CE}(\text{Danish}|\text{Swedish}) \ll \text{CE}(\text{Swedish}|\text{Danish})$
- How can we operationalize this?

How to Determine Conditional Entropy

1. Obtain bilingual texts, e.g. from *Europarl*
2. Extract the “cognate” (similar) words
3. Convert to phonemic representation
4. Align phonemes across languages
5. Extract statististics of correspondence

Danish Realizations of Swedish /a/

Tabel 1: Conditional probabilities for Danish sounds given Swedish /a/

Danish →	ə	a	ɒ	Others
Swedish ↓				
a	0.45	0.14	0.10	0.31
o				
u				
..etc				

Calculating CE for Phoneme Realizations

- Entropy H $(P(D \mid \text{/a/}))$

$$H = -\sum_{d \in D, \text{/a/}} p(d, \text{/a/}) log_2(d \mid \text{/a/})$$

$$H = -(0.45 * log_2 0.45) + (0.14 * log_2 0.14) + (0.10 * log_2 0.10) + (0.31 * log_2 0.31)$$

- H(D|a) = 1.775 bits of information
- Calculation above (incorrectly) uses $p(d|\text{/a/})$ to weight the $-\log_2(d \mid \text{/a/})$ for different d realizations. In genuine calculation, this will be weighted by $p(d, \text{/a/}) = p(d|\text{/a/}) \cdot p(\text{/a/})$
- If this is done for all phonemes, we derive predictions where intelligibility problems are, i.e. where errors are most likely to be made

Preliminary Results for Small Corpus

- Sample corpus: 206 Danish-Swedish word pairs, some non-cognates
- Due to insertions and deletions, the word length is sometimes different. Some sounds map to $\emptyset$ (corresponding to insertion and/or deletion)
- D means Danish and S Swedish
- H(D|S) = **2.29**
- H(S|D) = **2.22**
- With $\emptyset$:
- H(D|S) = **2.25**
- H(S|D) = **2.23**

Recall prediction: CE(Danish$|$Swedish) $\ll$ CE(Swedish$|$Danish)!

Preliminary Results for a Sample Corpus

- For 25 words: $H(D|S) = 1.22, H(S|D) = 1.11$
- For 200 words: $H(D|S) = 2.22, H(S|D) = 2.19$

...stay tuned!

(Pointwise) Mutual Information

$$MI(X,Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

- measure of association strength between two variables
  —compare to χ^2
- how often do x and y co-occur, compared to how often they'd be expected to co-occur if independent ($p(x)p(y)$)
- *Pointwise* where we use individual x, y (without summing over all values of X, Y)

RuG

Applying Mutual Information

Begoña Villada Moiron *Data-Driven Identification of Fixed Expressions and their Modifiability*, Diss. Groningen 2005

w_i / w_{i+1}	*house*	*shot*	*mess*	...	totals
big	2×10^{-4}	1×10^{-4}	1.5×10^{-4}	...	1.5×10^{-4}
small	2×10^{-4}	4×10^{-6}	1.1×10^{-5}	...	1.5×10^{-4}
red	2×10^{-5}	1×10^{-7}	1.5×10^{-7}	...	1.5×10^{-5}
...	...	...	...	...	...
totals	6×10^{-5}	8×10^{6}	1×10^{-5}	...	...

- we simply count how often $w_i w_{i+1}$ appear
- divide this by the number of bigrams to obtain relative frequencies (cell values)
- we use relative frequencies as *estimates* of relative probabilities $p(w_i, w_{i+1})$
- *marginal* values give us $p(w_i), p(w_{i+1})$

Mutual Information & Conditional Entropy

$$
\begin{aligned}
MI(X,Y) &= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \\
&= \sum_{x,y} p(x,y) \log \frac{p(x|y)}{p(x)} \\
&= -\sum_{x,y} p(x,y) \log p(x) + \sum_{x,y} p(x,y) \log p(x|y) \\
&= -\sum_{x,y} p(x,y) \log p(x) - (-\sum_{x,y} p(x,y) \log p(x|y)) \\
&= H(X) - H(X|Y)
\end{aligned}
$$

Mutual entropy of X, Y is just entropy of X less the conditional entropy of $(X|Y)$.

Cross Entropy

CROSS ENTROPY compares empirical X to theoretical m (model) distributions.

$$H(X,m) = -\sum_{x \in X} p(x) \log m(x)$$

where $m(x)$ is prob. of x according to model

- $\forall m, H(X) < H(X,m)$
- So we can use simple models to estimate (give a bound on) true entropy
- The more accurate the model, the more m resembles X

Cross Entropy, Example

You compare two coins, each under the model that the coin is honest. You obtain different empirical frequencies:

$$X = \{x_1, x_2\}, m(x_1) = m(x_2) = 0.5$$
$$X' = \{x'_1, x'_2\}, m(x'_1) = m(x'_2) = 0.5$$

$$p(x_1) = p(x_2) = 0.5$$
$$p(x'_1) = 0.25, p(x'_2) = 0.75$$

Cross Entropy, Example

We compare cross entropies of same model under different empirical frequencies.

	$p(x)$	$m(x)$	$-\log m(x)$	$-p(x)\log m(x)$
x_1	0.5	0.5	1	0.5
x_2	0.5	0.5	1	0.5

$$H(X, m) = 1$$

x'_1	0.25	0.5	1	0.25
x'_2	0.75	0.5	1	0.75

$$H(X', m) = 1$$

Example, Cross Entropy

Different models, same empirical frequencies.

	$p(x)$	$m(x)$	$-\log m(x)$	$-p(x)\log m(x)$
x_1'	0.25	0.5	1	0.25
x_2'	0.75	0.5	1	0.75

$$H(X', m) = 1$$

	$p(x)$	$m'(x)$	$-\log m'(x)$	$-p(x)\log m'(x)$
x_1'	0.25	0.25	2	0.5
x_2'	0.75	0.75	0.4	0.3

$$H(X', m') \approx 0.8$$

—to get $\log_2(x)$, remember that $\log_a(x)/\log_a(b) = \log_b(x)$

End of Entropy

QuantLing

RuG