Introduction to
Natural Language Processing (600.465)

Probability

Dr. Jan Hajič
CS Dept., Johns Hopkins Univ.
hajic@cs.jhu.edu
www.cs.jhu.edu/~hajic

Experiments & Sample Spaces

- Experiment, process, test, …
- Set of possible basic outcomes: sample space Ω
  - coin toss (Ω = {head,tail}), die (Ω = {1..6})
  - yes/no opinion poll, quality test (bad/good) (Ω = {0,1})
  - lottery ($|\Omega| \cong 10^7 .. 10^{12}$)
  - # of traffic accidents somewhere per year (Ω = N)
  - spelling errors ($\Omega = Z^*$), where Z is an alphabet, and Z^* is a set of possible strings over such and alphabet
  - missing word ($|\Omega| \cong$ vocabulary size)

Events

- Event A is a set of basic outcomes
- Usually $A \subset \Omega$, and all $A \in 2^{\Omega}$ (the event space)
  - Ω is then the certain event, $\emptyset$ is the impossible event
- Example:
  - experiment: three times coin toss
    - **Ω = {HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}**
  - count cases with exactly two tails: then
    - **A = {HTT, THT, TTH}**
  - all heads:
    - **A = {HHH}**

Probability

- Repeat experiment many times, record how many times a given event A occurred ("count" c_1).
- Do this whole series many times; remember all c_is.
- Observation: if repeated really many times, the ratios of $\mathbf{c_i/T_i}$ (where T_i is the number of experiments run in the *i-th* series) are close to some (unknown but) **constant** value.
- Call this constant a ***probability of A***. Notation: **p(A)**

Estimating probability

- Remember: ... close to an *unknown* constant.
- We can only estimate it:
  - from a single series (typical case, as mostly the outcome of a series is given to us and we cannot repeat the experiment), set

$$p(A) = c_1/T_1.$$

  - otherwise, take the weighted average of all c_i/T_i (or, if the data allows, simply look at the set of series as if it is a single long series).
- This is the **best** estimate.

Example

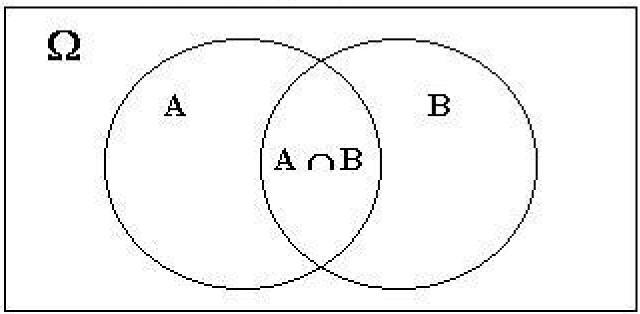
- Recall our example:
  - experiment: three times coin toss
    - Ω = {**HHH, HHT, HTH, HTT, THH, THT, TTH, TTT**}
  - count cases with exactly two tails: A = {**HTT, THT, TTH**}
- Run an experiment 1000 times (i.e. 3000 tosses)
- Counted: 386 cases with two tails (**HTT, THT, or TTH**)
- estimate: p(A) = 386 / 1000 = .386
- Run again: 373, 399, 382, 355, 372, 406, 359
  - p(A) = .379 (weighted average) or simply 3032 / 8000
- *Uniform* distribution assumption: p(A) = 3/8 = .375

Basic Properties

- Basic properties:
  - p: $2^{\Omega} \rightarrow [0,1]$
  - $p(\Omega) = 1$
  - Disjoint events: $p(\cup A_i) = \sum_i p(A_i)$
- [NB: <u>*axiomatic definition*</u> of probability: take the above three conditions as axioms]
- Immediate consequences:
  - $p(\emptyset) = 0$, $p(\overline{A}) = 1 - p(A)$, $A \subseteq B \Rightarrow p(A) \leq p(B)$
  - $\sum_{a \in \Omega} p(a) = 1$

Joint and Conditional Probability

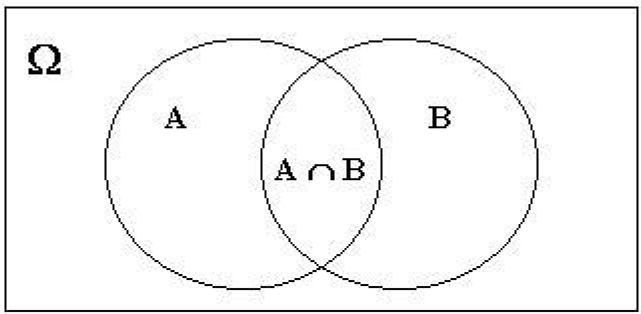
- $p(A,B) = p(A \cap B)$
- $p(A|B) = p(A,B) / p(B)$
  - Estimating form counts:
    - $\mathbf{p(A|B) = p(A,B) / p(B) = (c(A \cap B) / T) / (c(B) / T) =}$
      $\mathbf{= c(A \cap B) / c(B)}$



Bayes Rule

- $p(A,B) = p(B,A)$ since $p(A \cap B) = p(B \cap A)$
  - therefore: $p(A|B)\ p(B) = p(B|A)\ p(A)$, and therefore

$$p(A|B) = p(B|A)\ p(A) / p(B) \quad !$$



Independence

- Can we compute p(A,B) from p(A) and p(B)?
- Recall from previous foil:

$$p(A|B) = p(B|A)\ p(A) / p(B)$$

$$p(A|B)\ p(B) = p(B|A)\ p(A)$$

$$p(A,B) = p(B|A)\ p(A)$$

... we're almost there: how p(B|A) relates to p(B)?

- p(B|A) = P(B) iff A and B are **independent**

- Example: two coin tosses, weather today and weather on March 4th 1789;
- Any two events for which p(B|A) = P(B)!

Chain Rule

$$p(A_1, A_2, A_3, A_4, \ldots, A_n) = \quad !$$

$$p(A_1|A_2,A_3,A_4,\ldots,A_n) \times p(A_2|A_3,A_4,\ldots,A_n) \times$$
$$\times\, p(A_3|A_4,\ldots,A_n) \times \;\ldots\; p(A_{n-1}|A_n) \times p(A_n)$$

- this is a direct consequence of the Bayes rule.

The Golden Rule (of Classic Statistical NLP)

- Interested in an event A given B (where it is not easy or practical or desirable) to estimate p(A|B)):
- take Bayes rule, max over all Bs:
- $\text{argmax}_A\ p(A|B) = \text{argmax}_A\ p(B|A)\ .\ p(A) / p(B) =$

$$\text{argmax}_A\ p(B|A)\ p(A)\ !$$

- ... as p(B) is constant when changing As

Random Variables

- is a function $X: \Omega \rightarrow Q$
  - in general: $Q = R^n$, typically R
  - easier to handle real numbers than real-world events
- random variable is *discrete* if Q is countable (i.e. also if finite)
- Example: *die*: natural "numbering" [1,6], *coin*: {0,1}
- Probability distribution:
  - $p_X(x) = p(X=x) =_{df} p(A_x)$ where $A_x = \{a \in \Omega : X(a) = x\}$
  - often just p(x) if it is clear from context what X is

Expectation Joint and Conditional Distributions

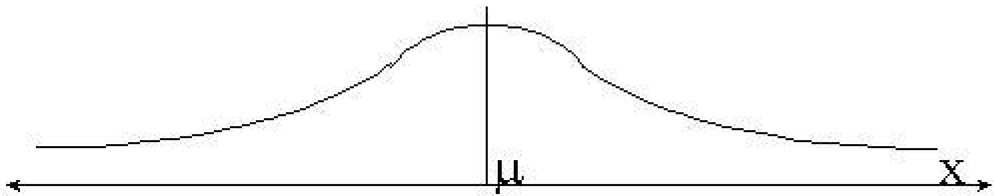
- is a mean of a random variable (weighted average)
  - $E(X) = \Sigma_{x \in X(\Omega)} x \cdot p_X(x)$
- Example: one six-sided die: 3.5, two dice (sum) 7
- Joint and Conditional distribution rules:
  - analogous to probability of events
- Bayes: $p_{X|Y}(x,y) =_{\text{notation}} p_{XY}(x|y) =_{\text{even simpler notation}}$
  $\mathbf{p(x|y) = p(y|x) \cdot p(x) / p(y)}$
- Chain rule: $\mathbf{p(w,x,y,z) = p(z).p(y|z).p(x|y,z).p(w|x,y,z)}$

Standard distributions

- Binomial (discrete)
  - outcome: 0 or 1 (thus: *bi*nomial)
  - make *n* trials
  - interested in the (probability of) number of successes *r*
- Must be careful: it's not uniform!
- $p_b(r|n) = \binom{n}{r} / 2^n$ (for equally likely outcome)
- $\binom{n}{r}$ counts how many possibilities there are for choosing r objects out of n; $= n! / (n-r)!r!$

Continuous Distributions

- The normal distribution ("Gaussian")
- $p_{norm}(x|\mu,\sigma) = e^{-(x-\mu)^2/(2\sigma^2)}/\sigma\sqrt{2\pi}$
- where:
  - μ is the mean (x-coordinate of the peak) (0)
  - σ is the standard deviation (1)



- other: hyperbolic, t

Introduction to
Natural Language Processing (600.465)

Essential Information Theory I

Dr. Jan Hajič
CS Dept., Johns Hopkins Univ.
hajic@cs.jhu.edu
www.cs.jhu.edu/~hajic

The Notion of Entropy

- Entropy ~ "chaos", fuzziness, opposite of order, ...
  - you know it:
    - **it is much easier to create "mess" than to tidy things up...**
- Comes from physics:
  - Entropy does not go down unless energy is used
- Measure of ***uncertainty:***
  - if low... low uncertainty; the higher the entropy, the higher uncertainty, but the higher "surprise" (information) we can get out of an experiment

The Formula

- Let $p_X(x)$ be a distribution of random variable X
- Basic outcomes (alphabet) Ω

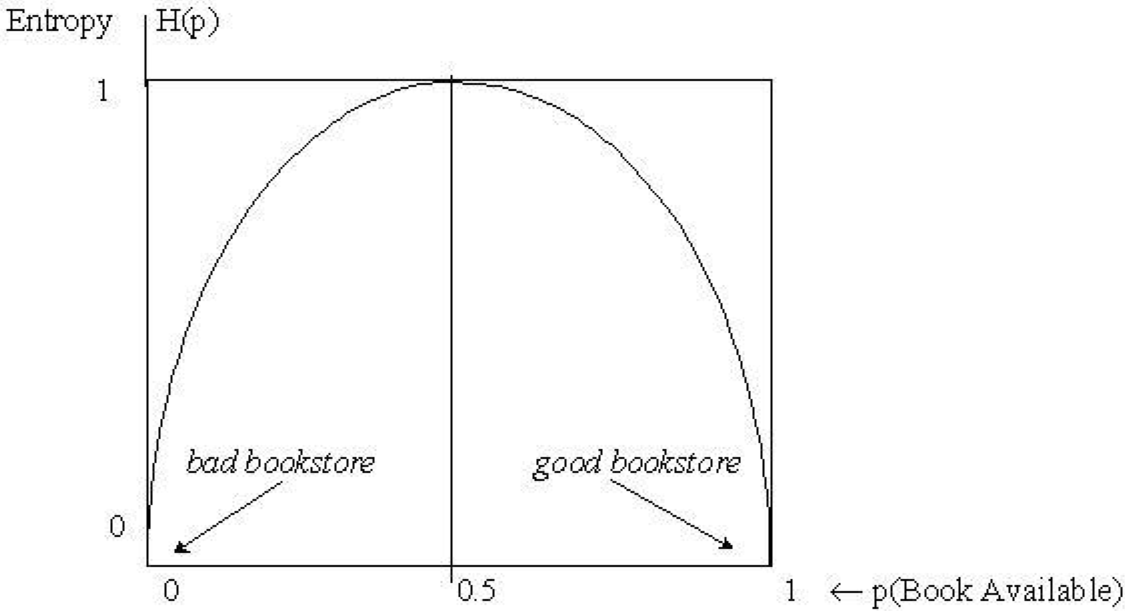
$$H(X) = - \sum_{x \in \Omega} p(x) \log_2 p(x) \quad !$$

- Unit: bits ($\log_{10}$: nats)
- Notation: $H(X) = H_p(X) = H(p) = H_X(p) = H(p_X)$

Using the Formula: Example

- Toss a fair coin: $\Omega = \{\text{head}, \text{tail}\}$
  - p(head) = .5, p(tail) = .5
  - **H(p)** $= - 0.5 \log_2(0.5) + (- 0.5 \log_2(0.5)) =$ $2 \times ((-0.5) \times (-1)) = 2 \times 0.5 = \mathbf{1}$
- Take fair, 32-sided die: p(x) = 1 / 32 for every side x
  - **H(p)** $= -\sum_{i = 1..32} p(x_i) \log_2 p(x_i) = - 32 (p(x_1) \log_2 p(x_1)$ (since for all i $p(x_i) = p(x_1) = 1/32$) $= -32 \times ((1/32) \times (-5)) = \mathbf{5}$ *(now you see why it's called **bits**?)*
- Unfair coin:
  - p(head) = .2 ... **H(p) = .722**; p(head) = .01 ... **H(p) = .081**

Example: Book Availability



The Limits

- When $H(p) = 0$?
  - if a result of an experiment is ***known*** ahead of time:
  - necessarily:

$$\exists x \in \Omega;\ p(x) = 1\ \&\ \forall y \in \Omega;\ y \neq x \Rightarrow p(y) = 0$$

- Upper bound?
  - none in general
  - for $|\Omega| = n$: $H(p) \leq \log_2 n$
    - **nothing can be more uncertain than the uniform distribution**

Entropy and Expectation

- Recall:
  - $E(X) = \sum_{x \in X(\Omega)} p_X(x) \times x$
- Then:

$$E(\log_2(1/p_X(x))) = \sum_{x \in X(\Omega)} p_X(x) \log_2(1/p_X(x)) =$$

$$= - \sum_{x \in X(\Omega)} p_X(x) \log_2 p_X(x) =$$

$$= H(p_X) =_{notation} H(p)$$

Perplexity: motivation

- Recall:
  - 2 equiprobable outcomes: H(p) = 1 bit
  - 32 equiprobable outcomes: H(p) = 5 bits
  - 4.3 billion equiprobable outcomes: H(p) ~= 32 bits
- What if the outcomes are not equiprobable?
  - 32 outcomes, 2 equiprobable at .5, rest impossible:
    - H(p) = 1 bit
  - Any measure for comparing the entropy (i.e. uncertainty/difficulty of prediction) (also) for random variables with *different number of outcomes*?

Perplexity

- Perplexity:
  - $G(p) = 2^{H(p)}$
- ... so we are back at 32 (for 32 eqp. outcomes), 2 for fair coins, etc.
- it is easier to imagine:
  - NLP example: vocabulary size of a vocabulary with uniform distribution, which is equally hard to predict
- the "wilder" (biased) distribution, the better:
  - lower entropy, lower perplexity

Joint Entropy and Conditional Entropy

- Two random variables: X (space Ω), Y (Ψ)
- Joint entropy:
  - no big deal: ((X,Y) considered a single event):

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

- Conditional entropy:

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

recall that $H(X) = E(\log_2(1/p_X(x)))$

(weighted "average", and <u>weights are not conditional</u>)

Conditional Entropy (Using the Calculus)

- other definition:

$H(Y|X) = \sum_{x \in \Omega} p(x)\, H(Y|X=x) =$

for H(Y|X=x), we can use the single-variable definition (x ~ constant)

$= \sum_{x \in \Omega} p(x) \left(- \sum_{y \in \Psi} p(y|x) \log_2 p(y|x) \right) =$

$= - \sum_{x \in \Omega} \sum_{y \in \Psi} p(y|x)\, p(x) \log_2 p(y|x) =$

$= - \sum_{x \in \Omega} \sum_{y \in \Psi} p(x,y) \log_2 p(y|x)$

Properties of Entropy I

- Entropy is non-negative:
  - $H(X) \geq 0$
  - proof: (recall: $H(X) = - \sum_{x \in \Omega} p(x) \log_2 p(x)$)
    - log(p(x)) is negative or zero for $x \leq 1$,
    - p(x) is non-negative; their product p(x)log(p(x) is thus negative;
    - sum of negative numbers is negative;
    - and $-f$ is positive for negative f
- Chain rule:
  - $H(X,Y) = H(Y|X) + H(X)$, as well as
  - $H(X,Y) = H(X|Y) + H(Y)$ (since $H(Y,X) = H(X,Y)$)

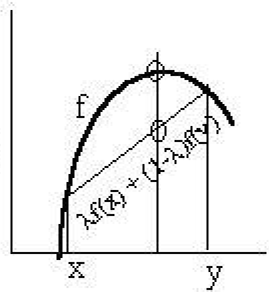
Properties of Entropy II

- Conditional Entropy is better (than unconditional):
  - $H(Y|X) \le H(Y)$ (proof on Monday)
- $H(X,Y) \le H(X) + H(Y)$ (follows from the previous (in)equalities)
  - **equality iff X,Y independent**
  - **[recall: X,Y independent iff $p(X,Y) = p(X)p(Y)$]**
- H(p) is concave (remember the book availability graph?)
  - concave function f over an interval (a,b):

$$\forall x,y \in (a,b),\ \forall \lambda \in [0,1]:$$

$$f(\lambda x + (1-\lambda)y) \ge \lambda f(x) + (1-\lambda)f(y)$$

  - **function f is convex if -f is concave**



- **[for proofs and generalizations, see Cover/Thomas]**

"Coding" Interpretation of Entropy

- The least (average) number of bits needed to encode a message (string, sequence, series,...) (each element having being a result of a random process with some distribution p): = H(p)
- Remember various compressing algorithms?
  - they do well on data with repeating (= easily predictable = low entropy) patterns
  - their results though have high entropy $\Rightarrow$ compressing compressed data does nothing

Coding: Example

- How many bits do we need for ISO Latin 1?
  - ⇒ the trivial answer: 8
- Experience: some chars are more common, some (very) rare:
  - **...so what if we use more bits for the rare, and less bits for the frequent? [be careful: want to decode (easily)!]**
  - **suppose: p('a') = 0.3, p('b') = 0.3, p('c') = 0.3, the rest: p(x)≅ .0004**
  - **code: 'a' ~ 00, 'b' ~ 01, 'c' ~ 10, rest: $11b_1b_2b_3b_4b_5b_6b_7b_8$**
  - **code acbbécbaac:** 00100101<u>1100001111</u>1001000010

    a c b b é c b a a c
  - **number of bits used: 28 (vs. 80 using "naive" coding)**
- code length ~ 1 / probability; conditional prob OK!

Entropy of a Language

- Imagine that we produce the next letter using

$$p(l_{n+1}|l_1,\ldots,l_n),$$

where $l_1,\ldots,l_n$ is the sequence of ***all*** the letters which had been uttered so far (i.e. ***n*** is really big!); let's call $l_1,\ldots,l_n$ the ***history*** h (h_{n+1}), and all histories H:

- Then compute its entropy:
  - $-\sum_{h \in H} \sum_{l \in A} p(l,h) \log_2 p(l|h)$
- Not very practical, isn't it?

Cross-Entropy

- Typical case: we've got series of observations
  $T = \{t_1, t_2, t_3, t_4, \ldots, t_n\}$ (numbers, words, ...; $t_i \in \Omega$);

  estimate (simple):

  $\forall y \in \Omega: \tilde{p}(y) = c(y) / |T|$, def. $c(y) = |\{t \in T; t = y\}|$
- ...but the true p is unknown; every sample is too small!
- Natural question: how well do we do using $\tilde{p}$ **[instead of p]**?
- Idea: simulate actual p by using a different T'

  (or rather: by using different observation we simulate the insufficiency of T vs. some other data ("random" difference))

Cross Entropy: The Formula

- $H_{p'}(\tilde{p}) = H(p') + D(p' \| \tilde{p})$

$$H_{p'}(\tilde{p}) = - \sum_{x \in \Omega} p'(x) \log_2 \tilde{p}(x) \quad !$$

- p' is certainly not the true p, but we can consider it the "real world" distribution against which we test $\tilde{p}$
- note on notation (confusing...): $p/p' \leftrightarrow \tilde{p}$, also $H_{T'}(p)$
- (Cross)Perplexity: $G_{p'}(p) = G_{T'}(p) = 2^{H_{p'}(\tilde{p})}$

Conditional Cross Entropy

- So far: "unconditional" distribution(s) $p(x)$, $p'(x)$...
- In practice: virtually always conditioning on context
- Interested in: sample space Ψ, r.v. Y, $y \in \Psi$;

  context: sample space Ω, r.v. X, $x \in \Omega$;:

  "our" distribution $p(y|x)$, test against $p'(y,x)$,

  which is taken from some independent data:

$$H_{p'}(p) = - \sum_{y \in \Psi, x \in \Omega} p'(y,x) \log_2 p(y|x)$$

Sample Space vs. Data

- In practice, it is often inconvenient to sum over the sample space(s) Ψ, Ω (especially for cross entropy!)
- Use the following formula:

$$H_{p'}(p) = - \sum_{y \in \Psi, x \in \Omega} p'(y,x) \log_2 p(y|x) = - 1/|T'| \sum_{i=1..|T'|} \log_2 p(y_i|x_i)$$

!

- **This is in fact the normalized log probability of the "test" data:**

$$H_{p'}(p) = - 1/|T'| \log_2 \prod_{i=1..|T'|} p(y_i|x_i)$$

Computation Example

- $\Omega = \{a,b,..,z\}$, prob. distribution (assumed/estimated from data):

  p(a) = .25, p(b) = .5, p(α) = 1/64 for $\alpha \in \{c..r\}$, = 0 for the rest: s,t,u,v,w,x,y,z
- Data (test): <u>barb</u> **p'(a) = p'(r) = .25, p'(b) = .5**
- Sum over Ω:

α	**a**	**b**	**c**	**d**	**e**	**f**	**g**	**...**	**p**	**q**	**r**	**s**	**t**	**...**	**z**	
$-p'(\alpha)\log_2 p(\alpha)$	**.5+**	**.5+**	**0+**	**0+**	**0+**	**0+**	**0+**	**0+**	**0+**	**0+**	**1.5+**	**0+**	**0+**	**0+**	**0+0**	**= <u>2.5</u>**

- Sum over data:

i / s_i	**1/b**	**2/a**	**3/r**	**4/b**		$1/\|T'\|$
$-\log_2 p(s_i)$	**1 +**	**2 +**	**6 +**	**1**	**= 10**	**(1/4) × 10 = <u>2.5</u>**

Cross Entropy: Some Observations

- $H(p)$?? $<, =, >$?? $H_{p'}(p)$: ALL!
- Previous example:

[p(a) = .25, p(b) = .5, p(α) = 1/64 for $\alpha \in$ {c..r}, = 0 for the rest: s,t,u,v,w,x,y,z]

$$H(p) = 2.5 \text{ bits} = H(p')\ (\text{barb})$$

- Other data: probable: (1/8)(6+6+6+1+2+1+6+6)= 4.25

$$H(p) < 4.25 \text{ bits} = H(p')\ (\text{probable})$$

- And finally: abba: (1/4)(2+1+1+2)= 1.5

$$H(p) > 1.5 \text{ bits} = H(p')\ (\text{abba})$$

- But what about: baby $-p'(\text{'y'})\log_2 p(\text{'y'}) = -.25\log_2 0 = \infty$ (??)

Cross Entropy: Usage

- Comparing data??
  - *NO!* (we believe that we test on ***real*** data!)
- Rather: comparing distributions (***vs.*** real data)
- Have (got) 2 distributions: p and q (on some Ω, X)
  - which is better?
  - better: has lower cross-entropy (perplexity) on real data S
- "Real" data: S
- $H_S(p) = - 1/|S| \sum_{i=1..|S|} \log_2 p(y_i|x_i)$?? $H_S(q) = - 1/|S| \sum_{i=1..|S|} \log_2 q(y_i|x_i)$

Comparing Distributions

Test data S: <u>probable</u>

- p(.) from prev. example: $H_S(p) = 4.25$

  p(a) = .25, p(b) = .5, p(α) = 1/64 for α ∈{c..r}, = 0 for the rest: s,t,u,v,w,x,y,z

- q(.|.) (conditional; defined by a table):

q(.|.)→ ↓	a	b	e	l	o	p	r	other
a	0	.5	0	0	0	.125	0	0
b	1	0	0	0	1	.125	0	0
e	0	0	0	1	0	.125	0	0
l	0	.5	0	0	0	.125	0	0
o	0	0	0	0	0	.125	1	0
p	0	0	0	0	0	.125	0	1
r	0	0	0	0	0	.125	0	0
other	0	0	1	0	0	.125	0	0

ex.: q(o|r) = 1

q(r|p) = .125

(1/8) (log(p|oth.)+log(r|p)+log(o|r)+log(b|o)+log(a|b)+log(b|a)+log(l|b)+log(e|l))

(1/8) (0 + 3 + 0 + 0 + 1 + 0 + 1 + 0)

$H_S(q) = .625$