Introduction to Symbolic and Statistical NLP in Scheme

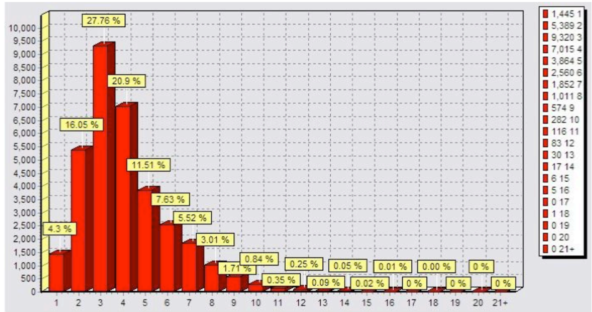
Damir Ćavar
dcavar@unizd.hr

ESSLLI 2006, Malaga

July/August 2006

Frequency Profiles

- Token-length on frequency



Information Theory

- Some background information on the bigram statistics from `countbigram*.ss` and `average-mi.ss/average-re.ss`.

Information Theory

- Surprise effect:
  - Coin tossing and observing the results
  - What is our prior believe or expectation about an outcome?
  - How surprised are we to see a certain outcome?
- Data compression:
  - Knowing about the distributional properties of some data
  - What is the best compression we can get by mapping it to bit-representations?
  - Is there a formal way to calculate the optimal representation for data transmission?

Information Theory

- Entropy:
  - Entropy as uncertainty
    - Tossing a coin = not knowing what the outcome will be.
    - Probability distribution:
      - Fair coin
      - Biased coin, unlimited probability distributions

Information Theory

- Entropy:
  - Entropy as uncertainty
    - Is there a way to calculate the uncertainty and formulate a function on the basis of a probability distribution?
    - Let us design such a function:
      - $H[X]$ is the measure for X, with X a probability distribution
      - H takes X, with $X = \{P(1), P(2), \ldots P(N)\}$ as an argument
      - and returns a real number, the value of uncertainty

Information Theory

- Designing a function for Entropy:

1. Maximum uncertainty in uniform distribution: every possible outcome is equally likely

→ This is the maximum H can return

2. H is a continuous function over the probabilities

→ changing the probabilities slightly leads to slight changes of H

Information Theory

- Grouping Probabilities:
  - $X = \{P(1) = .5, P(2) = .2, P(3) = (.3)\}$:
  - is equivalent to:
    - $X = \{P(1) = .5, P(Y) = .5\}$
    - $Y = \{P(2) = .4, P(3) = .6\}$

3. Uncertainty H cannot depend on the grouping of events for a random variable.

Information Theory

- Entropy: Formal reformulation of (1–3)
  - $H(p)$ is a real valued function of $P(1), P(2), \ldots P(N)$, with N the number of values for the random variable or length of *domain*, then
  1. $H(P(1), P(2), \ldots P(N))$ reaches a maximum if the distribution is uniform: $P(i) = 1/N, N = len(i), \forall\ i$.
  2. $H(P(1), P(2), \ldots P(N))$ is a continuous function of all $P(i)$'s.

Information Theory

- Entropy: Formal reformulation of (1–3)

3. Independence of subsets of probability groups: for N probabilities grouped into k subsets, w_k:

$$w_1 = \sum_{i=1}^{n_1} p_i; w_2 = \sum_{i=n_1+1}^{n_2} p_i; \ldots$$

Information Theory

- Entropy: Formal reformulation of (1–3)

3. Independence of subsets of probability groups: assumption

$$H[p] = H[w] + \sum_{j=1}^{k} w_j H[\{p_i/w_j\}_j]$$

- $\{p_i/w_j\}$ is: sum extends over p_i's that make up a particular w_j

Information Theory

- Entropy: Summary
  - Given the three requirements it follows that:

$$H[X] = k \sum_{x \in X} Pr(x) log Pr(x)$$

  - with k and arbitrary constant $[8, 40, 44]$. For $k = -1$ and log_2 the units are bit.

Information Theory

- Average Shannon Entropy: measured in bits

$$H[X] = -1 \sum_{x \in X} Pr(x) lg Pr(x)$$

$$H[X] = \sum_{x \in X} Pr(x) lg \frac{1}{Pr(x)}$$

Information Theory

- Average Shannon Entropy of one outcome: measured in bits

$$h[x] = Pr(x) lg \frac{1}{Pr(x)}$$

Joint Entropy

- For a pair of random variables: $X, Y \sim p(x, y)$

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

- $X = \{A = .4, B = .6\}$
- $Y = \{C = .2, D = .8\}$

Joint Entropy

- $X \wedge Y = \{AC = .4 \times .2, AD = .4 \times .8, BC = .6 \times .2, BD = .6 \times .8\}$
- $X \wedge Y = \{AC = .08, AD = .32, BC = .12, BD = .48\}$
- $Z = \{AC = .08, AD = .32, BC = .12, BD = .48\}$

Mutual Information

- Reduction of uncertainty of one random variable due to knowing about another.
- Amount of information one random variable contains about another.
- Symmetric, Non-negative
- $MI = 0$, if two random variables are independent
- MI is high, if two random variables are dependent, depending on their entropy.

Mutual Information

- MI over random variables!

→ Pointwise Mutual Information

  – Pointwise MI over selected values of random variables!

$$I(X;Y) = P(XY)lg\frac{P(XY)}{P(X)P(Y)}$$

- How many bits can we spare by storing two elements, rather than each single element alone?

Relative Entropy – KL Divergence

- Average number of bits that are wasted by encoding events from random variable X with a code based on random variable Y. How close are two pmf's?

$$D(y||x) = p(y)lg\frac{p(y)}{p(y|x)}$$

$$D(y||x) = p(y)lg\frac{p(y)}{\frac{p(xy)}{p(x)}} = p(y)lg\frac{p(y)p(x)}{p(xy)}$$

- How many bits more would we use by storing $< xy >$, rather than each single element alone?