Relative Entropy and Mutual Information

Gerald Penn

CSC 401
University of Toronto

http://www.cs.toronto.edu/~gpenn/csc401

Relative Entropy

Relative entropy, or *Kullback-Leibler divergence* measures the quality with which one distribution, q, approximates another distribution, p. In the discrete case:

$$D(p||q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)}$$

Some facts:

- $D(p||q) = 0$ iff $p = q$.
- $D(p||q) \geq 0$.
- D is not symmetric in p and q. This is why it's called a *divergence* — it's not a distance.

Resnik's Method

Let R be the set of all selectional relationships over functors and arguments.

- each relationship consists of a functor word and an argument synset.
- e.g., drink-BEVERAGE, elapse-INTERVAL

The *selectional preference strength* of w in R is:

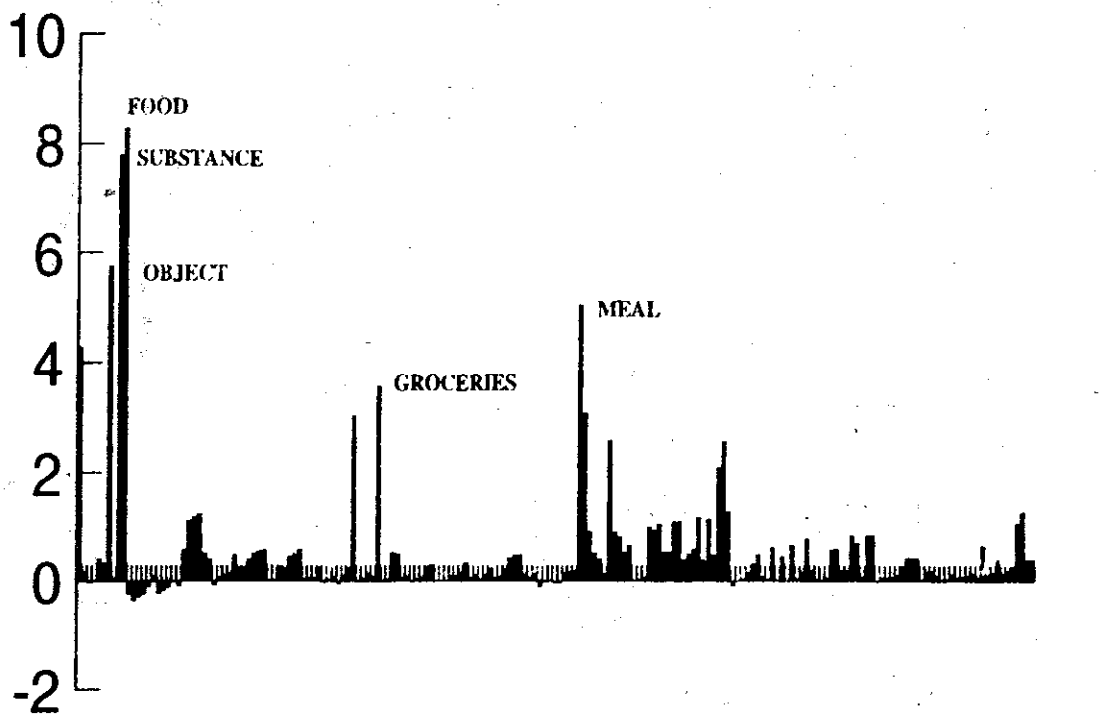
$$\begin{aligned} s_R(w) &= D(P(Syn|w)||P(Syn)) \\ &= \sum_{s \in \text{WordNet}} P(s|w) \log \frac{P(s|w)}{P(s)} \end{aligned}$$

The *selectional association strength* of w and sense s is then:

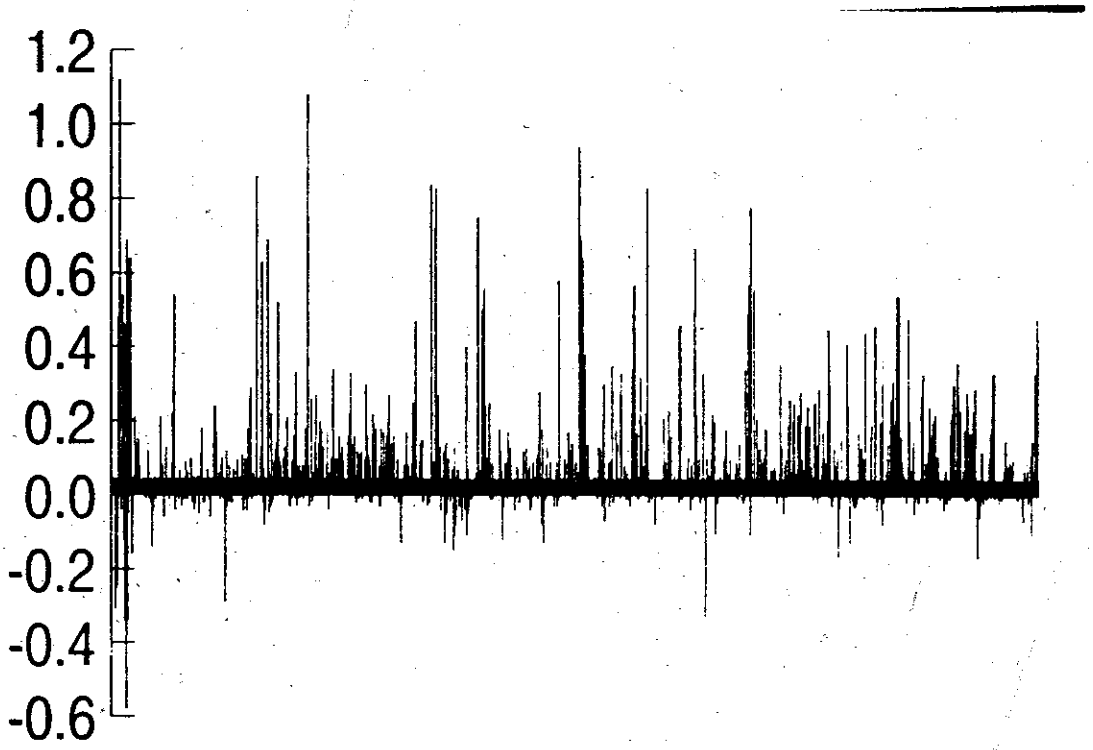
$$A_R(w,s) = \frac{1}{s_R(w)} P(s|w) \log \frac{P(s|w)}{P(s)}$$

Given sense-ambiguous argument word, w', with functor w, choose sense subsumed by $s' = \underset{s}{\text{argmax}}\, A_R(w,s)$.

eat-OBJ



find-OBJ



Performance of Resnik's method

Random choice:	$\sim 30\%$
Resnik's method:	$\sim 35 - 44\%$
Most frequent sense:	$\sim 58\%$

But on a type-by-type basis, it's actually better — frequent words are often polysemous (as Zipf told us)

Ambiguous word	Indicator	Examples: value → sense
prendre	object	*mesure* → *to take* *décision* → *to make*
vouloir	tense	present → *to want* conditional → *to like*
cent	word to the left	*per* → % number → *c.* [money]

Table 7.3 Highly informative indicators for three ambiguous French words.

senses of *drug* in the Hansard corpus. For example, *prices* is a good clue for the 'medication' sense. This means that $P(prices|\text{'medication'})$ is large and $P(prices|\text{'illicit substance'})$ is small and has the effect that a context of *drug* containing *prices* will have a higher score for 'medication' and a lower score for 'illegal substance' (as computed on line 14 in figure 7.1).

7.2.2 An information-theoretic approach

The Bayes classifier attempts to use information from all words in the context window to help in the disambiguation decision, at the cost of a somewhat unrealistic independence assumption. The information theoretic algorithm which we turn to now takes the opposite route. It tries to find a single contextual feature that reliably indicates which sense of the ambiguous word is being used. Some of Brown et al.'s (1991b) examples of indicators for French ambiguous words are listed in table 7.3. For the verb *prendre*, its object is a good indicator: *prendre une mesure* translates as *to take a measure*, *prendre une décision* as *to make a decision*. Similarly, the tense of the verb *vouloir* and the word immediately to the left of *cent* are good indicators for these two words as shown in table 7.3.

FLIP-FLOP ALGORITHM

In order to make good use of an informant, its values need to be categorized as to which sense they indicate, e.g., *mesure* indicates *to take*, *décision* indicates *to make*. Brown et al. use the *Flip-Flop algorithm* for this purpose. Let $t_1, \ldots, t_m$ be the translations of the ambiguous word, and $x_1, \ldots, x_n$ the possible values of the indicator. Figure 7.2 shows the Flip-Flop algorithm for this case. The version of the algorithm described here only disambiguates between two senses. See Brown et al. (1991a) for an extension to more than two senses. Recall the definition of mutual

Mutual Information

Given:

- 2 random variables, X, Y
- joint probability, $p(x, y)$
- marginal probabilities, $p_x(x)$ and $p_y(y)$

Then:

$$\begin{aligned} I(X;Y) &= \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p_x(x) p_y(y)} \\ &= D(p||p_x p_y) \end{aligned}$$

Example:

$p(x,y)$	1	2	3	4	p_x
1	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{32}$	$\frac{1}{4}$
2	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{32}$	$\frac{1}{32}$	$\frac{1}{4}$
3	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{4}$
4	$\frac{1}{4}$	0	0	0	$\frac{1}{4}$
p_y	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$	$I(X;Y) = \frac{3}{8}$

Useful because joint distribution is larger and harder to estimate on large outcome sizes.

The Flip-Flop Algorithm

Objective: find single "feature" of context that reliably indicates sense, using a bitext.
Given:

- $T = \{t_1, \ldots, t_m\}$: translations of ambiguous word, e.g., for *prendre*, {*take*, *make*, *rise*, *speak*}
- $X = \{x_1, \ldots, x_n\}$: possible values of feature, e.g., object of *prendre*: {*mesure*, *note*, *exemple*, *décision*, *parole*}

1. Randomly parition $T = P_1 \uplus P_2$. Let $P = \{P_1, P_2\}$.
2. Repeat:
   (a) Find $Q := \underset{\substack{X = Q_1 \uplus Q_2, \\ \hat{Q} = \{Q_1, Q_2\}}}{\text{argmax}} I(P; \hat{Q})$ [P fixed]
   (b) Find $P := \underset{\substack{T = P_1 \uplus P_2, \\ \hat{P} = \{P_1, P_2\}}}{\text{argmax}} I(\hat{P}; Q)$ [Q fixed]
   (c) Until $I(P; Q)$ stops improving (significantly)

Guaranteed to improve monotonically.

The Flip-Flop Algorithm

Solution to the *prendre* example:

- $P_1 = \{take\}, P_2 = \{make, rise, speak\}$
- $Q_1 = \{mesure, note, example\}, Q_2 = \{décision, parole\}$

Can then repeat over all possible features, e.g., object of *prendre*, subject of *prendre*, tense of *prendre* token, etc.

The problem with Flip-Flop is that it considers only one feature.

It's better to combine all of the evidence...