Probabilistic Models

- **Motivation**: People process language: rapidly, robustly, and accurately
- **SLCM**: Simple, robust account of lexical category disambiguation
- **Jurafsky**: Probabilistic parser that models a range of local ambiguities
- **Crocker & Brants**: High accuracy and fast performance with beam search
  - Also: Performance is maintained under assumptions of incremental processing and strict memory limitations
- **In common**: all models approximate a maximum likelihood function
- **Differences**: the underlying symbolic model (n-gram, cfg), and what units of structure are associated with statistical parameters.
Summary of Informativity

• Optimal function incorporates aspects of earlier models:
  • Basic cognitive limitations: serial interpretation + reanalysis
  • Maximising success of reaching correct interpretation
    \[
    P(\text{global success}) = \prod_{i=1}^{n} P(\text{success at } L_i) \quad I(H_i) = P(H_i) \cdot S(H_i) \quad S(H_i) = \frac{1}{P(\text{Confirm } H_i)}
    \]
  • Explains why people don’t always follow likelihood alone
  • Prefer to form testable (interpretable) dependencies
    • These can be evaluated as plausible, or trigger reanalysis quickly
  • Informativity is an idealisation of what the HSPM should approximate

Rational Models and Linking Hypotheses

• Rational Hypothesis 1: \[ \arg \max_{i} P(s_i) \text{ for all } s_i \in S \]
• Rational Hypothesis 2: \[ \arg \max_{i} P(s_i) \cdot S(s_i) \text{ for all } s_i \in S \]

• Implementing and evaluating more plausible “optimal functions”:
  • More linguistically informed probabilistic models (lexical, semantic ...)
  • Integration with non-probabilistic factors (recency, memory load)
  • Richer linking functions between parser and human processing measures
  • Relate the parsing mechanisms to observed processing difficulty, i.e. reading measures, event-related potentials, fMRI
Information Theoretic Approaches

• We can think of language as a communication system, in which information is transmitted from speaker to hearer.

• Rationality suggests that language, and language use, will be optimized to transmit information as efficiently as possible (speaker) while taking into account cognitive limitations of the hearer.

• The average amount of information conveyed by a linguistic unit.
  • Uncertainty of a random variable is measured by its entropy.

• Information Theory (Shannon).
  • Finding the best “code” for sending messages of a language.

Entropy

• Information: for a given language.
  • The number of bits needed to send a message, on average.

• Optimal code for an event having probability \( p(x) \) is: \[
\left\lfloor \log_2 \frac{1}{p(x)} \right\rfloor
\]

• The average number of bits needed to transmit a message in a language \( X \) is:
  • Entropy: \[
H(X) = \sum_{x \in X} p(x) \log_2 \frac{1}{p(x)}
\]
Example 1: 8-sided die

- Let $x$ represent the result of rolling a (fair) 8-sided die.

- Entropy:
  \[
  H(X) = \sum_{x \in X} p(x) \log_2 \frac{1}{p(x)}
  \]

\[
H(X) = \sum_{x \in \{001, 010, 011, 100, 101, 110, 111, 000\}} \frac{1}{8} \log_2 \frac{1}{\frac{1}{8}} = \log_2 8 = 3
\]

- The average length of the message required to transmit one of 8 equiprobable outcomes is 3 bits.

    “1” “2” “3” “4” “5” “6” “7” “8”
    001  010  011  100  101  110  111  000

Entropy of a Weighted Coin

- The more uncertain the result, the higher the entropy.

- Fair coin: $H(X) = 1.0$

- The more certain the result, the lower the entropy.

- Completely biased coin: $H(X) = 0.0$
Example 2: Simplified Polynesian

- Simplified Polynesian:

<table>
<thead>
<tr>
<th>P</th>
<th>T</th>
<th>K</th>
<th>A</th>
<th>I</th>
<th>U</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/8</td>
<td>1/4</td>
<td>1/8</td>
<td>1/4</td>
<td>1/8</td>
<td>1/8</td>
</tr>
</tbody>
</table>

\[
H(X) = -\sum_{x \in X} p(x) \log_2 p(x) \\
= -\left[ 4 \times \frac{1}{8} \log_2 \frac{1}{8} + 2 \times \frac{1}{4} \log_2 \frac{1}{4} \right] \\
= 2\frac{1}{2} \text{ bits}
\]

- Coding Tree:
Conditional Entropy

- In language, the likelihood of each “outcome” (word in the sequence), is not independent. Rather, it depends on the preceding outcomes.

- Conditional entropy: the amount of information needed to communicate Y, given that message X has been communicated.

\[
H(Y \mid X) = \sum_{x \in X} p(x)H(Y \mid X = x)
\]

\[
= \sum_{x \in X} p(x) \left[ -\sum_{y \in Y} p(y \mid x) \log_2 (y \mid x) \right]
\]

\[
= -\sum_{x \in X} \sum_{y \in Y} p(x, y) \log_2 p(y \mid x)
\]

Natural language lexica

- It’s long been known that word length correlates with word frequency: frequent words are generally shorter (Zipf).

- If lexica are optimized to take into account the likelihood of words in context, then average predictability should be a better predictor of word length:

\[
-\frac{1}{N} \sum_{i=1}^{N} \log P(W = w \mid C = c_i)
\]

Correlating Information with Length

- Higher average information content corresponds to greater length
- And is better predictor of length than unigram
Surprisal & Psycholinguistics

- In addition to measuring the average information for a language, we can of course measure the information conveyed by any given linguistic unit (e.g. phoneme, word, utterance) in context. This is often called surprisal:

\[
\text{Surprisal}(x) = \log_2 \frac{1}{P(x \mid \text{context})}
\]

- Surprisal will be high, when \( x \) has a low conditional probability, and low, when \( x \) has a high probability.

- Claim: Cognitive effort required to process a word is proportional to its surprisal (Hale, 2001).

Predictability & Integration

- Surprisal theory claims that expected words will be easier to process:
  - presumably due to their predictability or ease of integration

- This has broad empirical support from psycholinguistics, where Cloze probability (Taylor, 1953) correlate with reading times and N400 ERPs:
  - *My brother came inside to* ... chat? eat? play? rest?
  - *The children went outside to* ... chat? eat? play? rest?

- Evidence of anticipatory processing is also found in visual world experiments, where people look at the visual referents of words likely to be mentioned next:
  - *The boy will eat the* ... [more looks to cake, than other objects]
Computing Surprisal

\[ \text{Surprisal}_{k+1} = -\log P(w_{k+1} \mid w_1 \ldots w_k) \]

- There are various ways we can compute surprisal from different kinds of underlying probabilistic language models

- N-gram surprisal:

\[ \text{Surprisal}(w_{k+1}) = -\log_2 p(w_{k+1} \mid w_{k-2}, w_{k-1}, w_k) \]

Parse Surprisal

- We can also show how define surprisal in terms of the probabilities recovered by a probabilistic grammar/parser:

\[ \text{Surprisal}_{k+1} = -\log_2 P(w_{k+1} \mid w_1 \ldots w_k) \]

\[ = -\log_2 \frac{P(w_1 \ldots w_{k+1})}{P(w_1 \ldots w_k)} \]

\[ = \log_2 P(w_1 \ldots w_k) - \log_2 P(w_1 \ldots w_{k+1}) \]

\[ = \log_2 \sum_T P(T, w_1 \ldots w_k) - \log_2 \sum_T P(T, w_1 \ldots w_{k+1}) \]

\[ = \text{prefprob}_{w_k} - \text{prefprob}_{w_{k+1}} \]
Hale 2001

- Hale proposed that surprisal measures determined by an incremental probabilistic Earley parser offer a psychologically plausible index of effort.

\[
prefprob_{w_n} = -\log_2 \sum_T p(T \mid w_1 \ldots w_n)\\
\text{Surprisal}_{w_n} = \text{prefprob}_{w_{n-1}} - \text{prefprob}_{w_n}
\]

- When \textit{fell} is encountered, the higher probability parse is eliminated.

- This results in a large drop in the prefix probability as we process word \textit{n}.

\[
\begin{array}{c|c|c}
\text{prefprob}_{w_n} & \text{Surprisal}_{w_n} & \text{Example} \\
1.0 & S & \rightarrow \text{NP VP .} \\
0.876404494831 & \text{NP} & \rightarrow \text{DT NN} \\
0.123595505169 & \text{NP} & \rightarrow \text{NP VP} \\
1.0 & \text{PP} & \rightarrow \text{IN NP} \\
0.171428571172 & \text{VP} & \rightarrow \text{VBD PP} \\
0.752380952552 & \text{VP} & \rightarrow \text{VBN PP} \\
0.0761904762759 & \text{VP} & \rightarrow \text{VBD} \\
1.0 & \text{DT} & \rightarrow \text{the} \\
0.5 & \text{NN} & \rightarrow \text{horse} \\
0.5 & \text{NN} & \rightarrow \text{barn} \\
0.5 & \text{VBD} & \rightarrow \text{fell} \\
0.5 & \text{VBD} & \rightarrow \text{raced} \\
1.0 & \text{VBN} & \rightarrow \text{raced} \\
1.0 & \text{IN} & \rightarrow \text{past}
\end{array}
\]
Hale 2001: Results (toy)

- Hale proposed that surprisal measures determined by an incremental probabilistic Earley parser offer a psychologically plausible index of effort.

\[
prefprob_{w_n} = \log_2 \sum_T p(T \mid w_1 \ldots w_n)
\]

\[
\text{Surprisal}_{w_n} = prefprob_{w_n-1} - prefprob_{w_n}
\]
Hale proposed that surprisal measures determined by an incremental probabilistic Earley parser offer a psychologically plausible index of effort.

\[
\text{prefprob}_{w_n} = \log_2 \sum_T p(T \mid w_1 \ldots w_n)
\]

\[
\text{Surprisal}_{w_n} = \text{prefprob}_{w_n-1} - \text{prefprob}_{w_n}
\]

Unambiguous example

For example, it is well known that subject relative clauses are processed more easily than object relatives:

\begin{align*}
\text{The reporter who attacked the senator} & \less than \text{easier} \quad \text{The reporter who the senator attacked}
\end{align*}
Rational Communication

• Linguistic forms are being reduced/expanded at all linguistic levels

• Variation enables speakers to modulate the rate and linearization of message transmission
  
  • Evidence: Word length, speech, reading times

• Rational communication systems:
  
  • How is information communicated optimally?
  
  • Are speakers adapted to listeners constraints?

Hypotheses

• Rational language use is shaped by general information theoretic principles
  
  • There is an upper bound on the amount of information: Channel Capacity
  
  • Language users prefer to distribute information uniformly over a message
  
  • Variation in encoding serves to modulate information density

• Production choices are modulated by predictability:
  
  • Expand of high surprisal expressions, reduce predictable ones

![Graphs showing bad and good use of channel ID](image)
Information Density

\[ Information(event) = \log_2 \frac{1}{P(event)} \]

\[ = \log_2 \frac{1}{P(w_1)} + \log_2 \frac{1}{P(w_2 | w_1)} + \ldots + \log_2 \frac{1}{P(w_n | w_1 \ldots w_{n-1})} \]

- Uniform Information Density:
  - Maximizes information transmission
  - Avoids comprehender difficulty

Example: that-omission

- The complementizer “that” is optional in English:

  My boss confirmed (that) I am absolutely crazy.

- Uniform Information Density: Use of overt “that” increases with ID at onset of the CC “I ...”

Overt that

\[ = \log_2 \frac{1}{P(w_1 | CC, that, w_{n-1})} \]

Omitted that

\[ = \log_2 \frac{1}{P(CC | w_{n-1})} + \log_2 \frac{1}{P(w_1 | CC, w_{n-1})} \]

Jaeger, 2010
Example: *that*-omission

- N-gram estimates of ID predicted use of “that”

- Additionally, evidence that purely structural ID also predicts use of “that”

**Encoding and UID**

![Diagram](image)

$$\text{Utterance} = \arg\max_{\text{Enc}_i} UID(\text{Enc}_i)$$

Levy & Jaeger, 2007
Information Theoretic Approaches

• Information theory offers a (linguistic) theory neutral measure of the information conveyed by linguistic events: Surprisal

• The average surprisal of a word has been shown to correlate with word length, suggesting *lexica* have “evolved” towards an optimised encoding

  • predictable words are shorter

• Surprisal also offers a good index of on-line lexical and syntactic *comprehension* effort, both for ambiguous and unambiguous constructions (Hale, 2001; Levy, 2008).

• Finally, evidence suggests *speakers* may modulate surprisal to avoid peaks (and troughs) of information (UID: Levy & Jaeger, 2007).

Causal Bottleneck

• Surprisal Theory assumes difficulty is determined by a word’s predictability

  • Abstracts away from detailed representational or mechanistic accounts

  • Only depends on the quality of the conditional word probabilities

• If true, evidence regarding processing difficulty will shed little light on the nature of mental grammar

(a) Direct effect of representation on processing

(b) Surprisal as a causal bottleneck mediating effect of representation on processing
Information Theoretic Approaches

- The average surprisal of a word has been shown to correlate with word length, suggesting lexica have “evolved” towards an optimised encoding
  - Predictable words are shorter

- Surprisal of a word in a particular context is also a good predictor of online processing complexity: reading times, ERPs

- Some evidence that speakers seek to optimize production for listeners:
  - Uniform Information Density, avoiding peaks/troughs in Surprisal
  - Also evidence from word duration in speech, and lexical choice

- Communication strives for uniformity of information transmission