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.
Rational Models and Linking Hypotheses

• Rational Hypothesis 1: \( \arg \max_i P(s_i) \) for all \( s_i \in S \)

• Rational Hypothesis 2: \( \arg \max_i P(s_i) \cdot S(s_i) \) 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)

• Better linking functions between parser and human processing behaviour

• Better relate the parsing process to observed processing difficulty

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

• Entropy measures the uncertainty of a random variable

  • The average amount of information conveyed by a linguistic unit

• Information Theory

  • 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: \[ \log_2 \frac{1}{p(x)} \]
- 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: \( H(X) = \sum_{x \in X} p(x) \log_2 \frac{1}{p(x)} \)
  
  \[
  H(X) = \sum_{x \in X} \frac{1}{8} \log_2 \frac{1}{\frac{1}{8}} = \log_2 8 = 3
  \]
- The average length of the message required to transmit one of 8 equiprobable outcomes is 3 bits.

<table>
<thead>
<tr>
<th></th>
<th>“1”</th>
<th>“2”</th>
<th>“3”</th>
<th>“4”</th>
<th>“5”</th>
<th>“6”</th>
<th>“7”</th>
<th>“8”</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>001</td>
<td>010</td>
<td>011</td>
<td>100</td>
<td>101</td>
<td>110</td>
<td>111</td>
<td>000</td>
</tr>
</tbody>
</table>
Entropy of a Weighted Coin

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

- 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 Polynesia:

<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:

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

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:

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


Correlating Information with Length

- Higher average information content corresponds to greater length

- And is better predictor of length than unigram
Surprisal & Psycholinguistics

- Claim: Cognitive effort required to process a word is proportional to its surprisal.

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

- 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:

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

### Surprisal & Psycholinguistics

- Correlation with length

- Languages: Swedish, Spanish, Romanian, Portuguese, Polish, Italian, German, French, English, Dutch, Czech
Predictability & Integration

• Surprisal theory claims that predictable 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 prediction is also found in visual worlds experiments, where people looks a the visual referents of words likely to be mentioned next:
  • *The boy will eat the ... [more looks to cake, than other objects]*

Computing Surprisal

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

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

• N-gram surprisal:
  \[
  \text{Surprisal}(w_{k+1}) = -\log_2 p(w_{k+1} | w_{k-2}, w_{k-1}, w_k)
  \]
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 proposed that surprisal measures determined by an incremental probabilistic Earley parser offer a psychologically plausible index of effort.

\[
\text{prefprob}_{w_n} = - \log_2 \sum_T p(T \mid w_1 \ldots w_n) \\
\text{Surprisal}_{w_n} = \text{prefprob}_{w_{n-1}} - \text{prefprob}_{w_n}
\]
Hale 2001

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

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

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

- When *fell* is encountered, the higher probability parse is eliminated.

- This results in a large drop in the prefix probability as we process word \(n\)

Hale 2001: Results (toy)

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

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

\[
\text{Surprisal}_{w_n} = \text{prefprob}_{w_{n-1}} - \text{prefprob}_{w_n}
\]
Hale 2001: Results (Brown)

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

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

Surprisal_{w_t} = prefprob_{w_{n-1}} - prefprob_{w_n}

Figure 4: Mean 10.5

Figure 5: Mean: 16.44
Unambiguous example

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

![Graph showing easier processing of subject relative clauses compared to object relatives.]

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

Information Density

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

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

- 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 ...”

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

\[
\text{Omitted } that \quad = \log_2 \frac{1}{P(CC | w_{-1})} + \log_2 \frac{1}{P(w_1 | CC, w_{-1})}
\]

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**

\[ \text{Utterance} = \arg \max_{\text{encoding } i} \text{UID}(\text{Enc}_i) \]

---

**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 processing effort, both for ambiguous and unambiguous constructions
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

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 on-line 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