Predictive processing in sentence processing

Shravan Vasishth

August 2017

Introduction

- I will discuss the Linzen and Jaeger paper with slides and by referring to the paper.

Surprisal

Suppose we (somehow) compute the conditional probability of word n+1 given words 1. . . ,n: $Prob(w_{n+1} \mid w_1, \ldots, w_n)$.

One empirical method for doing this is by computing cloze probability. A computational method is to use a corpus of text. Here's a crude method:

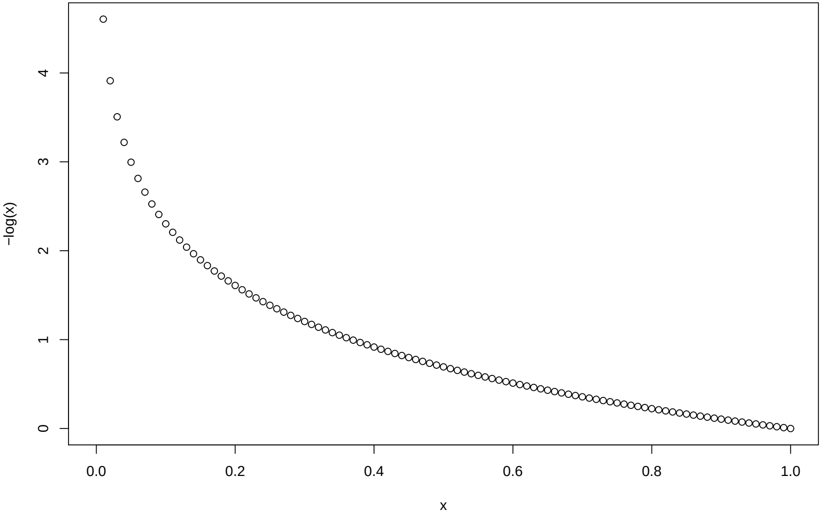
- List all sentences that begin with part of speech

Det

- List all possible continuations $i = 1, \ldots, k$, count number of cases n_i and calculate relative frequencies: $n_i / \sum_{i=1}^{k} n_i$. This gives you a bigram surprisal for each possible continuation.

Surprisal

```
x<-seq(0,1,by=0.01)
plot(x,-log(x))
```



Entropy

Entropy is the average length of a message, measured in bits, that would have to sent to describe a sample.

- Example 1: Consider a fair coin, with Prob(Heads)=0.5

In one trial, entropy is

```
-(0.5*log2(0.5)+0.5*log2(0.5))
```

```
## [1] 1
```

You need to transmit only one bit to express the outcome.

Entropy

Entropy is the average length of a message, measured in bits, that would have to sent to describe a sample.

- Example 2: Consider a two-headed coin, with Prob(Heads)=1

In one trial, entropy is 0:

```
-(1*log2(1))
```

```
## [1] 0
```

You need to transmit zero bits to express the outcome.

Entropy vs entropy reduction

Applying entropy to sentence processing

Suppose that only two continuations are possible after a determiner:

- $\text{Prob}(\text{N} \mid \text{Det}) = 0.9$
- $\text{Prob}(\text{Adjective} \mid \text{Det}) = 0.1$

[Note: these have to sum to 1.]

Then, entropy at the determiner is:

```
(ent_det<- -(0.9*log2(0.9)+0.1*log2(0.1)))
```

```
## [1] 0.4689956
```

Entropy and entropy reduction

Now, if the high probability continuation occurs, the next possible continuation might be

$\text{Prob(Verb}|\text{ Det N)} = 1$

The new entropy is:

```
## total certainty:
(ent_detN_CERTAIN<- -(1*log2(1)))
```

```
## [1] 0
```

We get a large entropy reduction:

```
ent_detN_CERTAIN - ent_det
```

```
## [1] -0.4689956
```

Entropy and entropy reduction

Consider an alternative scenario:

Suppose that, as before, only two continuations are possible after a determiner:

- Prob(N| Det) = 0.9
- Prob(Adjective | Det) = 0.1

Entropy and entropy reduction

But after seeing a noun, suppose there was a very likely continuation, vs an unlikely continuation:

-Prob(Verb| Det N) = 0.999

-Prob(Relative Pronoun | Det N) = 0.001

```
## low entropy:
(ent_detN_UNCERTAIN<- -(0.999*log2(0.999)+0.001*log2(0.001)
```

```
## [1] 0.01140776
```

There is a smaller reduction in uncertainty at the Noun:

```
ent_detN_UNCERTAIN - ent_det
```

```
## [1] -0.4575878
```

Surprisal vs Entropy Reduction

Surprisal is all about how surprised we are *at the current point in the sentence*

Entropy reduction is also about our uncertainty at the current point but can be computed (i) with respect to entropy at the current point (call it ER1) or (ii) with respect to entropy relating to what the *rest* of the sentence looks like (ER2).

There is an interesting theoretical question here: *how much forward prediction do we really do*? In other words, Is ER1 or ER2 (my terms) the correct metric?

Linzen and Jaeger 2015

Verbatim quote from the abstract:

- In a self-paced reading study, we use lexical subcategorization distributions to factorially manipulate both the strength of expectations and the uncertainty about them.
- We compare two types of uncertainty: uncertainty about the verb's complement, reflecting the next prediction step; and uncertainty about the full sentence, reflecting an unbounded number of prediction steps.
- We find that uncertainty about the full structure, but not about the next step, was a significant predictor of processing difficulty: Greater reduction in uncertainty was correlated with increased reading times (RTs).
- We additionally replicated previously observed effects of expectation violation (surprisal), orthogonal to the effect of uncertainty. This suggests that both surprisal and uncertainty affect human RTs.

Linzen and Jaeger 2015

Example of uncertainty calculation

He accepted

possible continuations: 80% NP 20% S

Entropy:

```
-(0.8 * log2(0.8) + 0.2 * log2(0.2))
```

```
## [1] 0.7219281
```

Linzen and Jaeger 2015

He forgot

possible continuations: 55% NP 9% S 18% PP 14% Inf

```
-(0.55 * log2(0.55) + 0.09 * log2(0.09)+
    0.18 * log2(0.18)+0.14 * log2(0.14))
```

```
## [1] 1.629445
```

Linzen and Jaeger's question: Does the difference in entropy between *accepted* and *forgot* lead to differences in processing difficulty?