Bias in the cloze task

Nathaniel J. Smith <njsmith@cogsci.ucsd.edu>
UC San Diego

Roger Levy <rlevy@ling.ucsd.edu>
UC San Diego

What does the cloze task actually measure?

Studies often use the cloze task (Taylor, 1953) to estimate the predictability of a continuation given some context. This takes advantage of participants' rich knowledge of their language. But, it also assumes that participants' probability match, that is, that they produce responses by sampling from their subjective distribution over potential continuations.

Is this assumption justified, or do the peculiarities of the cloze task produce distortions? We compare experimentally cloze responses to "gold standard" corpus-based probability distributions, and find that many psycholinguistic properties bias cloze responses. This produces the potential for spurious results in studies which use cloze to estimate predictability.

What do we want to know about cloze responses?

We want to know:
1. How well do cloze responses predict the probability of the true language?
2. How do computational models of language predict the probability of the true language?
3. How do we estimate these probabilities in our experiments?

What errors - if any - do cloze task demands introduce? To answer, we compare cloze responses directly to rich, unsmoothed, corpus statistics.

Results

- Linear space
  - $R^2 = 0.34$

- Log space
  - $R^2 = 0.27$

Methods

Stimuli

4-word sentence initial stems from the Google Web 1-T corpus of 5-grams, with 5000 entries. Enough occurrences in the corpus that no common sense assumptions were needed.

Participants

24 native English speakers from UCSD subject pool.

Analysis

We examine two types of effects:

- An overall shift in probability mass, the degree to which may be affected by stem:
  - Bias towards or away from individual words based on their psycholinguistic properties.

- Without AoA and With AoA
  - Attractiveness
  - Familiarity
  - Concreteness
  - Imageability
  - Word length
  - Contextual diversity
  - Frequency
  - Age of Acquisition

Discussion

We find that cloze probabilities differ greatly from web corpus probabilities, and that these differences correlate with common psycholinguistic variables.

Some results here may be artifacts of using web-derived stimuli, but suppose some are correct. Then:
- If participants in a reading study are sensitive to true probability...
- But we control for cloze, rather than true probability...
- Then the above variables will correct our cloze estimates and appear to matter...
- Even if readers don’t care about them at all.

(Or, if readers are sensitive to cloze, then corpus probabilities have this problem.)

Next step: use these stimuli in a self-paced reading study, to check whether corpus or cloze probability better predicts reading time.