Surprisal as a Predictor of Essay Quality

Gaurav Kharkwal[1], Smaranda Muresan[2]

gaurav.kharkwal@gmail.com, smara@ccls.columbia.edu

1. Department of Psychology and Center for Cognitive Science, Rutgers University, New Brunswick
2. Center for Computational Learning Systems, Columbia University

Summary

- We bring together work from psycholinguistics and NLP.
- Through corpora studies, we examine the relation between sentence processing complexity and essay quality.
- Essays of greater overall complexity tend to have lower scores, and vice versa.

Surprisal Theory

- ***Surprisal*** is a psycholinguistic model of sentence processing complexity (Hale, 2001; Levy, 2008).
- Word-level processing cost estimated as negative log-prob of word given preceding context:

$$Surp(w_i) \propto -\log P(w_i | w_{1\ldots i-1}, \text{CONTEXT})$$

Computing Surprisal

- We used a top-down parser trained on WSJ corpus (Roark, 2009), which provided three measures:
  - **Syntactic surprisal**: unexpectedness of POS cat of word given sentential context.
  - **Lexical surprisal**: unexpectedness of word given sentential context and POS cat.
  - **Total surprisal**: sum of Syntactic and Lexical.

Experiment 1

Introduction

- Investigated whether EFL training improves essay quality, using essays written by EFL students across various terms.
- Examined whether essays' surprisal values decrease after training.

Corpus

- Uppsala Student English corpus (Axelsson, 2000).
  - 1,489 essays written by 440 EFL students.
- 116 essays were randomly selected:
  - 38 pairs on topic *Analysis*
  - 20 pairs on topic *Argumentation*
  - Each pair written by the same student across 2 terms

Methods

- Computed surprisal values using Roark's parser.
- Evaluated group mean differences across the two terms using linear mixed-effects regression models for the two topics:

$$Surp \sim Term + (1|Subject)$$

Results and Discussion

Topic	Term	Mean (Syn)	SD (Syn)	Mean (Lex)	SD (Lex)	Mean (Total)	SD (Total)
Analysis	Term1	2.37	1.86	3.97	3.24	6.34	3.32
	Term2	2.34	1.85	3.94	3.23	6.28	3.30
Argumentation	Term1	2.34	1.85	3.90	3.23	6.24	3.29
	Term2	2.28	1.85	3.87	3.24	6.15	3.36

- Despite trends, no consistent indication of an effect of EFL training on essays' surprisal scores.
- Absence of essay scores prevented direct evaluation of the link between surprisal and essay quality.

Experiment 2

Introduction

- Directly investigated link between surprisal and essay quality using a pre-scored set of essays.
- Evaluated whether surprisal values are correlated with essays' scores.

Corpus

- ETS's corpus used for NLI (Blanchard, et al, 2013).
  - 12,100 essays on 8 topics scored as *High*, *Medium*, or *Low*.
- 3,975 essays were randomly selected:
  - 1,325 per score category.

Methods

- Computed surprisal values as before.
- Performed correlation tests and group mean evaluations using a linear mixed-effects model:

$$Surp \sim EssayScore + (1|Topic)$$

Results

Score	Mean (Syn)	SD (Syn)	Mean (Lex)	SD (Lex)	Mean (Total)	SD (Total)
Low	2.46	.22	3.76	.29	6.22	.39
Medium	2.35	.17	3.75	.26	6.10	.34
High	2.27	.14	3.82	.24	6.09	.28

Surprisal Measure	ρ	t-value	p-value
Syntactic	$-.39$	-26.53	$< .001$
Lexical	$.08$	5.35	$< .001$
Total	$-.15$	-9.87	$< .001$

- Although all measures were found to be correlated, only Syntactic Surprisal had a high correlation coeff.

Conclusion

- Inverse relation between surprisal values and essay scores, with Syntactic Surprisal most promising.

Future Work

- How do findings vary across different datasets?
- Does greater processing complexity cause lower essay score?
- How important is training corpus used for computing surprisal?