Psycholinguistic models of sentence processing improve sentence readability ranking

David M. Howcroft[1] Vera Demberg[1,2]
Saarland Informatics Campus, Saarland University, Germany

Motivation

- Most work on readability to date has focused on *document level* measures of text difficulty.
- Work in natural language generation in general, and on automatic text simplification in particular, requires a notion of *sentence level* readability.

Question

Can psycholinguistic theories of on-line (human) sentence processing be leveraged to rank sentences by their 'difficulty'?

Corpora

English – Simple English Wikipedia Corpus (ESEW)

- ≈120k pairs of English and Simple English sentences
- noisy due to inter-author variation with respect to notions of 'simplicity'

One Stop English Corpus (OSE)

- ≈1,500 triples of English sentences at 3 levels:
  - *Elementary*, *Intermediate*, and *Advanced*
- less noisy – professionally edited

Psycholinguistic Theories of Sentence Processing

Surprisal (Hale 2001; Levy 2008)

- a.k.a. Shannon information
- measures the unpredictability of a word in context

Embedding Depth and Difference (van Schijndel et al. 2012)

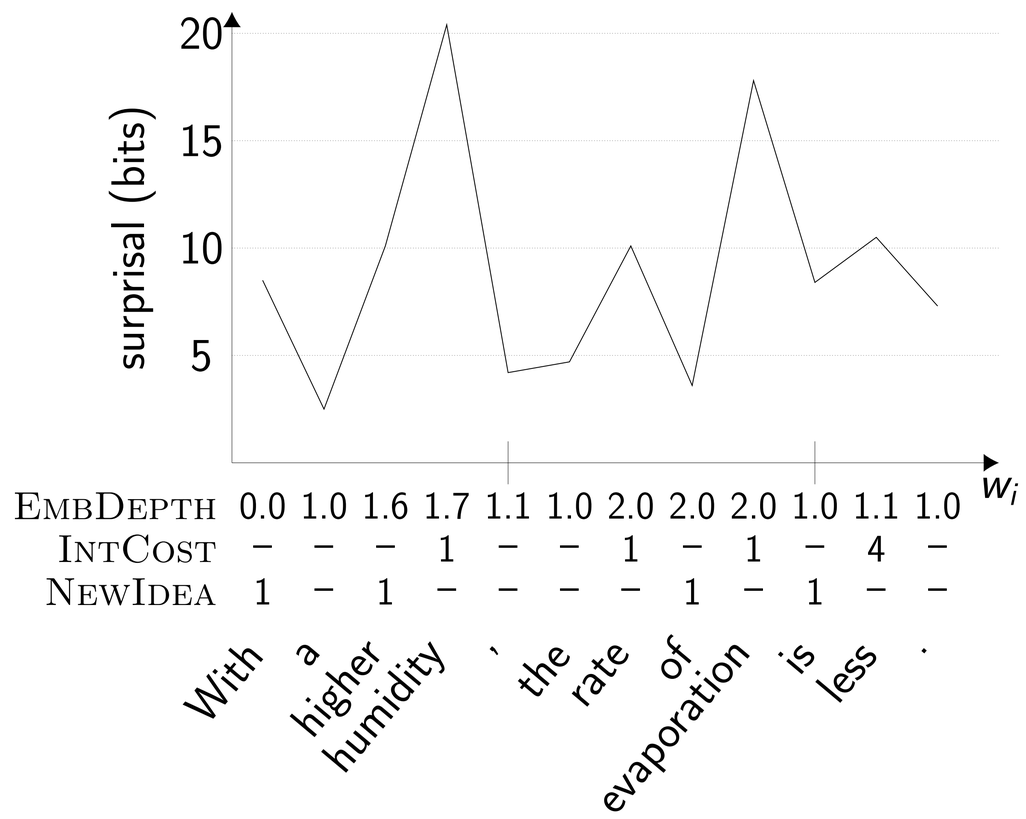
- estimates the amount of memory required to parse the sentence

Integration Cost (Gibson 1998, 2000)

- estimates the difficulty of integrating a new discourse referent

Idea Density (Kintsch 1972; Kintsch & Keenan 1973)

- estimates the number of propositions being expressed
- approximated by proportion of words which are adjectives, verbs, adverbs, and propositions



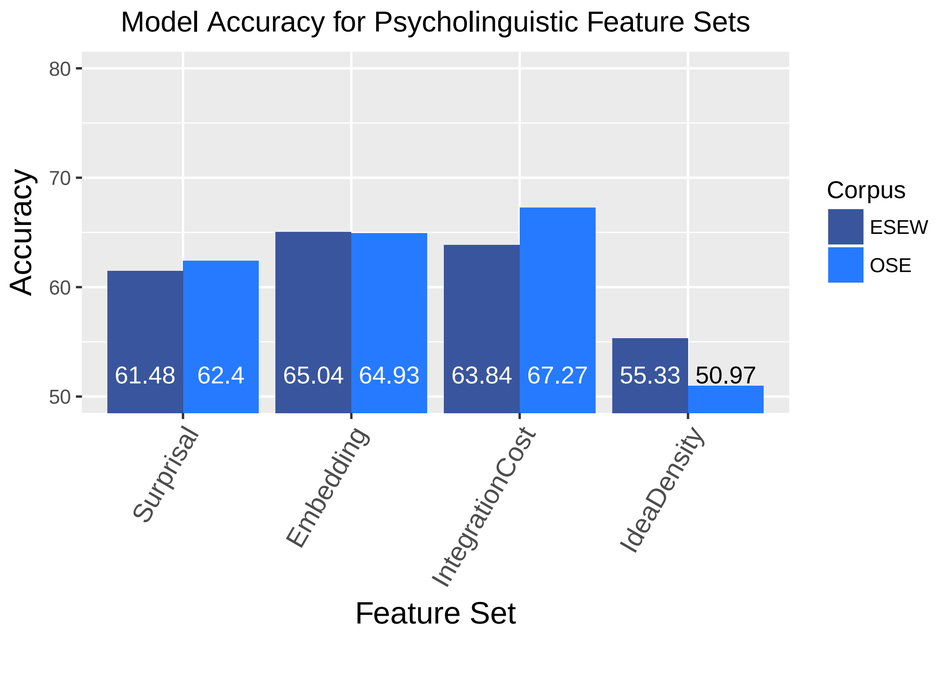
Features and Models

Feature Sets

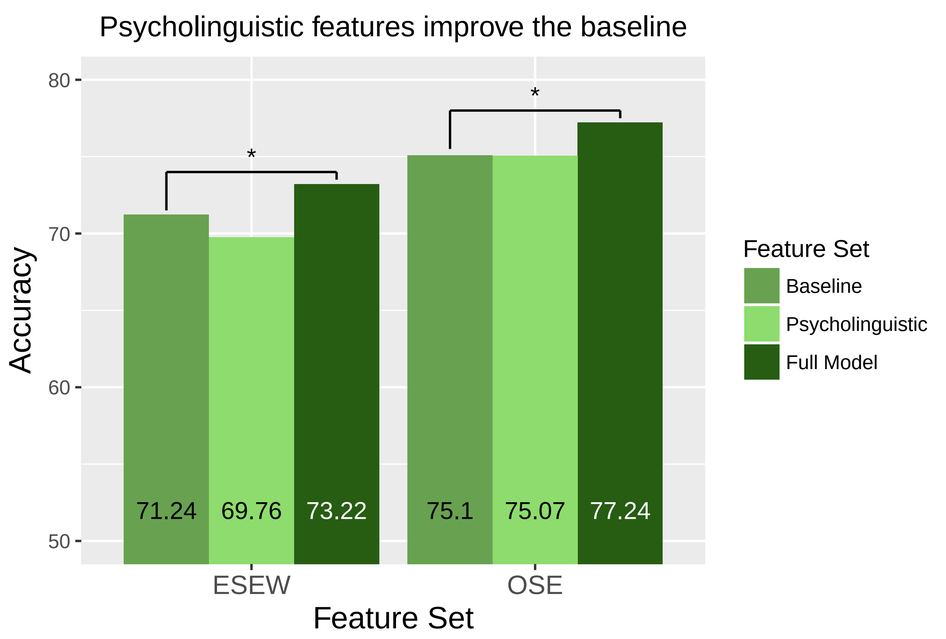
- Surprisal: avg. and max. lexical and syntactic surprisal
- Embedding: avg. and max. embedding depth and difference
- Integration Cost: avg. and max. integration cost
- Idea Density: avg. number of propositions per word
- Baseline: word length and sentence length
- Psycholinguistic: combines surprisal, embedding, integration cost, and idea density features
- Full Model: combines baseline and psycholinguistic models

Averaged Perceptron Model

- ranking treated as classification of difference features
- chance performance = 50



Results



Overall Results

- Individual and combined psycholinguistic features perform worse than the baseline
- Combined model with baseline and psycholinguistic features outperforms baseline by ≈2 percentage points.

Conclusion

Psycholinguistic features such as surprisal and embedding depth can improve performance on a readability ranking task.

[1]Department of Language Science and Technology
[2]Department of Computer Science
{howcroft,vera}@coli.uni-saarland.de