Some of Them Can be Guessed!
Exploring the Effect of Linguistic Context in Predicting Quantifiers

Sandro Pezzelle¹, Shane Steinert-Threlkeld², Raffaella Bernardi¹,³, Jakub Szymanik²

¹,³CIMEC - Center for Mind/Brain Sciences, ³DISI, University of Trento
²ILLC - Institute for Logic, Language and Computation, University of Amsterdam
sandro.pezzelle@unitn.it | quantit-clic.github.io

Motivation

Quantifiers (‘few’, ‘some’, ‘all’) are interesting because:

- They are typically considered as function words (as opposed to nouns, verbs, etc.), but they have a rich semantics
- They are of central importance in linguistic semantics and its interface with cognitive science [1,2,3]
- Their choice depends on both local and global context [4]
- Larger contexts are claimed to be detrimental for the prediction of function words in cloze test [5]

Hypotheses

- Human performance boosted by more context (proportional Qs)
- Models (very) effective with local context, hurt by broader context

Task & Datasets

Cloze test

<qnt> the island’s breeding birds are endemic.

The island is one of the world’s most biologically diverse areas, with many endemic species.

<qnt> the island’s breeding birds are endemic.

Other endemic species include the red-bellied lemur, the indri, and the aye-aye.

Datasets

1-Sent 10350 target sentences (quantifier+of at beginning): <s,>

3-Sent 10350 preceding + s + following: <s, s, s,>

Human Evaluation

Crowdsourcing

- Two experiments, one per condition (1-Sent, 3-Sent)
- 506 examples from validation set (same in two conditions)
- 3 judgments/datapoint; correctly-guessed w/ agreement > 0.66
- Higher accuracy in 3-Sent (0.258) compared to 1-Sent (0.221)

Models & Results

Models

8 models tested: 3 BoW baselines, 1 CNN, 4 LSTMs
2 conditions: 1-Sent, 3-Sent
Data: 80% train, 10% val, 10% test

Humans vs Models

Discussion & References

Discussion

Humans do better w/ broader contexts especially on proportional Qs; models suffer due to their inability to handle longer sequences

Models capitalize more on lexical cues compared to humans; 41% cases in 3-Sent (hum. 24%) and 50% cases in 1-Sent (hum. 44%)

References