Which distributional cues help the most?

Unsupervised context selection for lexical category acquisition

Giovanni Cassani, Robert Grimm, Walter Daelemans, Steven Gillis
name.surname@uantwerpen.be

1. Introduction

Distributional bootstrapping hypothesizes that children start grouping words into lexical categories using patterns of co-occurrences. In the acquisition literature, computational models have been used to test this hypothesis and assess the effectiveness of a handful of different cues, most notably:

- ***frequent frames (FF)*** [1]: 45 most-frequent A_X_B trigrams.
- ***flexible frames (ff)*** [2]: 45 most-frequent words, used as left and right bigrams that can be combined on the fly to provide frame-like information

However, they both display some problems:
*__arbitrariness__: what is frequent? why only a specific type of cue?
*__poor scalability__: frequent contexts may always occur with the same word
*__category bias__: in English, FF occur with more verbs than nouns
*__low coverage__: few types occur in FF
*__biased evaluation__: train and test on on the same data, with serious risk of overfitting

2. Model

Beyond token frequency, we suggest other distributional features of words - that children track - may play a role, including type frequency (number of different words a cue occurs with) and association strength (how predictable is the cue given the word).

$$token\_F = \frac{log_2(count(c_i))}{avg(log_2(count(c)))} \quad (1)$$

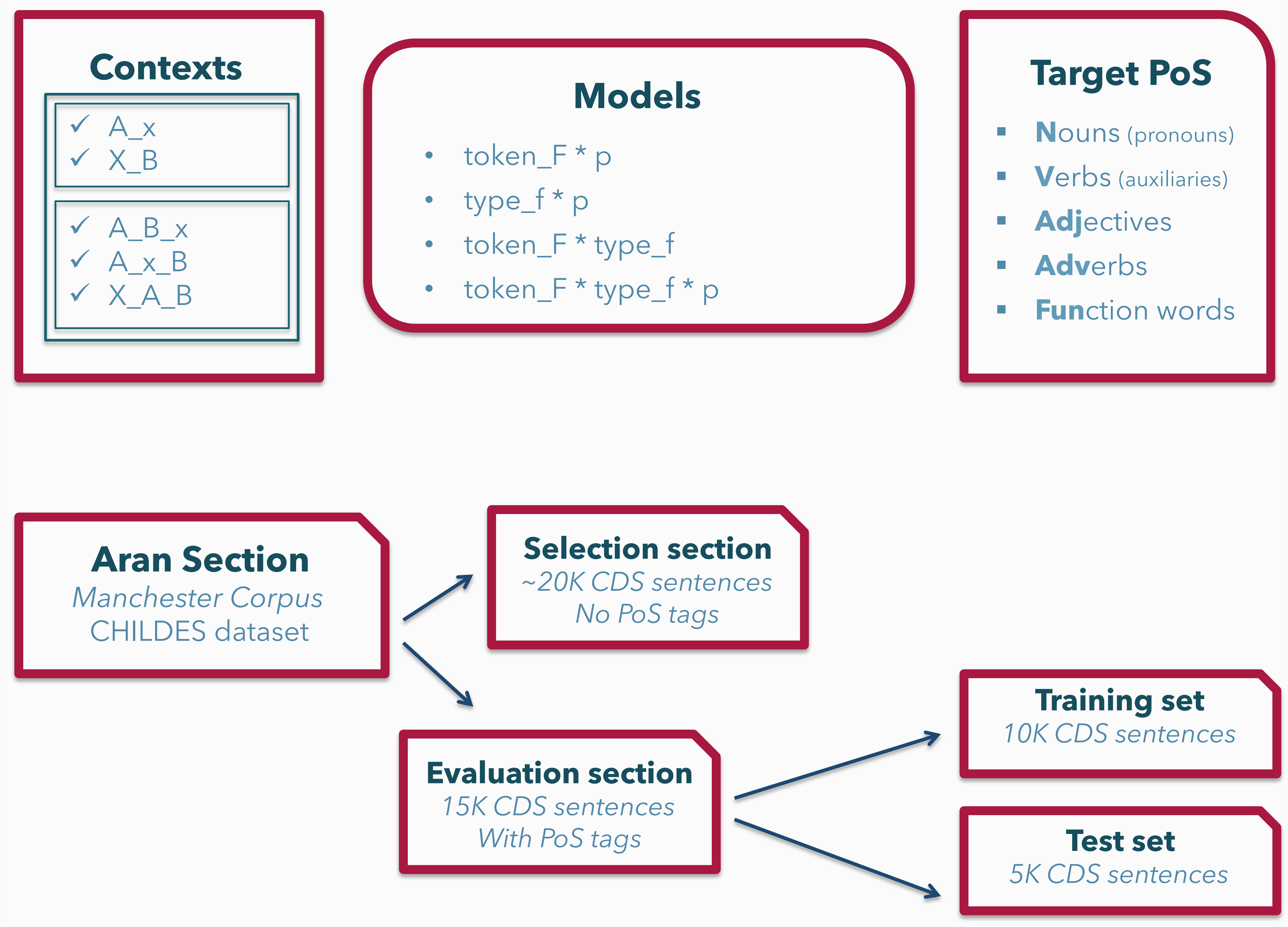
$$type\_f = \frac{log_2(\|W_{ci}\|)}{avg(log_2(\|W_c\|))} \quad (2)$$

$$p = \frac{1}{\|W_{ci}\|} \sum_{j=1}^{\|W_{ci}\|} \frac{log_2(count(w_j, c_i))}{log_2(count(w_j))} \quad (3)$$

$$score = token\_F \cdot type\_f \cdot p \quad (4)$$

A context is salient if $score > 1$.
Raw counts are log-transformed since every new occurrence is a little less important and to emphasize the search for structure: hapaxes have log 0 and are not considered.

3. Experimental setting



We evaluate performance on **5 dimensions**:
*__number of selected contexts__: more parsimonious sets make search faster
*__number of useless contexts__: how many of the selected contexts don't appear or only occur with one word in the training set
*__coverage__: how many types from the training set occur with the selected contexts
*__number of hits__: number of correctly categorized types in the test set
*__accuracy__: micro-F1 score of a supervised PoS experiment

4. Results

Context type	*# contexts*	*Useless*	*Missed words (%)*	*Hits*	*Acc.*
frequent frames	45	3 (6.7%)	83.7	290	**.83**
flexible frames	90	0	16.6	1405	.66
		$p \cdot token\_F$			
2grams	75	0	10.2	1559	.671
3grams	348	13 (3.7%)	37.3	1073	.681
all	490	11 (2.2%)	3.8	**1669**	.664
		$p \cdot type\_f$			
2grams	21	0	19.5	1377	.674
3grams	42	0	56.7	788	.756
all	97	0	8.7	1611	.679
		$p \cdot token\_F \cdot type\_f$			
2grams	211	0	2.6	1624	.641
3grams	659	7 (1%)	25.5	1249	.653
all	964	8 (0.8%)	**1.2**	1562	.609

Table 1: Evaluation of several sets of distributional cues, with baselines at the top and our models grouped according to the included pieces of information.
Column 1 specifies the type of context used
Column 2 shows the number of salient contexts
Column 3 shows how many of them could not be used for categorization
Column 4 provides the percentage of words from the training set (total = 3191) that could not be categorized by the contexts.
Column 5 gives the raw number of hits (test set = 2600 words)
Column 6 shows accuracy on supervised PoS tagging.

*The model including *Token_F* and *type_f* only is not shown since results were markedly worse than all other models, on all dimensions except for coverage.

5. Conclusions & future work

There is a **trade-off** between coverage, accuracy, and scalability: evaluating on one dimension without considering interactions is likely to lead to biased inferences.
Type frequency seems to be better than token frequency, because it ensures that a cue is *systematic* and not idiosyncratic.
Currently, we are
(i) evaluating models on more corpora from **typologically different languages**
(ii) evaluating **learning curves**
(ii) testing models on **core vocabulary**
(iv) training models on core vocabulary, to evaluate **generalization**

A. References

[1] Toben H. Mintz. Frequent frames as a cue for grammatical categories in child-directed speech. *Cognition*, 90(1):91–117, 2003.

[2] Michelle C St Clair, Padraic Monaghan, and Morten H Christiansen. Learning grammatical categories from distributional cues: Flexible frames for language acquisition. *Cognition*, 116(3):341–360, 2010.

B. Acknowledgements

The presented research was supported by a BOF/TOP grant (ID 29072) of the Research Council of the University of Antwerp.

The poster was designed on Overleaf with the fundamental help of Chris Emmery and is based onto the template developed by Brian Amberg.