Bootstrapping for Syntactic Categories

John Hewitt Jordan Kodner Mitch Marcus Charles Yang

Department of Linguistics and Computer and Information Science, University of Pennsylvania

Problem

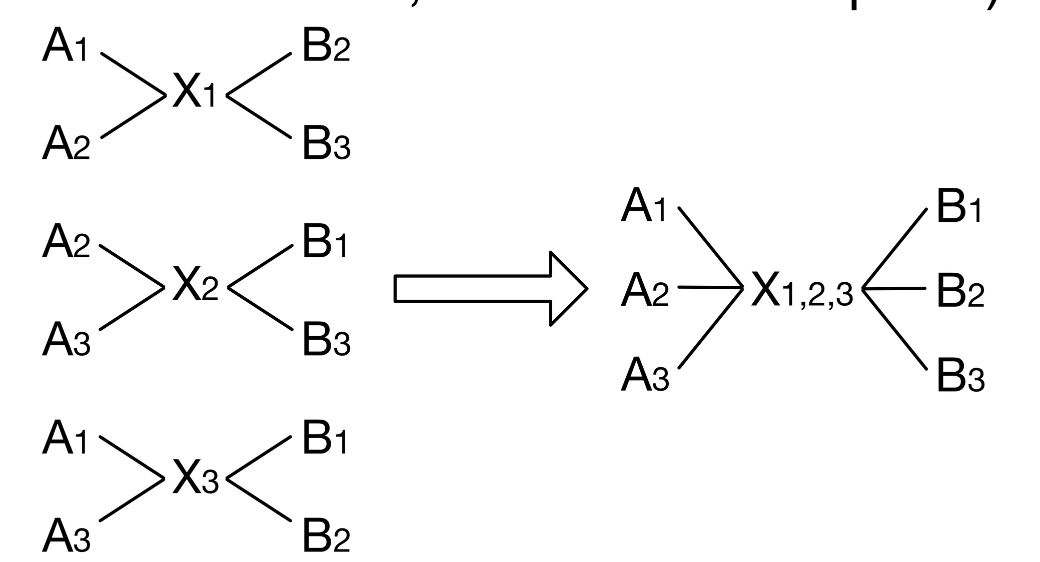
- Children learn syntactic categories very early and use them well (Valian 1986, Yang 2013)
- Whether categories are innate (Chomsky 1965) or emergent (Chomsky 1955), they must be acquired from language specific distributional information.
- Popular methods in cognitive modeling (e.g., Redington et al. 1996) and unsupervised NLP (e.g., Haghighi & Klein 2006) are computationally expensive yet still perform poorly
- Simple and empirically motivated methods (Mintz 2000) do not scale well (Chemla et al. 2009)
- Need an effective model with better connections to language development.

Bootstrapping for Categories

- **Semantic bootstrapping**: prototypical exemplars for syntactic categories (Grimshaw 1979, Pinker 1984)
- Strongly supported by analysis of child-directed input (Rondal & Cession 1990)
- **Distributional learning**: young children can use distributional regularities, including word order and morphology, to identify the categories of novel words (Brown 1957, Naigels 1990, Mintz 2002, Shi & Melacon 2010)
- **Proposal**: starting from a small set of labeled words, iteratively construct classifiers on the basis of **distributional distance** with the seeds
- Seed set increases and classifiers go from concrete to abstract

Method

From word frames to category frames (Reeder et al. 2013, Schuler et al. in press)

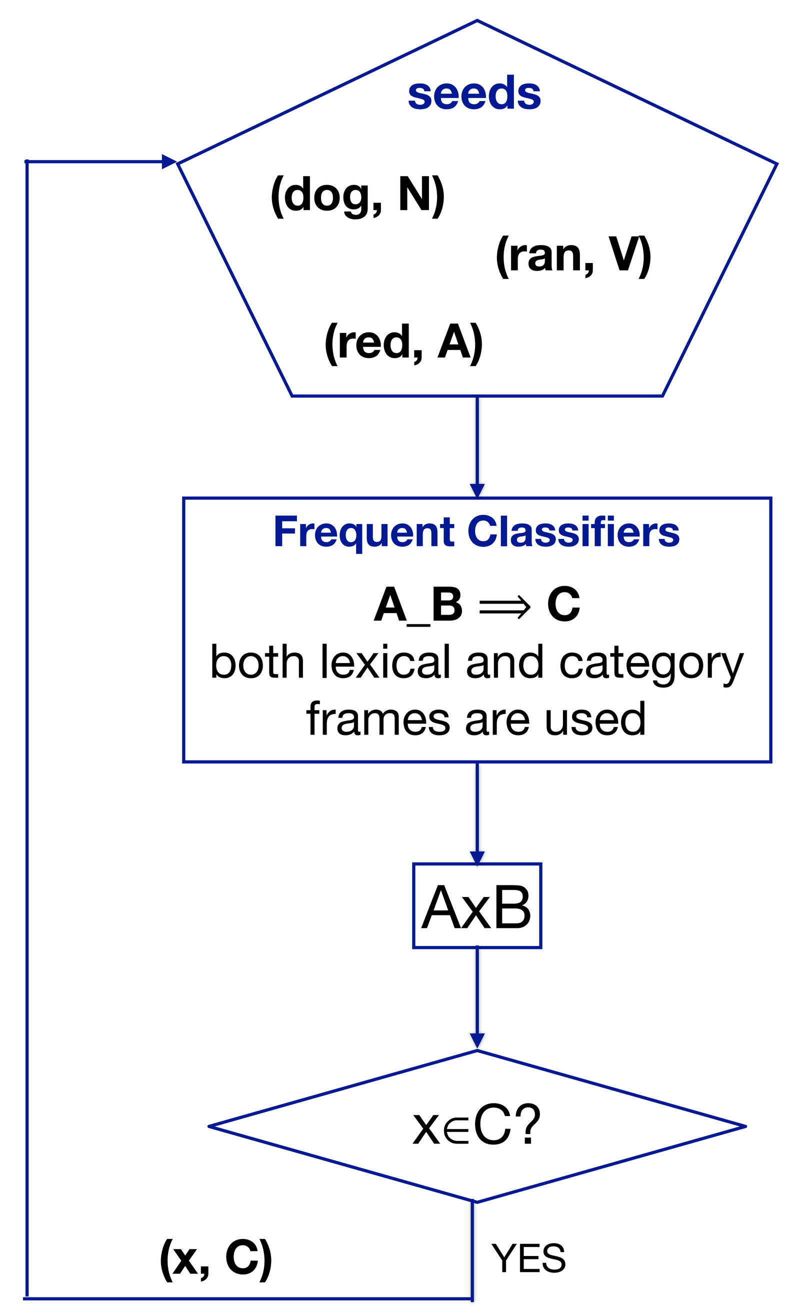


lexical and **category** **frames**

the_N_is: (**the_is**, **D_is**, **the_V**, **D_V**)$\Rightarrow$N

Two frame-based distance metrics

Identity	KL-clusters
w$\in$C, AwB, AxB $\Rightarrow$x$\in$C	KL_{AB}(w, x) clusterers $\Rightarrow$x$\in$argmin(x, C)
online with developmental support	offline still efficient (quadratic time)



Results

All results measured by 1-to-1 accuracy compared with baseline (initial seeds only)

Exp 1: Child-directed English

- 86 salient seed words from the Chicago early language corpus (e.g., selected **this**, not **the**, as a determiner seed)
- 220,000 words mapped to 7 categories
- 25-fold cross-validation
- Accuracy **62.44%** (baseline: **20.5%**)

Exp 2: WSJ/Chinese Treebanks

- Haghighi & Klein (2006): large word vectors
- Top 3 words per category (WSJ 45, Chinese 33): current model uses only text

	HK2006	Current	CRF
WSJ	68.8%	55.6%	41.3%
CTB	39.0%	46.7%	34.4%

Exp 3: KL-clusters

- A range of languages (top 1000 words only)
- Hierarchical clustering (Parkes et al. 1998)
- Scoring by types

	# seeds	Baseline	Accuracy
CHILDES	100	10.0%	70.4%
WSJ	89	8.9%	76.2%
CTB	85	8.5%	73.1%
German	35	3.5%	52.9%
Indonesian	34	3.4%	71.8%
Spanish	34	3.4%	62.8%

Conclusion

- Cognitive and perceptual cues (a la semantic bootstrapping) can be effectively combined with distributional linguistic cues (a la lexical and category frames)
- Robustness of results with respect to seed selection (few seeds, ambiguity allowed)
- The online learning algorithm leads to developmental predictions and can incorporate other mechanisms of word learning (e.g., Lederer et al. 1999, Stevens et al. 2016)
- Future work: incorporation of structural information (e.g., morphology, non-local contexts, abstract syntax)

Selected References

Chemla et al. (2009) Developmental Science. **Haghighi & Klein (2006)** EMNLP. **Lederer et al. (1999)** Cognition. **Mintz (2002)** Cognition. **Naigels (1990)** Language Acquisition. **Parkes, Malek, & Marcus (1998)** ACL. **Redington, Chater, & Finch (1996)** Cognitive Science. **Reeder, Newport, & Aslin (2013)** Cognitive Psychology. **Rondal & Cession (1990)** J. Child Language. **Schuler et al. (in press)** Language Learning and Development **Shi & Melacon (2010)** Infancy. **Stevens et al. (2016)** Cognitive Science. **Valian (1986)** Cognitive Psychology. **Yang (2013)** PNAS.

Acknowledgements

Partially funded by the DARPA LORELEI program under Agreement No. HR0011-15-2-0023.