

A computational model of the simultaneous learning of grammatical structures and statistics

Statistical learning for language is extensively studied in the artificial language learning paradigm (Gomez, 2002; Thompson & Newport, 2007; Wonnacott, Newport & Tanenhaus, 2008). Implicit in these studies is the learning of the appropriate ‘units’ on which to keep statistics. Extending on Chang (2008), this work presents a unified computational model in which both the grammatical units and usage statistics are learned simultaneously from naturalistic, contextually-grounded Mandarin Chinese input.

Using construction grammar (Goldberg, 1995; Fillmore & Kay, 1999; Bergen & Chang, 2005), this model learns argument structure constructions in a language with argument omission, assuming as pre-existing knowledge words for objects and actions, and embodied knowledge about motor programs and participants associated with each action. The input consists of dialogues taken from the CHILDES Tardif Beijing corpus (Tardif, 1993; MacWhinney, 2000). Based on transcribed dialogues, the author coded for events that can reasonably occur prior to each utterance. Each utterance is additionally assigned a rudimentary speech act (e.g. *explaining, requesting-action, admonishing*).

The events preceding each child-directed utterance make up the situational context for interpreting and learning from the utterance. For example, encountering the utterance *XiXi1 chi1 yao4* (Xixi eat medicine) while the father gestures towards some cough syrup with a requesting tone allows the model to guess that Xixi is asked to take the medicine. The model then creates new *structures* in the grammar by hypothesizing a new, lexicalized, phrasal construction, *XIIXII-CHII-YAO4*. This construction specifies strict ordering constraints between its constituents and denotes a eating process in which the eater is a child supplied by its first constituent, *XIIXII*, and the eatee is some medicine supplied by its third constituent, *YAO4*.

Statistics are gathered by the model as each learned construction is used to interpret new utterances: (1) unigram frequency, e.g. how many times *XIIXII-CHII-YAO4* is used, (2) bigram probability, e.g. the probability that *XIIXII* is followed by *CHII*, (3) the probability that each constituent is omitted, e.g. the probability that *YAO4* is omitted in *XIIXII-CHII-YAO4*, and (4) the probability that each constituent is filled by a particular construction, e.g. the probability that *XIIXII* fills the first constituent of *XIIXII-CHII-YAO4*. The combinatorial power of the grammatical structures and statistics comes through when the model generalizes across syntactically and semantically aligned constructions. The generalization process creates increasingly bigger grammatical categories of words as well as more general constructions that use them. For these constructions, (4) is exactly the probabilistic subcategorization knowledge that adult speakers possess (Bybee, 2006).

After exposure to about 750 utterances, the model learned a number of proto argument structure constructions, including an NP-VP-like construction denoting a solid object in an intransitive state. The NP-like constituent has a probability of 0.54 of being omitted, consistent with subject-drop in Mandarin Chinese, and the VP-like constituent is filled most frequently by the word for *broken*, followed by *good*, *small*, and *amusing*. This model was further evaluated on a 300-utterance validation corpus in terms of its ability to create semantic interpretations of unseen utterances and showed significant improvements as learning progressed.

References

- Bergen, B. K. & Chang, N. (2005). Embodied Construction Grammar in simulation-based language understanding. In J.-O. Östman & M. Fried (eds.), *Construction Grammars: Cognitive Groundings and Theoretical Extensions*. Philadelphia, PA: John Benjamins.
- Bybee, J. (2006). From usage to grammar: The mind's response to repetition. *Language* 82, 711.
- Chang, N. (2008). Constructing grammar: A computational model of the emergence of early constructions. Ph.D. dissertation, University of California Berkeley.
- Fillmore, C. & Kay, P. (1999). *Construction grammar*. Stanford, CA: CSLI.
- Goldberg, A. E. (1995). *Constructions: A construction grammar approach to argument structure*. Chicago: University of Chicago Press.
- Gomez, R. (2002). Variability and detection of invariant structure. *Psychological Science* 13, 431-6.
- MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk*. Third Edition. Mahwah, NJ: Lawrence Erlbaum Associates.
- Tardif, T. (1993). Adult-to-child speech and language acquisition in Mandarin Chinese. Unpublished doctoral dissertation, Yale University.
- Thompson, S. P. & Newport, E. L. (2007). Statistical Learning of Syntax: The Role of Transitional Probability. *Language Learning and Development* 3, 1-42.
- Wonnacott, E., Newport, E. L. & Tanenhaus, M. K. (2008). Acquiring and processing verb argument structure: Distributional learning in a miniature language. *Cognitive Psychology* 56, 165-209.