

An automated distribution-driven prefix learner

Marco Baroni, University of California, Los Angeles

During language acquisition, children must discover which strings constitute the morphemes of their language and which words of the language can be decomposed into morphemes. For example, children acquiring English must learn that *re-* is a prefix, and that *refry* can be decomposed into *re+fry*, but *retail* is not composed of *re+tail*.

Learners undoubtedly use various morpheme discovery strategies, exploiting semantic, syntactic and phonological information. It is also likely that learners use the frequency and distribution of words and their substrings as sources of evidence. Distributional information can be straightforwardly extracted from the data and it can be used prior to any linguistic analysis: learners can later use more sophisticated linguistic information to refine the coarse guesses on morphological structure made on the basis of distributional information.

Brent and Cartwright (1996) used a computational learning model to show that a distributional approach can be successful in the related task of sentence segmentation into words. To assess the effectiveness of a distributional approach to morpheme discovery, I designed an automated distribution-driven prefix learner. The learner takes a list of words from a language as its input, tries to discover the prefixes of the language and decides which input words are prefixed.

The learner uses a “greedy” algorithm (Cormen, Leiserson and Rivest 1990: 329-355) to generate a set of lexica compatible with the input data (where a lexicon is a list of prefixes and stems), and it evaluates the competing lexica using the Minimum Description Length (MDL) principle (Rissanen 1978). The MDL formula used here favors lexica in which frequent initial substrings are treated as prefixes, and strings frequently occurring as independent words or as final substrings are treated as stems (stem+suffix combinations are treated as stems). A string is more likely to be a morpheme (prefix or stem) if it is long and/or if it combines frequently with other potential morphemes. Furthermore, substrings are more likely to be treated as morphemes when they occur in low frequency words than when they occur in high frequency words (intuitively, frequent etymologically complex words are more likely to develop lexicalized meanings and become opaque: a preliminary lexico-statistic analysis of a corpus of English prefixed words confirmed this intuition).

The learner was tested on a list of 25,000 orthographically transcribed English word tokens randomly selected from the PHLEX database (Seitz, Bernstein, Auer and MacEachern 1998). The learner selected a lexicon containing 12 prefixes, 9 of which are actual English prefixes (*re-*, *con-* and its allomorph *com-*, *pro-*, *de-*, *un-*, *dis-*, *in-*, *ex-*). In a pilot analysis, the morphological parses (prefixed vs. non-prefixed) assigned by the learner to the corpus words beginning with *re* were compared to morphological complexity ratings assigned to the same words by three native English speakers. The two sets are significantly correlated. Interestingly, the correlation between the learner and the judges is higher when only semantically opaque forms are considered. This suggests that humans are sensitive to distributional cues similar to the ones exploited by the automated learner, especially when they have to parse forms for which they cannot rely on semantic cues. A larger survey, involving more subjects and forms beginning with other strings, is in progress.

The preliminary results show that a distribution-driven morpheme discovery procedure is effective and that it assigns morphological parses matching humans’

morphological intuitions. If these results are confirmed by the analysis in progress, our research will provide further support for the claim that often-overlooked distributional cues play an important role in language acquisition (see Redington and Chater 1998 for a review of several studies providing support for this claim in various linguistic domains).

References

- Brent, M. and T. Cartwright 1996. Distributional regularity and phonotactic constraints are useful for segmentation, Cognition 61: 93-125.
- Cormen, T., C. Leiserson and R. Rivest 1990. Introduction to algorithms, Cambridge, MA: MIT Press.
- Redington, M. and N. Chater 1998. Connectionist and statistical approaches to language acquisition: A distributional perspective, Language and Cognitive Processes 13: 129-191.
- Rissanen, J. 1978. Modeling by shortest data description, Automatica 14: 456-471.
- Seitz, P., L. Bernstein, E. Auer and M. MacEachern 1998. The PHLEX Database, Los Angeles, CA: House Ear Institute.