Available online at www.sciencedirect.com
SCIENCE DIRECT®
Cognitive Systems Research 6 (2005) 17–25
Cognitive Systems RESEARCH
www.elsevier.com/locate/cogsys

On the resolution of ambiguities in the extraction of syntactic categories through chunking

Action editor: Christian Schunn

Daniel Freudenthal [a,*], Julian M. Pine [a], Fernand Gobet [b]

[a] School of Psychology, University of Liverpool, Liverpool L69 7ZA, UK
[b] Department of Human Sciences, Brunel University, Uxbridge, Middlesex UB8 3PH, UK

Received 16 September 2004; accepted 16 September 2004

Abstract

In recent years, several authors have investigated how co-occurrence statistics in natural language can act as a cue that children may use to extract syntactic categories for the language they are learning. While some authors have reported encouraging results, it is difficult to evaluate the quality of the syntactic categories derived. It is argued in this paper that traditional measures of accuracy are inherently flawed. A valid evaluation metric needs to consider the well-formedness of utterances generated through a production end. This paper attempts to evaluate the quality of the categories derived from co-occurrence statistics through the use of MOSAIC, a computational model of syntax acquisition that has already been used to simulate several phenomena in child language. It is shown that derived syntactic categories that may appear to be of high quality quickly give rise to errors that are not typical of child speech. A solution to this problem is suggested in the form of a chunking mechanism that serves to differentiate between alternative grammatical functions of identical word forms. Results are evaluated in terms of the error rates in utterances produced by the system as well as the quantitative fit to the phenomenon of subject omission.

Keywords: Distributional learning; Co-occurrence statistics; Syntactic categories; MOSAIC; Chunking; Language acquisition; Cognitive modelling

1. Introduction

In recent years, several authors have argued that co-occurrence statistics can provide powerful cues that may aid children in extracting syntactic

* Corresponding author.
E-mail addresses: d.freudenthal@liverpool.ac.uk (D. Freudenthal), julian.pine@liverpool.ac.uk (J.M. Pine), fernand.gobet@brunel.ac.uk (F. Gobet).

1389-0417/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.cogsys.2004.09.003

categories for the language they are learning (Edelman, Solan, Horn, & Ruppin, 2004; Mintz, 2003; Redington, Chater, & Finch, 1998). Redington et al. (1998) analysed large corpora of child-directed speech and performed a cluster analysis on vectors describing the lexical context in which words occurred. They found that words that occurred in linguistically similar contexts (tended to be preceded and followed by the same words) had a high likelihood of belonging to the same syntactic class.

Mintz (2003) expanded on the work of Redington et al. Rather than analysing vectors describing lexical context, Mintz's unit of analysis was a frame: two jointly occurring words with one word in between. Mintz restricted his analysis to the 45 most frequent frames that occurred in a large corpus.

While both Redington et al. and Mintz showed that their procedure resulted in apparently good syntactic categories, there is an inherent difficulty with the use of co-occurrence statistics to derive syntactic categories. As Pinker (1987) points out, words that occur in similar contexts may not be of the same category. Pinker argues that a distributional learning mechanism faced with utterances 1a, b and c, would produce an ungrammatical utterance like 1d.

1a. John ate fish.
1b. John ate rabbits.
1c. John can fish.
1d. *John can rabbits.

Mintz (2003) claims that 'in children's actual input, these problems do not significantly undermine the informativeness of distributional patterns' (p. 92). He also suggests that 'although problematic environments may exist, there is nonetheless enough "signal" in the distributional patterns compared to the noise created by the problematic environments that categorization from distributional patterns is not intractable' (p. 93).

However, the approach taken by Mintz and Redington et al. may obscure the extent of the problem identified by Pinker. Mintz and Redington et al. evaluated the quality of the extracted categories using criteria of accuracy and completeness. Accuracy was computed by classifying every word-pair within a category as a hit (same syntactic class), or miss (different syntactic class). Where the grammatical class of a word was unclear, the corpus was consulted to disambiguate and label the word. Mintz used two types of labeling. In standard labeling, all nouns and pronouns were classed as nouns, and all verbs (lexical verbs, auxiliaries and the copula) were classed as verbs. In expanded labelling, nouns and pronouns were labeled as distinct categories, as were lexical verbs, auxiliaries and the copula. While Mintz achieved high levels of accuracy with both types of labelling, closer inspection of his categories reveals that they may not be as accurate as his analyses suggest. One of Mintz's verb categories contains verbs in present tense and past tense as well as progressive participles, verbs that can and cannot be used in an imperative frame, and verbs such as *do* and *have* that can be used both as a main verb and as an auxiliary.

This heterogeneity of the derived word classes may not appear problematic since neither Mintz nor Redington et al. concern themselves with production. (Mintz views the process of extracting distributional categories as a precondition for a (relatively unspecified) process of bootstrapping into a parametrized universal grammar). When one considers how the extracted categories might be used in production, however, it quickly becomes apparent that heterogeneous word classes will result in utterances that deviate considerably from child speech. The simplest way in which a child producing speech could use the categories arrived at through a distributional analysis of the input is by considering the members of a category as equivalent. That is, if words *a* and *b* occur in the same category, the child may simply substitute *a* for *b* in a context where it knows *b* has occurred. Taking the words *do, have* and *put* (which were classed together in Mintz's analysis) as an example, such a substitution mechanism will result in (clearly incorrect) utterances such as *Do you got an ice-cream* and *Put you want a drink*.

However, more subtle problems, that are not apparent with the use of an evaluation metric based on a researcher's intuition about a word's syntactic class, emerge as well when syntactic categories derived from co-occurrence statistics are