Simulation of Lexical Acquisition

James Kilbury

This paper describes ongoing research at the University of Düsseldorf in a project funded by the Deutsche Forschungsgemeinschaft. The goal of the work is to simulate the acquisition of lexical items under restricted conditions in an implemented system. We thus are investigating specific aspects of human intelligence within the framework of cognitive computational linguistics.

The investigation is based on the premise that the use and acquisition of language involve the same principles and that the representation of new words (more precisely: lexemes) can be established on the basis of given redundancy conditions and contextual information. The identification of new lexemes is regarded as a complex process in which the various modules (i.e. components) of the system interact with each other.

The central questions we address involve (1) the representation and processing of new lexical knowledge, and (2) the interaction of system components during processing. Our formalism for the encoding of syntactic rules and lexical entries is unification-based: a version of PATR-II (cf Shieber 1986) extended with disjunction and negation. A distinctive feature of the system is its use of the default-based inheritance mechanism of DATR (cf Evans and Gazdar 1989) to capture generalizations for which PATR-systems normally employ macros or templates. Morphological alternations are dealt with by a morphological parser based on the work of Bear (1988). We employ our own novel treatment of morphosyntax based on a notion of nonlocal subcategorization. In the near future the system will be extended to deal with lexical semantics.

Since the PATR formalism is "tool oriented" in the sense of Shieber, the choice of a linguistically motivated syntactic formalism is not prejudiced, because a wide variety of formalisms can be implemented in PATR. Likewise, although our system is being used for the analysis of German, its modular structure guarantees that it can be used with syntactic rules and lexica for other languages.

The research described in Zernik (1989) shows that a wide variety of approaches to lexical acquisition is presently being investigated. Our approach involves no "tutoring" of the system by an interacting human. On the contrary, the system attempts to make the best of all available contextual information for the identification of a new lexical item as soon as it appears during the analysis of input.

The informal notion of a "new lexical item" is to be understood in the two distinct senses of (1) neologisms, possibly arising through productive processes of word formation, and (2) existing lexical items which are not included in a given lexicon. With an adequate component for word formation at its disposal, however, the system treats both senses as the same since information involving the preexistence of a lexical item is unavailable to it.

From a cognitive viewpoint our goal is the partial simulation of an adult reader's ability to read a text containing unfamiliar vocabulary. Rather than treating new lexical items simply as noise, the reader partially learns the items from contextual information. As an idealization, we assume that the reader has a complete structural command of the language in question. Thus, we do not address the question of lexical acquisition in children, where structural acquisition takes place concurrently.

The following points summarize the fundamental principles that we consider to govern the assignment of contextual linguistic information to new lexical items:

(1) All linguistic information is represented with the feature- structure matrices of Unification Grammar (cf Shieber 1986).

(2) Unification (cf Shieber 1986) serves as the single operation for combining information in the identification of new lexemes.

(3) As a consequence of (2), the construction of

new lexical representations is regarded as an incremental combination of partial (i.e. incomplete) information. Previously constructed representations can be extended through unification with information gained from additional contexts. Lexical acquisition in this restricted situation thus involves the instantiation of parameters rather than the construction of hypotheses.

(4) Assumption (3) constrains the form of possible lexical representations.

(5) The syntactic parser uses a mixed top-down and bottom-up analysis strategy in the form of a left-corner algorithm (cf Johnson-Laird 1983) with a top-down filter. Consequently, expectations involving the next syntactic constituent to be analyzed play a central rule in the analysis process.

(6) Unification allows a simulation of the parallel interaction of all linguistic levels (i.e. morphological, syntactic, and semantic) during anlysis.

(7) It is presupposed that new lexical items are morphologically and syntactically regular and that all irregular forms are already included in the existing lexicon.

(8) The lexicon as a whole as well as its individual entries are regarded as being incomplete and extendible. New entries can be constructed, and those already present can be further specified. This principle is restricted by (9) and (10), however.

(9) The distinction between open and closed lexical classes is essential in order to restrict the structurally possible analyses of unknown lexemes. We assume that new lexical items can only be assigned to open classes; existing entries for lexemes in open classes can be extended (e.g. for new subcategorizations). The closed lexical classes, in contrast, consist of lexemes with complete and nonextendible entries. Whether or not a lexeme belongs to an open class depends on its semantic and other features.

(10) All lexical items encountered in an analysis are regarded as "new" if they have no appropriate entry in the lexicon. In general, a new lexical entry can be constructed even if an inappropriate entry for the same word form already exists. But in order to restrict the variety of possible structural analyses it must be assumed that there are reserved word forms (i.e. structure words) which allow no further entries. Thus, the system should not be allowed to discover a new English substantive homonymous with the definite article 'the'.

The system is implemented in Arity/PROLOG and runs under MS-DOS.

REFERENCES

Bear, John (1988) "Morphology with Two-Level Rules and Negative Rule Features," Proceedings of COLING 88, 28-31.

Evans, Roger / Gazdar, Gerald (1989) The DATR Papers: May 1989. Cognitive Science Research Paper 139. Brighton: University of Sussex.

Johnson-Laird, P. N. (1983) Mental Models. Cambridge et al.: Cambridge University Press.

Shieber, Stuart M. (1986) An Introduction to Unification-Based Approaches to Grammar. Stanford, California: CSLI.

Zernik, Uri (1989) Proceedings of the First International Lexical Acquisition Workshop. Detroit, Michigan.

JAMES KILBURY
UNIVERSITÄT DÜSSELDORF
SEMINAR FÜR ALLGEMEINE
SPRACHWISSENSCHAFT
D-4000 DUESSELDORF 1
FED. REP. OF GERMANY
E-MAIL: KILBURY@DD0RUD81