Machine Learning 2: 5-8, 1987

EDITORIAL

Machine Learning and Grammar Induction

Language and the Acquisition of Syntax

Language is a major component of cognition, and as such, its acquisition has been a central concern of machine learning researchers. Some of the earliest AI learning work focused on this topic, and interest has continued to the present. However, progress in this area has been much slower than in most other learning tasks, undoubtedly due to the inherent complexity of natural language.

Despite its complexity, the task of natural language processing can be divided into a number of well-defined subtasks, and one of these centers on *syntax.* This aspect of language has been studied in detail by linguists, and developmental studies have provided a variety of empirical generalizations about the stages that children traverse in their acquisition of grammar. Because our knowledge of syntax is more complete than that for other components of language, the vast majority of language-related research in machine learning has focused on the task of grammar induction.

Two Views of Grammar Induction

Within this effort, two different paradigms have emerged for describing the grammar induction task. The first approach was formulated by Solomonoff (1959) and others in the early days of AI. It assumed only a set of legal sentences as input, from which the learner induced a grammar that would parse those sentences. This approach was quite popular during the 1960's, during which Gold (1967) and others formulated a number of formal results about the task. This paradigm has sometimes been called *grammatical inference.*

The second approach did not appear until the late 1960's, when some researchers noted that in natural languages, grammars were used for more than simply parsing sentences - they also *mapped* sentences onto meaning structures. This led to an alternative view of the grammar induction task: given pairs of sentences and their associated meanings, the learner induced grammatical rules for mapping legal sentences onto meaning structures. Siklóssy (1968) and Klein and Kuppin (1970) carried out early work along these lines, which we may call the *grammatical mapping* paradigm. This new framework emerged as the standard view of grammar acquisition in machine learning during the 1970's, though a few continued to work in the grammatical inference tradition.

Researchers who favored the grammatical mapping paradigm rejected the earlier approach for a number of reasons. First, it was clear that language involved more than syntax, despite the insights into grammar that had been provided by linguistics. Second, there was considerable evidence that the task confronting children corresponded more closely to the mapping model. At least in the early stages of language acquisition, parents provide sample sentences that describe objects and events in the immediate environment. Thus, it seemed reasonable to infer that first language learners had available not only legal sentences, but also the meanings of those sentences.

Despite these arguments, the grammatical inference approach remains active and productive, as evidenced by the two papers in this issue of *Machine Learning*. In one paper, VanLehn and Ball show that a variation on Mitchell's (1982) version space method can be applied to grammar induction. In the other, Berwick and Pilato show that a grammar learning algorithm proposed by Angluin (1982) can be applied to significant portions of English grammar. Given the resurgence of activity in this paradigm, the arguments made against the approach deserve some response.

Reasons for Studying Grammatical Inference

One motivation for studying grammatical inference is completely independent of psychological and linguistic issues – it holds significant interest as an abstract learning task. In their paper, VanLehn and Ball show that grammar induction has features that make it more challenging than many machine learning problems. New challenges lead to new representations and new algorithms, and these constitute progress for our developing field. VanLehn and Ball explicitly state that they are not modeling human learning, and they note that the task of grammatical inference arises in other contexts than the acquisition of natural language. Thus, solutions to this problem may have real-world applications, and they supply one example that involves inducing the command language for an operating system.

Returning to the issue of language acquisition, there is no question that human language use involves more than making judgements about the grammatical correctness of sentences, and in many cases people can determine a sentence's meaning even when it is ungrammatical. Nevertheless, people *can* make grammaticality decisions, and the acquisition of this ability deserves some explanation. In fact, humans can make such decisions even about 'nonsense' sentences like *colorless green ideas sleep furiously.*

This apparent decoupling suggests two separate but interrelated processes – one for syntax and the other for semantics – each with its own learning mechanisms. This in turn suggests that one might study these two aspects of language acquisition separately, at least in the early stages of theory construction.

Another argument against the grammatical inference paradigm involved children's use of semantic feedback from the surrounding environment. This undoubtedly holds for the earliest periods of language acquisition, from 18 to 30 months, when the child is determining the relative order of content words and similar matters. However, it becomes progressively less true of later stages, such as those modeled by Berwick and Pilato in this issue. The aspects of grammar learned during these later periods are much less closely linked to semantics than are those aspects learned during the earliest stages. Thus, the grammatical inference view may be a poor model for early grammar induction but, at the same time, an ideal model for late syntactic acquisition.

General mechanisms and domain-specific knowledge

Another long-standing dichotomy – again within the linguistic and psychological literature on language acquisition – concerns the amount of knowledge initially available to the learner. *Nativists* argue that humans come to the grammar learning task with considerable knowledge of the domain, including the basic form of sentences and the basic word classes. According to this view, a weak learning method suffices to acquire syntax because the space of possibilities is so constrained. In contrast, *empiricists* claim that humans have little innate knowledge of language and that they acquire grammar using the same powerful inductive techniques they employ for other domains. In this scheme, the learning task is constrained not by prior knowledge, but rather by experience itself.

Few modern-day researchers would take either the extreme nativist or the extreme empiricist position, though most students of language acquisition lean towards one end of the spectrum or the other. Yet machine learning researchers have seldom taken an explicit stance on this issue, and some computational models of grammar acquisition have merged these two world views with profitable results. For instance, Anderson's (1977) LAS system used some quite general inductive techniques to acquire ATNs from sentence-meaning pairs, but it also made important assumptions about the space of grammars to constrain its learning process.

The two papers in this issue provide further evidence of the advantages inherent in combining the nativist and empiricist frameworks. Both approaches rely on general inductive methods but combine them with constraints on the grammar space that reduce search to manageable proportions. VanLehn and Ball identify a special case of context-free grammars – the class of *simple* and *reducible* grammars – and show that it meets the requirements for a modified version space scheme. They also propose additional biases that further limit search. Berwick and Pilato focus on another constrained class of grammars, in this case a subset of regular languages called k-reversible grammars.

The basic approach in both papers is the same – to identify limits on grammar that reduce search while retaining significant expressive power, whether for modeling human language use or for other tasks. The authors then apply a general learning scheme to the reduced space of grammars, using 'empiricist' learning methods to explore a 'nativist' set of hypotheses. This approach is not a compromise but a *synthesis*, with implications reaching beyond the domain of grammar induction into many other branches of machine learning. We encourage other researchers in the field to consider analogous solutions for their own tasks.

Pat Langley
University of California, Irvine
Langley@CIP.UCI.EDU

References

Anderson, J. R. (1977). Induction of augmented transition networks. *Cognitive Science*, *1*, 125–157.

Angluin, D. (1982). Inference of reversible languages. *Journal of the Association for Computing Machinery*, *29*, 741–765.

Gold, E. M. (1967). Language identification in the limit. *Information and Control*, *10*, 447–474.

Klein, S., & Kuppin, M. A. (1970). *An interactive, heuristic program for learning transformational grammars* (Technical Report No. 97). Madison: University of Wisconsin, Department of Computer Science.

Mitchell, T. M. (1982). Generalization as search. *Artificial Intelligence*, *18*, 203–226.

Siklóssy, L. (1972). Natural language learning by computer. In H. A. Simon & L. Siklóssy (Eds.), *Representation and meaning: Experiments with information processing systems.* Englewood Cliffs, NJ: Prentice–Hall.

Solomonoff, R. (1959). A new method for discovering the grammars of phrase structure languages. *Proceedings of the International Conference on Information Processing.*