Abduction of Semantic Patterns

U. Grenander
Division of Applied Mathematics, Brown University
Providence, RI 02912, USA

1. Introduction

In the study of regular structures the mappings between them play a central role, just as is the case in algebra in general. One therefore examines homomorphisms, isomorphisms, and deformations between regular structures in order to better understand the structures. Here we shall also investigate mappings, but from another point of view, namely that of learning the mapping from a finite experience.

To fix ideas we shall choose two special types of structures. One will be chosen as a highly specialized "world model $\mathcal{G}$", it is intended to represent idealized fragments of the world around us. The second one will be chosen as a simple formal language $L$ generated by a finite state grammar. It is intended to represent a small fragment of an idealized linguistic environment.

Both of these structures are too restricted to be able to represent even a part of the real world or of real language, and this is not our purpose. This paper is limited to the mathematical study of a formalized notion of semantics; it is not empirical but is in the nature of a thought experiment.

Semantics will be understood as a mapping from $L$ to $\mathcal{G}$, and we shall try to identify the mathematical problems that it gives rise to. This paper reports work in progress of the author and two of his graduate students, P. Flanagan and W.-C. Chen, and tries to give some preliminary and partial answers to some of the questions raised.

2. Abduction of Syntax

Scientific reasoning has traditionally be enclassified as inductive or deductive depending upon whether it involves arguing from special cases - say observations and measurements - or uses syllogistic procedures. In the natural sciences a mixture of both is used. Mathematics employs only deductive methods.

It has been argued by C.S. Peirce, however, that one should add a third mental activity to these two. Naming it abduction he meant the conscious or subconscious act by which the scientist forms hypotheses, conjectures, guesses.

How can syntax be learnt? Can one suggest a priori reasonable organization of the abduction process? Several years ago the

1) Supported by Office of Naval Research Contract N00014-77-C-0248 with Brown University.
following thought experiment was carried out by the author and his coworkers; for some details see [1].

A learner is "listening" to syntactically correct sentences and is "speaking" sentences and told when they are not correct. On the basis of this information he tries to form hypotheses about the syntax.

To fix ideas consider a language \( L(\mathcal{G}) \) generated by a finite state grammar \( \mathcal{G} \) with a terminal vocabulary \( V_T = \{a, \beta, \gamma, \ldots\} \), a non-terminal vocabulary \( V_N = \{1, 2, \ldots F\} \), and a set \( \mathcal{P} \) of productions. We shall use the convention that 1 is the initial state, \( F \) is the final state. Productions are written in the form \( p: i \rightarrow x j \), where \( i, j \in V_N, x \in V_T \). The grammar can be visualized by the wiring diagram of a finite automaton which can be assumed without loss of generality to be deterministic.

For a given \( i \in V_N \) we associate probabilities to each production rewriting \( i \). Of course these probabilities shall be non-negative and add to one. This will determine a probability distribution over the language \( L \).

As an example see the wiring diagram in Figure 1 where

\[
\begin{align*}
V_T &= \{a, \beta, \gamma, \delta\} \\
V_N &= \{1, 2, 3, \ldots F\}; \quad F=12 \\
\mathcal{P} &= \{1 \rightarrow 2, 2 \rightarrow 2, 2 \beta 3, 2 \gamma F, \ldots \}
\end{align*}
\]

It generates, for example, the sentence \( a \beta \delta \gamma \beta \delta \) by successive application of the productions \( 1 \rightarrow 2, 2 \rightarrow 3, 3 \beta 4, 4 \gamma 5, 5 \beta 10, 10 \beta F \); compare with the wiring diagram.

It is not difficult to suggest abduction processes that are guaranteed to converge to the true syntax and many have been proposed. For example, the well known maximum likelihood method can be applied to the present case. The application takes some care since the parameter space is here infinite dimensional, but it was shown by the author in 1967 how this technical problem can be solved.

This result is of little value, since the resulting abduction process would require enormous computing and memory capacity; also it would be quite slow. It is too mechanical and does not take into account the grammatical structure of the problem.

It is more appealing to base the abduction on an important grammatical concept: congruence. Consider two strings from \( V_T \), say \( x_1 x_2 \ldots x_n = u \) and \( y_1 y_2 \ldots y_m = v \). Complete both strings with some arbitrary string \( w = z_1 z_2 \ldots z_p \). If for any \( w \) the completed strings \( uw = x_1 \ldots x_n z_1 \ldots z_p \) and \( vw = y_1 \ldots y_m z_1 \ldots z_p \) are either both syntactically correct or both wrong, then \( u \) and \( v \) are said to be congruent.