pp. 511-525.

Machine Learning of Natural Language: Problems and Prospects

Patrick Suppes, Michael Böttner, Lin Liang and Raymond Ravaglia
Stanford University, Ventura Hall, Stanford, CA 94305-4115

May 18, 1995

Abstract

We are developing a theory of machine learning of natural language applicable to various sublanguages used in special domains. For several years our work concentrated on robotic constructions in elementary assembly actions. Recently we have turned to physics word problems. The general axioms of the theory stated in section 2 are applicable without change to both of these rather different applications.

We also give in Section 2 some detailed grammatical results for the robotic environment based on a test set of 460 commands. In Section 3 we outline and illustrate our current research on physics word problems.

1 Introduction

We have taken what we believe is a new tack in the approach to machine learning by using in a very explicit way principles of association and generalization derived from classical psychological principles. On the other hand, we hasten to add that the actual principles we used are not mere copies of classical psychological principles, but are much more specific and technical. In fact, it is a firm conviction of ours that one of the great failures of the classical psychological tradition in theories of association and conditioning as applied to language is the absence of serious technical development of the principles. What is missing are precise and explicit formal statements that could be used as a basis for either extensive theoretical investigations or, for what we have emphasized in this report, experiments in machine learning. To guard against the principles being too strictly tailored to learning English, we have simultaneously applied the principles to English, German, and Chinese. We have been fortunate in that these three major languages has been represented by a native speaker on the project staff.

There are two general points about our work that we want to emphasize. First, our viewpoint is, in many respects semantic, rather than syntactic. The formation of grammatical forms, for example, in each of the languages, is driven by a uniform set of semantic categories, rather than grammatical categories — uniform here meaning uniform across the three languages. The second general point, one to be emphasized quite strongly, is that the current experiments have been solely concerned with language comprehension. We have made no effort whatsoever to construct a learning environment that can produce utterances in any natural language. It is apparent on a moment's reflection that the construction of a purely comprehension program by machine learning is a very much easier task than including production as well.

Principles of Learning: Association and Generalization. In a general sense the principles of learning we use are classical and well-known in psychology. In fact, it can be claimed that the principle of association goes back at least to Aristotle, and certainly was used extensively by eighteenth-century philosophers like Hume long before psychology had become an experimental science. The fundamental role of association as a basis for conditioning is thoroughly recognized in modern neuroscience and is essential to the experimental study of the neuronal activity of a variety of animals. For similar reasons its role is just as central to the learning theory of neural networks, now rapidly developing in many different directions.

We have not, however, made explicit use of neural networks, but have worked out our theory of language learning at a higher level of abstraction. In our judgment the difficulties we face need to be solved before a still more detailed theory is developed. Choice of some level of abstraction is inevitable in all work of the present kind — a fact not always appreciated by some interested bystanders.

Whatever the level of abstraction there is one general issue about association that must be faced. Is each given kind of association postulated to be incremental or all-or-none? It is intuitively natural in many ways to suppose that strength of association between a stimulus and a response, or between two stimuli increases with practice. Without doubt most motor skills show improvement, i.e., continued learning, with practice. On the other hand, there is a large body of experimental studies of many different kinds to support the view that in simple experimental situations association is all-or-none. For an example of detailed statistical analysis of this issue applicable to the learning of simple concepts, see Suppes & Ginsberg (1963).

In the present work we have almost exclusively assumed that associations are formed on an all-or-none basis, although we have had extensive discussions of where it would be valuable to use a strength-of-association measure to produce a natural probability measure for retrieval from memory of different associations, or, in another direction, to produce a measure for selecting weak associations that are to be changed.

The classical psychological principles of learning used here have been thought by linguists to be wholly inadequate as the basis for a theory of language learning. In fact, many linguists probably still believe that the *coup de grâce* to such ideas was given by Chomsky (1959) in his famous review of Skinner's (1959) *Verbal Behavior*. Nothing could be further from the truth. It is rather like saying that Newton's refutation of Cartesian vortex theory delivered a fatal blow to field theories in physics. Skinner's naive formulation of the problems of language learning was rightly attacked by Chomsky, but no serious alternative learning theory has been offered by linguists even today.

Recent work. Because this last remark is controversial, it needs to be amplified, which we can do by examining several fairly recent proposals. In principle a work that is very close to what we have attempted here is Wexler & Culicover (1980). Wexler & Culicover have an ambitious program and part of it is executed in great detail and provides considerable insight into language learning. Their work is regarded as one of the most important contributions thus far, to what has come to be called the theory of learnability, with special reference to language learning. The authors bring together two sets of theoretical assumptions. One concerns rather simple ideas about learning mechanisms, for example, hypothesis testing and learning only from errors. The other concerns linguistic constraints on the nature of the grammar that the child must learn. We emphasize by the way, that their work is wholly concerned with grammar

and not with learning of meaning. The reason for this is that they assume that the child has the deep structure of the language already available and that what is being learned is the transformational grammar that produces the surface structure. What they show is that on the assumption that the deep structure is already available to the child — an assumption not too far from our own assumptions about the internal representation of the semantics of a command —, can lead to learning of a transformational grammar producing the surface structure.

On the other hand their work is very different from ours because it is purely theoretical and does not work out the details of a particular example. What is especially important to contrast in the nature of the work is that they prove an asymptotic theorem that in the limit the transformational grammar will be learned, but they do not give any estimates as to the rate of learning. In contrast, our own work has been concerned with a constrained situation in which actual learning does take place and which the number of trials for learning is relatively small.

The second and more important difference is that their learning mechanisms are quite simple and do not have anything like our complicated process of association and generalization. There are several reasons for this. One is that the deep structure they assume is already in the form of a context-free grammar to which transformations are applied. In our case, there is no close connection, as we remark later, between the syntactic structure of the internal representation, and the syntactic structure of the various natural languages we study. We have therefore been concerned to state principles of learning that have a more general character and that can be applied to the details of the three natural languages we consider. In fact, a proper way to characterize the difference is that we introduce specific ideas of learning, whereas the learning mechanisms used by Wexler and Culicover are classical stochastic models of learning. On the other hand, they introduce specific principles concerned with transformations and this is what is original about their work. For example, the *freezing principle*, which they introduce, asserts that if a transformation changes the structure of a node so that part of the base structure is no longer a base structure, then no transformations may be applied to subparts of the structure of the node. Second, they introduce a *binary principle* which restricts transformations to applying to constituents that cut across more than two embedded sentences in the base structure. Third, they introduce a *raising principle* which asserts that if a node is raised, a transformation cannot be applied to a node beneath this node. They also have some other principles of a more technical nature. One can see how different these principles are from the principles of association and generalization we use.

Pinker (1984) moves closer to the data by building his ideas on the theory of lexical functional grammars (Kaplan & Bresnan, 1982), which are generalizations of the phrase-structure grammars characterized above, and more recently (Pinker, 1989) by using Chomsky's government-binding theory. Pinker's work is full of insightful remarks about the problems of theorizing in this area. He considers a variety of psycholinguistic experimental and naturalistic data, but he does not formulate a detailed formal theory. Consequently, he cannot establish whether his theoretical ideas are sensible from the standpoint of computations about rates of learning, formal structure of semantics, formal grammatical generalizations, etc. Wexler also has moved closer to linguistic data in his recent work, e.g. Avrutin & Wexler (1993), Hyams & Wexler (1993).

In the past decade or so there has been a relatively small number of articles or books on machine learning of natural language. Langley and Carbonell (1987) provide an excellent

overview as of the date of their publication. This semantic commitment is also shared by the recent work of Feldman, Lakoff, Stolcke & Hollbach Weber (1990), Feldman, Lakoff, Bailey, Narayanan, Regier & Stolcke (1995) and Siskind (1992,1994). Feldman et al. (1990) describe in direct and simple terms their original idea. First, the learning system is presented pairs of pictures and true natural language statements about the pictures. Second, the system is to learn the language well enough to determine whether or not a new sentence is true of the accompanying picture. Feldman et al.'s (1995) approach to language learning separates the learning of the grammar from the learning of the lexical concepts. The grammar is learned by use of Bayesian inference over a set of possible grammars and model merging. Siskind's original work, Siskind (1992), his dissertation, was in the context of naive physics, but focused also on the algorithms children may use in learning language. This work is continued in Siskind (1994), but with any assumption of prior language knowledge eliminated. The concentration is on lexical acquisition via possible internal representations of meaning.

We are not critical of these continual changes. They are required in dealing with a subject so complex as language. It is important to state how our own strategy differs on our initial work on robotic languages (Suppes, Liang & Böttner, 1992; Suppes, Böttner & Liang, 1995).

We have introduced the idea of grammatical forms as a simplifying framework for the initial acquisition of language. The set of grammatical forms generated by our learning process can properly serve as the data for writing a more general and more thorough generative grammar. We are skeptical about the problems of writing fully adequate grammars at this time for different languages. Even English, the most studied language, is still lacking in anything like a fully satisfactory grammar. It is the reason that we introduce various functional forms, for example, such phrases as *which is* and corresponding phrases in the other languages without any real grammatical analysis but as retained surface parts of grammatical forms.

The defect of our formulation of principles of language learning is the opposite of Skinner's. We have formulated precise technical principles directly applicable to our special environments. We have in no sense aimed at the most general formulation, and we have made no systematic effort to relate our principles of learning to specific principles applicable to human language learning. On the other hand, in formulating our principles we have often been guided by empirical results on child language learning, some of which we cite.

2 Theory

The theory that underlies our learning program is given in terms of a system of axioms. We begin with a general formulation, which is then made more special and technical for the robotic framework we have studied in the past by giving some detailed examples to illustrate the various axioms.

2.1 Background Assumptions

We state informally as background assumptions two essential aspects of any language learning device. First, how is the internal representation of an utterance heard, for example, for the first time, generated by the learner. Second, at the other end of the comprehension process, so to speak, is that of generating a semantic interpretation of a new utterance, but one that falls within the grammar and semantics already constructed by the learner.

Both of these processes ultimately require thorough formal analysis in any complete theory, but, as will become clear, this analysis is not necessary for the framework of this article. We give only a schematic formulation here.

1. *Association by contiguity.* When a learner is presented a robotic command or word problem that it cannot interpret then it associates the utterance to patterns in its contiguous environment whose internal representation may, but not necessarily, be induced by its own free or coerced actions.

2. *Comprehension-and-response axiom.* If a learner is presented a robotic command or word problem, then using the associations and grammatical rules stored in long-term memory, the learner attempts to construct a semantic interpretation of the sentence and respond accordingly.

2.2 General Learning Axioms

We now turn to our learning axioms, which naturally fall into two groups, those for computations using working memory and those for changes in the state of long-term memory. We use, as is obvious, a distinction about kinds of memory that is standard in psychological studies of human memory, but the details of our machine-learning process are not necessarily faithful to human learning of language, and we make no claim that they are. On the other hand, our basic processes of association, generalization, specification and rule-generation almost certainly have analogues in human learning, some better understood than others at the present time. In the general axioms formulated in this section we assume rather little about the specific language of the internal representation. The exposition follows the latest version of the axioms. (Suppes, Böttner & Liang, to appear).

Notation. Concerning notation used in the axioms, we generally use Latin letters for sentences or their parts, whatever the natural language, and we use Greek letters to refer to internal representations of sentences or their parts. Turning now to specific notation, the letters $a, b,$ ... refer to words in a sentence, and the Greek letters α, β, ... refer to internal symbols. The symbol s refers to an entire sentence, and correspondingly σ to an entire internal representation. Grammatical forms—either sentential or term forms—are denoted by g or also $g(X)$ to show a category argument of a form; correspondingly the internal representations of a grammatical form are denoted by γ or $\gamma(X)$. We violate our Greek-Latin letter convention in the case of semantic categories or category variables X, X', Y, etc. We use the same category symbols in both grammatical forms and their internal representations.

To show how our axioms are applied, we intersperse their statement with simple examples from our robotic experiments. We begin with *Get the screw!* which has the internal Lisp representation $(fa_1$ \$*get* (*io* (*fo* \$*screw* $*$))), where fa_1 is the semantic operation *form-action*, \$*get* is the internal representation of *get*, *io* is the semantic operation of identifying an object, *fo* is the semantic operation of finding a class of objects. \$*screw* is the internal representation of *screw* and $*$ denotes the objects in the robot's visual field.

Axioms of Learning

1. Computations using Working Memory

1.1 PROBABILISTIC ASSOCIATION. *On any trial, let s be associated to σ, let a be in the set of words of s not associated to any internal symbol of σ, and let α be in the set of internal symbols not currently associated with any word of s. Then pairs (a, α) are sampled, possibly using the current denotational value, and associated, i.e. $a \sim \alpha$.*

The probabilistic sampling in the case *Get the screw* could lead to the incorrect associations *get* $\sim$ *\$screw*, *the* $\sim$ *\$get* and no association for *screw*, for there are only two symbols to be associated to in the internal representation.

1.2 FORM GENERALIZATION. *If $g(g_i') \sim \gamma(\gamma_i')$, $g_i' \sim \gamma_i'$, and γ' is derivable from X, then $g(X_i) \sim \gamma(X_i)$, where i is the index of occurrence.*

From the associations given after Axiom 1.1 we would derive the incorrect generalization

$$OBJ\ A_1\ screw \sim (fa_1\ A_1\ (io\ (fo\ OBJ\ *))). \tag{1}$$

The correct one is

$$A_1\ the\ OBJ \sim (fa_1\ (io\ (fo\ OBJ\ *))). \tag{2}$$

1.3 GRAMMAR-RULE GENERATION. *If $g \sim \gamma$ and γ is derivable from X, then $X \rightarrow g$.*

Corresponding to 1.3 we now get the incorrect rule

$$A \rightarrow OBJ\ A_1\ screw. \tag{3}$$

The correct one is

$$A \rightarrow A_1\ the\ OBJ. \tag{4}$$

1.4 FORM ASSOCIATION. *If $g(g') \sim \gamma(\gamma')$ and g' and γ' have the corresponding indexed categories, then $g' \sim \gamma'$.*

We get from (1) the incorrect form association

$$OBJ \sim (io\ (fo\ OBJ\ *)). \tag{5}$$

The correct one – to be learned from more trials – is derived from (2)

$$the\ OBJ \sim (io\ (fo\ OBJ\ *)). \tag{6}$$

1.5 FORM SPECIFICATION. *If $g(X_i) \sim \gamma(X_i)$, $g' \sim \gamma'$, and γ is derivable from X, then $g(g_i') \sim \gamma(\gamma_i')$.*

As the inverse of 1.2 using the incorrect generalization given after 1.2, we use 1.5 to infer

$$Get\ the\ screw \sim (fa_1\ \$get\ (io\ (fo\ \$screw\ *))).$$

1.6 CONTENT DELETION. *The content of working memory is deleted at the end of each trial.*

2. Changes in State of Long-term Memory

2.1 DENOTATIONAL VALUE COMPUTATION. *If at the end of trial n a word a in the presented verbal stimulus is associated with some internal symbol α of the internal representation σ of s, then $d_{n+1}(a) = (1-\theta)d_n(a) + \theta$, and if a is not associated with some denoting internal symbol α of the internal representation $d_{n+1}(a) = (1-\theta)d_n(a)$. Moreover, if a word a does not occur on trial n, then $d_{n+1}(a) = d_n(a)$, unless the association of a to an internal symbol α is broken on trial n, in which case $d_{n+1}(a) = (1-\theta)d_n(a)$.*

Because our denotations are conceptually less familiar, we give a more detailed example.

To show how the computation of denotational value (Axiom 2.1) works, let us consider further the associations given are *get* $\sim$ \$*screw*, *the* $\sim$ \$*get*. Let us further assume that at the end of this trial

$$\begin{aligned} d(get) &= 0.900 \\ d(screw) &= 0.950 \\ d(the) &= 0.700. \end{aligned}$$

On the next trial the verbal command is

Get the nut.

As a result, we end this trial with

$$get \sim \$get,\ nut \sim \$nut$$

and with the association of *the* deleted (Axiom 2.6 (a)). Using $\theta = 0.03$, as we usually do, we now have $d(get) = 0.903$, $d(the) = 0.679$. After, let us say, three more occurrences of *the* without any association being formed the denotational value would be further reduced to 0.620.

The dynamical computation of denotation value continues after initial learning even when no mistakes are being made. As a consequence high-frequency words like *a* and *the* in English and *ba* in Chinese have their denotational values approach zero rather quickly. (From a formal point, it is useful to define a word as *nondenoting* if its asymptotic denotational value is zero, or, more realistically, below a certain threshold.)

2.2 FORM FACTORIZATION. *If $g \sim \gamma$ and g' is a substring of g that is already in long-term memory and g' and γ' are derivable from X, then g and γ are reduced to $g(X)$ and $\gamma(X)$. Also $g(X) \sim \gamma(X)$ is stored in long-term memory, as is the corresponding grammatical rule generated by Axiom 1.4.*

Let

$$\begin{array}{rrcl} g \sim \gamma: & A_1\ the\ OBJ & \sim & (fa_1\ A_1\ (io\ (fo\ OBJ\ *))) \\ g' \sim \gamma': & the\ OBJ & \sim & (io\ (fo\ OBJ\ *)) \\ X: & O & \rightarrow & the\ OBJ \\ \hline & A_1\ O & \sim & (fa_1\ A_1\ O) \\ & A & \rightarrow & A_1\ O \end{array}$$

2.3 FORM FILTERING. *Associations and grammatical rules are removed from long-term memory at any time if they can be generated.*

In the previous example, $g \sim \gamma$ can now be removed from long-term memory, and so can $A \rightarrow A_1$ *the OBJ*.

2.4 CONGRUENCE COMPUTATION. *If w is a substring of g, w' is a substring of g' and they are such that*

(i) $g \sim \gamma$ and $g' \sim \gamma$,

(ii) g' differs from g only in the occurrence of w' in place of w,

(iii) w and w' contain no words of high denotational value,

then $w' \approx w$ and the congruence is stored in long-term memory.

Using Axiom 2.4, reduction of the number of grammatical rules for a given natural language is further achieved by using congruence of meaning (Suppes 1973, 1991). Consider the following associations of grammatical forms:

$$die\ Mutter \sim (io\ (fo\ \$nut\ *)) \tag{7}$$

$$der\ Mutter \sim (io\ (fo\ \$nut\ *)). \tag{8}$$

Association (7) and (8) differ only with respect to the article. The article in (7) is in the nominative and accusative case, the article in (8) is in the genitive and dative case. What is important here is that there is no difference in the respective internal representations. We therefore call (7) *congruent* with (8) and collect the differing elements into a congruence class $[DA] = \{die, der\}$ where DA = definite article. This allows us to reduce the two grammatical forms (7) and (8) into one:

$$[DA]\ Mutter \sim (io\ (fo\ \$nut\ *)). \tag{9}$$

Notice that reduction by virtue of congruence is risky in the following way. We may lose information about the language to be learned. For instance, collapsing the gender distinction exhibited by the difference between (7) and (8) will make us incapable of distinguishing between the following sentences:

$$Leg\ die\ Mutter\ auf\ die\ Platte \tag{10}$$

$$Leg\ die\ Mutter\ auf\ den\ Platte. \tag{11}$$

Whereas (10) is grammatical, (11) is not. As long as our focus is on comprehension grammar, a command like (11) will probably not occur, but for purposes of production of the present kind, congruence should not be used.

2.5 FORMATION OF MEMORY TRACE. *The first time a form generalization, grammatical rule or congruence is formed, the word associations on which the generalization, grammatical rule or congruence is based are stored with it in long-term memory.*

Using our original example after 1.3, the incorrect associations would be stored in long-term memory, but with more learning, later deleted (2.6 (i)).

2.6 DELETION OF ASSOCIATIONS.

(i) *When a word in a sentence is given a new association, any prior association of that word is deleted from long-term memory.*

(ii) *If $a \sim \alpha$ at the beginning of a trial, a appears in the utterance s given on that trial but α does not appear in the internal representation σ of s, then the association $a \sim \alpha$ is deleted from long-term memory.*

(iii) *If no internal representation is generated from the occurrence of a sentence s, σ is then given as the correct internal representation, and there are several words in s associated to an internal symbol α of σ such that the number of occurrences of these words is greater than the number of occurrences of α in σ, then these associations are deleted.*

2.7 DELETION OF FORM ASSOCIATION OR GRAMMATICAL RULE. *If $a \sim \alpha$ is deleted, then any form generalization, grammatical rule or congruence for which $a \sim \alpha$ is a memory trace is also deleted from long-term memory.*

2.3 Some Robotic Language Results

In Table 1 we show the grammatical rules derived from machine learning of a training set of approximately 400 robotic commands. The three languages, English, Chinese and German, were learned independently. The results are taken from (Suppes, Böttner & Liang, 1995). Using the congruence axiom (axiom 2.4) to collapse rules that differ only in the occurrence of semantically equivalent nondenoting words, and thereby identifying certain grammar rules as common to the three languages, we obtained the rules shown in Table 1. As can be seen, there are nine rules common to the three languages, there are two rules common just to English and Chinese, and seven rules common just to English and German. There are no rules common only to Chinese and German. Moreover, Table 1 includes all the rules derived for English, but there seven additional rules not shown for Chinese only, and four additional rules not shown for German only.

The bracketed expressions refer to congruence classes on nondenoting words. For example, [DA] is the class of definite articles {*the*, ... *die*}. The main notation in Table 1 has the following intuitive interpretation. O is the category of object, S is the category of a set of objects, P is the category of physical properties, OBJ is the category of special properties denoted by common nouns, G is the category of spatial regions, R is the category of spatial relations, A is the general category of physical actions (to be performed by the robot), A_1, A_2, A_3 and A_4 are special subcategories of actions, and D is the category of directions. As should be clear from their descriptions, the semantical character of each category is primary, rather than its purely syntactic role [1].

[1] **The bracketed expressions for congruence classes have the following intuitive interpretation. [DA] = definite articles, [IA] = indefinite articles, [PO] = possessive prepositions, [&] = conjunctions, [¬] = negations, [V] = disjunctions, [Adv] = adverbs, [COP] = copulas, [RP] = relative pronouns. Several of these classes include $\in$ for the empty string, to extend congruences to more cases.**

Chinese, English, German	
1. $O \rightarrow [DA]\ S$	
2. $O \rightarrow [IA]\ S$	
3. $S \rightarrow P\ S$	
4. $S \rightarrow OBJ$	
5. $G \rightarrow R\ [PO]\ O$	
6. $D \rightarrow R$	
7. $P \rightarrow P\ [\&]\ P'$	
8. $P \rightarrow [\neg]\ P$	
9. $P \rightarrow P\ [\vee]\ P'$	
Chinese, English	English, German
10. $A \rightarrow [ADV]\ A_4\ D\ O$	12. $A \rightarrow [ADV]\ A_1\ O\ [COP]$
11. $A \rightarrow A_4\ O$	13. $A \rightarrow [ADV]\ A_2\ G\ [COP]$
	14. $A \rightarrow G\ A_3\ O$
	15. $A \rightarrow A_3\ O\ [COP]\ G$
	16. $A \rightarrow [ADV]\ A_4\ [COP]\ O\ D$
	17. $S \rightarrow S\ [RP]\ P$
	18. $S \rightarrow S\ [RP]\ G$

Table 1. Comprehension Grammars

3 Physics Word Problems

After devoting several years to machine learning of robotic language, we have recently turned our attention to machine learning of physics word problems, a domain of language use very different from the robotic one. Two features of this new work make it intrinsically more difficult than the robotic effort. First, in the robotic case we constructed ourselves the training set of commands, and consequently they have many special features to facilitate learning, which a corpus collected in the field would not have. In contrast, we are using problems from physics textbooks in English, Chinese and German to make up the training set in each language, possibly augmented by some additional variants constructed by us.

Second, and even more fundamental, the computational semantics we have developed for the physics word problems is meant to be complete enough to construct the equations derived from the word problem and compute a solution. In the robotic case, in contrast, our Lisp internal representations remained at an abstract level, and our actual robotic implementation relied on several levels of programming languages below the Lisp implementation that did not enter into our machine learning procedures, which were developed in a simulation environment. What this meant is that in the robotic case we concentrated only on language learning, and the much harder task of linking language and robotic visual perception and motor control was bypassed. Aspects of machine learning of perceptual concepts needed for this enlarged robotic learning program are the subject of a recent paper of ours (Suppes & Liang (in press)).

3.1 Examples of our computational semantics

In the elementary problems with which we have begun, the basic background assumptions are these: (i) Only kinematics of bodies is analyzed; no dynamics. (ii) Physical bodies are treated as point particles. (iii) Acceleration is always at a constant or uniform rate. (iv) The problems are all one-dimensional. (v) No derivatives of positions or velocity are introduced. Consequently the only relevant data or answers to questions are in terms of the following physical quantities: elapsed time Δt, initial or final times t_0 and t_1; elapsed distance Δx, initial or final positions $x(t_0)$ and $x(t_1)$; change in velocity Δv, initial or final velocity $v(t_0)$ and $v(t_1)$; acceleration a, or, more generally but not needed here, $a(t_i)$.

Here is a very simple problem chosen only to illustrate our method of analysis. We show immediately under the statement of the problem the internal representation used to make semantic computations. Each part of the notation is explained.

A	car	accelerates	from	3.1 m/s
ND	O	a	$([t_i = t_o]$	$[v(t_i) = 3.1\ m/s])$

to	6.9 m/s	in	5.0 s.
$([t_i' = t_1]$	$[v(t_i') = 6.9\ m/s])$	$([\Phi_1 = \Delta X]$	$([\Phi_1 = 5.0\ s][X = t]))$

What	is	its	acceleration?
$[\Phi_2 = ?]$	ND	O	$[\Phi_2 = a]$

From left to right, here is the meaning of the notation:

ND is a word or phrase that is nondenoting in the sense of Axiom 2.1 above. Intuitively such words do not have a relevant physical meaning for the problems considered.

O denotes the physical object under study.

$[t_0 = t_i]$ is the denotation or temporal interpretation of *from*.

$[v(t_i) = 3.1\ m/s]$ is the denotation of $3.1\ m/s$. We use the units to determine the physical quantity v, but the time argument is left variable, to be determined by using the denotation of *from*. Thus, the phrase *from* $3.1\ m/s$ has the denotation (note the outside parentheses to fix order of computation):

$$([t_0 = t_i][v(t_i) = 3.1\ m/s]) \rightarrow (v(t_0) = 3.1\ m/s).$$

The identity uses a simple logical inference about identities.

The analysis of *to* $6.9\ m/s$ is very similar, so that after the same sort of computation the denotation is $(v(t_0) = 6.9\ m/s)$.

The analysis of *in 5.0 s* is somewhat different. The possible values of the variables X are x, v and t (for these problems). From $[\Phi_1 = \Delta X]$ and $[X = t]$ we infer $\Phi_1 = \Delta t$ and then using $[\Phi_1 = 5.0\ s]$ we infer $\Delta t = 5.0\ s$.

The computation $(a = ?)$ for the question posed at the end of the problem goes along the same lines.

The computations given above are trivial, but having a program that learns this special computational semantics and its associated grammar to be able to solve this given class of problems is not trivial.

We turn to remarks about learning in the next section.

3.2 Illustration of learning axioms

In section 1.2 we illustrated our general leanring axioms for the robotic case. Here we do it for physics word problems. of the elementary kind described in the previous section. When the application of an axiom is quite similar to the robotic use, we say this and give no further illustration.

Axiom 1.1. *Probabilistic Association.* Similar to the robotic use.

Axiom 1.2. *From Generalization.* Since, as our example in the previous section showed,

$$from\ 3.1\ m/s \sim [t_0 = t_i][v(t_i) = 3.1\ m/s]) \tag{12}$$

we infer

$$T(t_0)\ R\ U \sim ([t_0 = t_i][\Phi(t_i) = R\ U]) \tag{13}$$

where T is a temporal category, R is the category of real numbers. U is the category of physical units and Φ is the category of physical magnitude functions, which in the present context are x, v, a and t.

Axiom 1.3. *Grammar-Rule Generation.* In the internal equational language for the physics word problems we have the derivation

$$P \rightarrow ([t_0 = t_i][\Phi(t_i) = R\ U]). \tag{14}$$

From this and (13), we infer by Axiom 1.3 the grammatical rule for English word problems

$$P \rightarrow T(t_0)\ R\ U$$

Axiom 1.4. *From Association.* From (13) we infer

$$R\ U \sim [\Phi(t_i) = R\ U], \tag{15}$$

which gives widely used form association between the natural language and internal representation of physical quantities.

Axiom 1.5. *Form Specification.* From (14) and the terminal grammatical rules

$$R \rightarrow 3.1$$
$$U \rightarrow m/s$$

we may, by using Axiom 1.5, derive as a necessary kind of intermediate step in many problem analyses

$$T(t_0)\ 3.1\ m/s \sim ([t_0 = t_i][\Phi(t_i) = 3.1\ m/s]).$$

Axiom 2.1. *Denotational Value Computation.* Similar to the robotic use.

Axiom 2.2. *Form Factorization.* Similar to the robotic use.

Axiom 2.3. *Form Filtering.* Similar to the robotic use.

Axiom 2.4. *Congruence computation.* As shown in our example in the previous case

$$a\ car \sim O$$

and we treat *the car* the same way, i.e.,

$$the\ car \sim O,$$

a typical abstraction in physics word problems. so *a car* is treated as congruent to *the car.* We can also easily extend congruence to denoting words. whose ordinary meanings are quite distinct, to a single abstract case. So in our restriction to problems about point particles *a car* is congruent to *a ball* or *a truck.*

Axioms 2.5 - 2.7. Similar to the robotic use.

4 Problems and Prospects

For the kind of machine learning of language we and others have been involved in. it is easy to list many important unsolved problems. We restrict ourselves to five.

Problems. First, at the present stage of development intense study is needed of each special domain of language to be learned. Special peculiarities of language use must be explicitly attended to and thought about in creating an appropriate internal representation. This is not at all the way we think of children everywhere learning the special language of elementary mathematics, for example.

Second, we know it is hard for an outsider to learn with any thoroughness a specialized domain of science, but it is much harder from machine learning, even though the general learning axioms exhibited, or those used by others, work in several domains. The detailed equational semantics. for example, we are creating for the internal computation of physics equations expressing problem data require us to develop new methods of language-based computation. The extent of this kind, as required, for a variety of special domains. is not yet clear.

Third, as far as we know there is as yet no example from any research group of machine learning of a large corpus of language. Examples of machine translation of large corpora are much closer to realization, but in much of the translation work there is little emphasis on adequate semantic comprehension of the language in question. Given that, independent of any question of machine learning , we do not detailed semantic-based grammars for any large corpus of language, it is clear that scaling up to such a venture in machine learning will present many detailed problems, even if there is no single problem that seems in principle unsolvable.

Fourth, we need to be able better to mix concept and language learning. Children do it well, but our theory of how they do it is still lacking in many essential details.

Fifth, for robots especially, language learning must be deeply connected to visual, auditory and haptic perception. Much practical talk in many important applications is closely linked to ongoing perception. Waht is said can only really be understood in a detailed way by a robot that can link language comprehension to perception of surrounding events.

Prospects. In spite of the problems, as with other scientific and technological developments , prospects for use are not only a distant hope. We list three. First, in the relatively near future real applications to well-defined domains of activity and their relevant sublanguages will appear. Some of the earliest successful technical examples are like to be in medicine, from the emergency room to the office dictation of medical records. Second, it is likely, even if not anything like certain, that within the next decade oral communication with computers, as with people, will be the most important and most used form of communication. If so, single words and phrases will not be good enough. A rich natural sublanguage will be used, and computers will need to

learn it. No doubt, the progress on comprehension may be faster than on production in the early years.

Third, important applications early in the next century at least will be in two domains, reflected in the early work reported here. There will be robots that talk and, above all, listen to instructions, and immobile computer-tutors that also talk and listen, and in the process teach the way a good tutor should.

References

Avrutin, S. & Wexler, K. (1993) Developments of principle B in Russian: coindexication at LF and coreference. *Language Acquisition* 2, 259-306.

Chomsky, N. (1959) Review of B. F. Skinner *Verbal Behavior*. *Language* 35, 26-58.

Feldman, J. A., Lakoff, G., Stolcke, A., & Hollbach Weber, S. (1990) Miniature Language Acquisition: A touchstone for Cognitive Science. In *Proceedings of the 12th Annual Conference of the Cognitive Science Society*, 686-693. MIT, Cambridge, Mass.

Feldman, J. A., Lakoff, G., Bailey, D., Narayanan, S., Regier, T., Stolcke, A. (1995) L_0 - The First Four Years. *AI Review* 8, to appear.

Hyams, N. & Wexler, K. (1993) On the grammatical basis of null subjects in child language. *Linguistic Inquiry* 24, 421-459.

Kaplan, R. & Bresnan, J. (1982) Lexical-functional grammar: a formal system for grammatical representation. In *The Mental Representation of Grammatical Relations*. ed. by J. Bresnan, Cambridge, Mass, 173-281.

Langley, P., & Carbonell, J. G. (1987) Language Acquisition and Machine Learning. In *Mechanisms of Language Acquisition*. edited by B. MacWhinney. 115-155. Hillsdale, NJ: Erlbaum.

Pinker, S. (1984) *Language Learnability and Language Development*. Cambridge, Mass: Harvard University Press.

Pinker, S. (1989) *Learnability and Cognition*. Cambridge, Mass: MIT Press.

Siskind, J. M. (1992) *Naive Physics, Event Perception, Lexical Semantics, and Language Acquisition*. M.I.T. Ph.D. Dissertation.

Siskind, J. M. (1994) Lexical Acquisition in the Presence of Noise and Homonymy. *Proceedings of the Twelfth National Conference on Artificial Intelligence. AAAI-94*, Vol. I, 760-766.

Skinner, B. F. (1959) *Verbal Behavior*. New York: Appleton.

Suppes, P. (1973) Congruence of meaning. *Proceedings and Addresses of the American Philosophical Association* 46, 21-38.

Suppes, P. (1991) *Language for Humans and Robots*. Oxford: Blackwell.

Suppes, P., Böttner, M., & Liang, L. (1995) Comprehension grammars generated from machine learning of natural languages. *Machine Learning* 19(2) (To appear). [Published in preliminary form in *Proceedings of the Eighth Amsterdam Colloquium, December 17-20, 1991*, edited by P. Dekker and M. Stokhof, 93-112. Institute for Logic, Language and Computation, University of Amsterdam 1992.]

Suppes, P., Böttner, M. & Liang, L. (to appear) Machine Learning comprehension grammars for ten languages.

Suppes, P. & Ginsberg, R. (1963) A fundamental property of all-or-none models, binomial distribution of responses prior to conditioning, with application to concept formation in children. *Psychological Review* 70, 139-161.

Suppes, P., & Liang, L. (in press) Concept Learning Rates and Transfer Performance of Several Multivariate Neural Network Models. In *Progress in Mathematical Psychology*, edited by C. Dowling, F. Roberts and P. Theuns.

Suppes, P., Liang, L., & Böttner, M. (1992) Complexity issues in robotic machine learning of natural language. In *Modeling Complexity Phenomena*, edited by L. Lam and V. Naroditsky, 102-127. New York: Springer.

Wexler, K., & Culicover, P. W. (1980) *Formal Principles of Language Acquisition*. Cambridge MA: MIT Press.