

Grammar Acquisition by Probabilistic Model Transformation

Eugene Koontz

Department of Computer Science
State University of New York at Buffalo

226 Bell Hall

Buffalo, NY 14260-2000

U.S.A.

(716) 645-6164 x 139

ekoontz@cs.buffalo.edu

1 Introduction

A crucial requirement of a linguistic theory is that it have a plausible theory of learnability. As [Green, 1994] points out, HPSG is an attractive framework in which to propose theories of language acquisition because it seems to require positing few language-specific learning mechanisms. Instead, language learning is seen as a specific instance of a general process of drawing distinctions and generalizations among the objects in the learner's experience.

In this paper, I present an implemented computational model of statistical HPSG grammar acquisition. The contributions of this paper are threefold. First, I show how language learning may be thought of as a sequence of transformations on a type hierarchy. Second, I develop methods of estimation by which two type hierarchies may be compared with respect to their simplicity and their ability to describe a language. Finally, I show how Mutual Information can be used to find optimal transformations among alternatives.

2 Background

[Stolke, 1994] has shown how grammars can be induced from corpora by means of model merging. By estimating the probabilities of a training corpus given a model, together with *a priori* biases toward simple models over complex ones, grammatical rules and constraints are combined in order to derive a more general grammar from one more specific to a training corpus.

A method of inducing HPSG-like grammars from a training corpus was presented in [Brew, 1994]. However, there was no consideration there for *a priori* preferences for simplicity over complexity in the induced grammar. In addition, the system of linguistic types is fixed; no types may be added or removed as a result of the training process. This paper addresses these two issues, and in doing so attempts to provide a realistic account of human language acquisition.

3 Architecture of the Acquisition System

3.1 Goals

The language learner is motivated by two opposing goals :

1. Constructing a simple model that expresses generalities of the language
2. Constructing a complex model that covers the diversity of the input

The first of these goals corresponds roughly to the notion of *explanatory adequacy*; the second to *descriptive adequacy*. In both the construction of a linguistic theory and the development of a grammar by a language learner, there is a fundamental tension between these two goals. An overly simple grammar will fail to describe exceptional phenomena of the input; an overly complex grammar will be redundant and fail to capture generalities of the language.

In statistical language learning of the type developed in [Stolke, 1994], these two goals are given a quantified interpretation. The first goal is achieved by maximizing $P(M)$, the *a priori* probability of the model M ; the second by maximizing $P(X|M)$, the *a posteriori* probability of the data X given the model.

Learning in such a system is by *Bayesian Inference*; the goal is to find the model that maximizes the quantity expressed in Bayes' Theorem : $P(M|X) = \frac{P(X|M) \times P(M)}{P(X)}$

Since the probability of the input ($= P(X)$) is the same for all models; it is a constant. We are in effect maximizing $P(M|X) = cP(X|M) \times P(M)$, where c is a constant.

As developed in [Stolke, 1994], this quantity is incrementally changed by the application of operations on the model that simplify the model and generalize it beyond its training set. In the acquisition system presented here, the model is a type hierarchy; the system learns by incrementally changing the topology of the type hierarchy.

3.2 Estimating $P(M)$ and $P(X|M)$

Learning begins with a type hierarchy of the kind considered by [Green, 1994] to be a plausible starting point for language learning. Figure 1 is an example of such a hierarchy.

Intuitively, one type hierarchy is more probable than another if it is simpler. This intuition may be quantified in terms of information theory – one type hierarchy is simpler than another if its description can be encoded in a shorter string. The probability of a type hierarchy is thus inversely proportional to its encoding length. Each symbol in the encoding refers to a type, a feature, or a relation of subsumption, and must be just long enough to uniquely identify its referent. n types will require $\log_2 n$ bits to distinguish them. We can recursively define the encoding length of a type hierarchy as the encoding length of its most general type (τ), where the encoding length is defined recursively as :

$$EL(\tau) = \log_2 n + \sum^i EL(\sigma_i) + \sum^j EL(\rho_j)$$

Here τ subsumes each of σ_i , and feature F_j having value of type ρ_j is appropriate to τ .

The probability of the data given the model ($P(X|M)$) relates the model to input data. This quantity acts as a barrier to over-generalization – a model that is too general will assign a low probability to its input, because the model will have too few constraints – it assigns high probabilities to too many ungrammatical signs.

Subdividing the input X into syntactic units such as sentences $\langle x_1, \dots, x_n \rangle$, $P(X|M)$ is defined as the product of the probabilities of each syntactic unit. The probability of a syntactic unit $P(x|M)$ is in turn defined as the sum of the probabilities of all possible derivations d_j of x according to the model : $P(x|M) = \sum^j P(d_j(x)|M)$.

Finally, the probability of a derivation $d_j(x)$ is defined much in the same way as in [Brew, 1994]; briefly, a derivation of sign x is a successive specification of a feature structure from \top until we have reached a maximally specific type that subsumes x .

3.3 Transformations on the Type Hierarchy

I now describe the operation on the type hierarchy, *feature raising*, which is fundamental to generalizing the model beyond its training corpus.

3.3.1 Feature Raising

In this type of transformation, a new type β is introduced as a supertype of a type α that bear a certain feature F . The new type also has this feature, and so the feature and its value have been “raised” from α to β .

Feature Raising Rule 1 Given γ subsuming $\alpha: \begin{bmatrix} F & \mu \\ \dots & \dots \end{bmatrix}$ introduce a new type $\beta: \begin{bmatrix} F & \mu \end{bmatrix}$ such that γ subsumes β and β subsumes α .

It turns out to be useful to allow sublists of type $list(\sigma)$ to be raised as the values of features in new supertypes; and therefore the following rule is formulated. The rule for sets is similar; I omit it here in the interest of brevity.

Feature Raising Rule 2 Given γ subsuming $\alpha: \begin{bmatrix} F < \lambda\mu\nu > \\ \dots & \dots \end{bmatrix}$, introduce a new type $\beta: \begin{bmatrix} F | \text{INFIX } < \mu > \\ \dots & \dots \end{bmatrix}$, such that γ subsumes β and β subsumes α , and replace α 's F feature value with $\begin{bmatrix} \text{PREFIX } < \lambda > \\ \text{SUFFIX } < \nu > \end{bmatrix}$, where λ, μ, ν are of type $list(\sigma)$.

The notation **raise** $(\alpha, F, \gamma) \rightarrow \beta$ will be used below to indicate a feature raising operation, where α, F, β, γ are as above.

3.4 Criteria for application of transformations

The **raise** $(\alpha, F, \gamma) \rightarrow \beta$ transformation is the driving force in grammar induction as it is presented in this paper. However, It must be restricted to particular situations in which it causes the grammar to

incorporate a generalization about the input. I here present briefly a method for determining the optimal feature raising to apply at a particular point in the grammar's development. In order to do this, I employ the concept of *Mutual Information*, as it is presented in [Magerman and Marcus, 1990]. Briefly, the mutual information of two events x and y is the log ratio of the probability of x and y occurring together over the product of x 's and y 's separate probabilities of occurrence.

Intuitively, given a type $\alpha: [F \ \mu]$ subsumed by γ , feature F should be attributed to a new supertype β of α if $[F \ \mu]$ is encountered not only in α , but in other types subsumed by γ . In other words, if $[F \ \mu]$ is rare outside of instances of α , the mutual information between α and F will be high; if it is common among other types, the mutual information between α and F will be lower due to F 's occurrence in other subtypes of γ . The optimal feature raising (if any) for a particular type hierarchy can thus be found by searching for the triple $\langle \alpha, F, \gamma \rangle$ for which

$$MI(\alpha, [F \ \mu]) = \log \frac{P(\alpha, [F \ \mu])}{P(\alpha)P([F \ \mu])}$$

is minimized.

4 Application of Feature Raising to English Verb Morphology

I now show how model merging by transformation may be used to acquire English present tense verb morphology. Suppose the system is exposed to a number of examples of present tense verbs. For each type of word in the input, a new type is introduced. After hearing a certain number of such verbs, the type hierarchy will have changed from Figure 1 to Figure 2.

After a certain number of such inputs, mutual information quantities associated with the common `CONT` and `PHON` values will cause the system to perform **raise** () , placing these common values in the values of new supertypes. For example, one transformation necessary to derive Figure 3 from Figure 2 is **raise** (*walks*,`CONTENT`,*walk-stem*) \rightarrow *sign*.

A sequence of transformations based on mutual information can be shown to exist that will change the topology of the type hierarchy from Figure 2 to Figure 3. Note that this exposition has been simplified considerably for this abstract. There are a few non-probabilistic transformations (such as type introduction) not discussed which are necessary for the learning described here.

5 Conclusion

In this paper, I have presented a computational model of language learning within the framework of HPSG. The model treats learning as a process of successive generalizations over information provided in the input. It provides a theory for how acquisition of an HPSG-like grammar could occur based on constraints of simplicity and corpus likelihood maximization.

References

[Brew, 1994] Brew, C. (1994). Stochastic HPSG. presented at Edinburgh HPSG Seminar/Workshop.

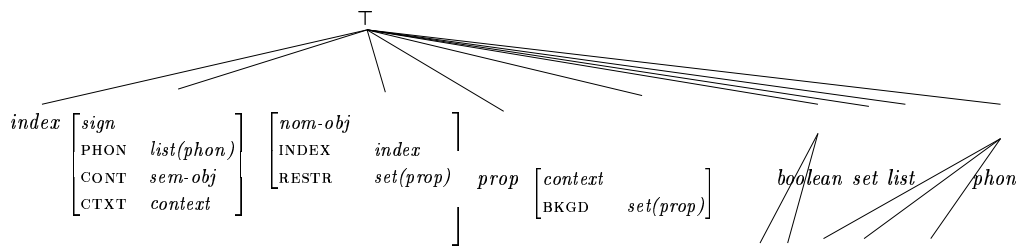


Figure 1: The initial type hierarchy + - /a/ /b/ ... /z/

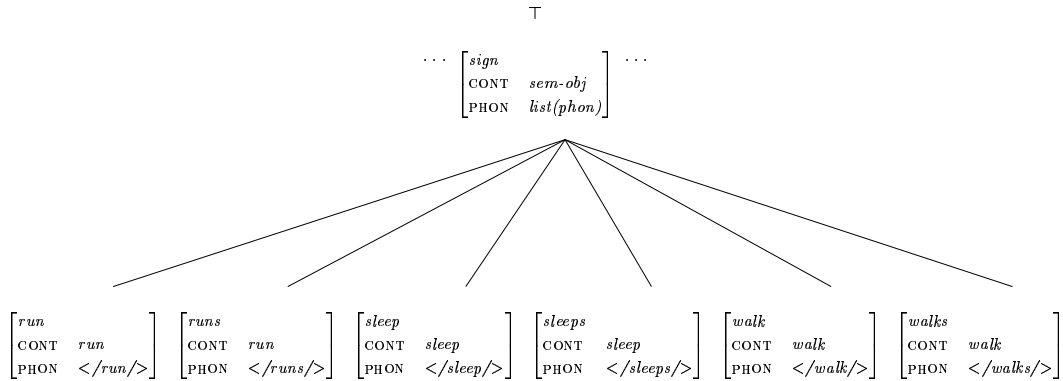


Figure 2: Type hierarchy after initial exposure to data (not all structure shown)

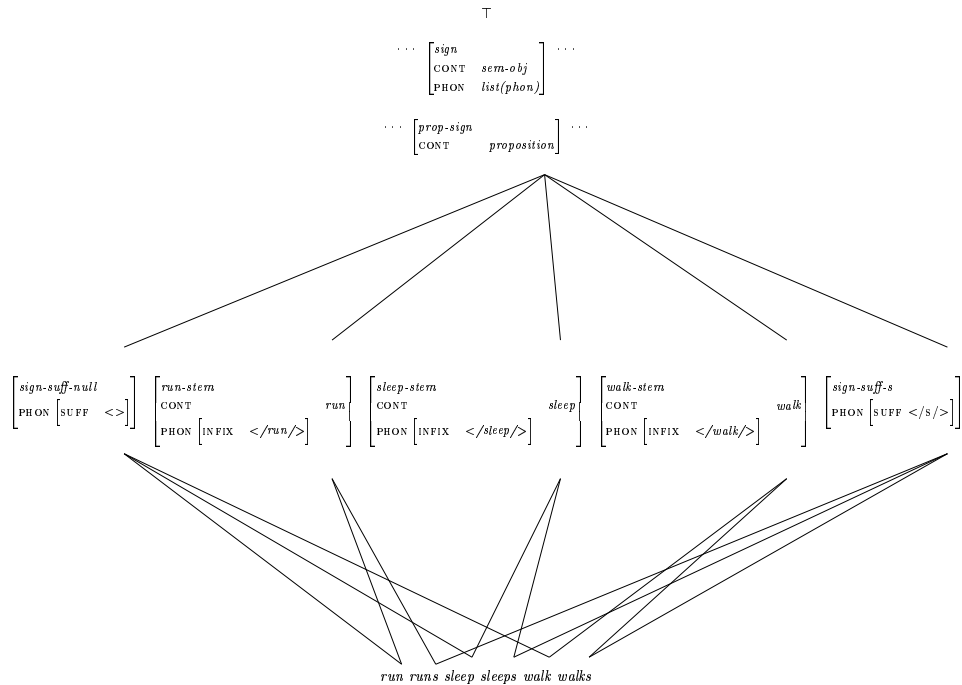


Figure 3: Type hierarchy after several **raise()** operations (not all structure shown)

- [Green, 1994] Green, G. M. (1994). Modelling grammar growth; universal grammar without innate principles or parameters. Unpubl. ms., University of Illinois.
- [Magerman and Marcus, 1990] Magerman, D. and Marcus, M. (1990). Parsing a natural language using mutual information statistics. In *Proceedings of AAAI-90*, Boston, MA.
- [Stolke, 1994] Stolke, A. (1994). *Bayesian Learning of Probabilistic Language Models*. PhD thesis, UC Berkeley.