# Approximate Maximum-Entropy Integration of Syntactic and Semantic Constraints

**Dekai Wu***
Department of Computer Science
University of Toronto
Toronto, Canada M5S 1A4
*dekai@cs.toronto.edu*

## Abstract

Statistical approaches to natural language parsing and interpretation have a number of advantages but thus far have failed to incorporate compositional generalizations found in traditional structural models. A major reason for this is the inability of most statistical language models being used to represent *relational* constraints, the connectionist variable binding problem being a prominent case. This paper proposes a basis for integrating probabilistic relational constraints using maximum entropy, with standard compositional feature-structure or frame representations. In addition, because full maximum entropy is combinatorically explosive, an approximate maximum entropy (AME) technique is introduced. As a sample problem, the task of integrating syntactic and semantic constraints for nominal compound interpretation is considered.

## Introduction

The importance of statistical methods in natural language processing has recently been rediscovered for a number of reasons. A major motivation is the automated acquisition of requisite information for language processing. The hope from the engineering standpoint is to bypass impractically intensive manual analysis by mechanically analyzing large online corpora (Church and Hanks 1990; Hindle 1990; Magerman and Marcus 1990; Smadja 1991). On the other hand, from the cognitive standpoint, statistical models could overcome many of the traditional difficulties faced by pure logic-based language acquisition models, such as learning without negative evidence.

Another of the statistical paradigm's main advantages is that quantitative measures facilitate integrating factors along many different dimensions, a difficult problem in purely symbolic parsing and interpretation models. Syntactic, lexical, semantic, conceptual, and many other factors all enter in as tendencies rather than rules (Marslen-Wilson and Tyler 1987). Integrating different knowledge sources is particularly important for disambiguation problems that arise with nominal compounds, adverbial modification, and prepositional phrase attachment.

A third major attraction is the grounding of numeric measures. Since Quillian (1969), many quantitative approaches to language understanding have been proposed. However, a major problem with all these approaches is the assumption of large numbers of *ad hoc* weights (Wu 1989). To make it plausible that a quantitative model will generalize to interestingly large domains, the numbers should be justified on some statistical basis; that is, the weights in principle should be derivable from a set of sample points.

It is a weakness of many statistical NLP proposals, however, that they do not yet exploit the advances of several decades of structural theories. Any structure must be automatically induced by the models, given only the surface features of the input strings. On the plus side, this can act as independent validation of linguistic theories if the same structural categories are induced. But on the minus side, the complexity and kinds of structures are heavily constrained by the induction model chosen. The current models do not induce generalizations involving the sorts of interacting structural constraints in knowledge-based parsers and semantic interpreters. Grammar induction methods, for example, work with probabilistic context-free formalisms (Lari and Young 1990; Fujisaki *et al.* 1991) rather than unification-based formalisms. Despite impressive successes, solving the hardest, "last 10%" problems always seems to demand the additional

*Beginning Fall 1992: Dept. of Computer Science, Hong Kong University of Science and Technology. Much of this research was done at the Computer Science Division, University of California at Berkeley and was sponsored in part by the Defense Advanced Research Projects Agency (DoD), monitored by the Space and Naval Warfare Systems Command under N00039-88-C-0292, the Office of Naval Research under contract N00014-89-J-3205, and the Sloan Foundation under grant 86-10-3. Preparation of this document was partially supported by the Natural Sciences and Engineering Research Council of Canada.

structure. Structural generalizations, if known, should be built into the induction model *a priori* rather than discarded.

In part efforts along these lines have been hindered by one major deficiency among the statistical methods presently being employed. This is the difficulty of integrating constraints arising from multiple knowledge sources when there are *relational constraints* of the kind found in symbolic models such as unification-based grammars. In this paper I propose a statistically grounded model that can integrate probabilistic relational constraints. Two novel contributions are presented:

- An idealized maximum-entropy treatment of evidential inference in a hierarchical, *compositional* feature-structure space.
- An approximate maximum-entropy (AME) technique that estimates conditional distributions by making *structural* approximations to the ideal maximum-entropy case.

## Evidential Interpretation

Consider the nominal compound *coast road*.[1] An informal survey produced as the most common interpretations a road (either generic or Highway 1) in or along the coastal area. A less preferred interpretation was a road amenable to coasting. In addition, though no informant volunteered the interpretation of a road leading to the coast, all agreed when asked that this was perfectly normal in contexts like *Since the earthquake damaged the only Interstate to the coast, old Highway 17 will temporarily serve as the main coast road.* We will use *coast road* as an example throughout this paper.

**Relational constraints and feature-structures** The choice of structures is at least as critical to the success of a model as any probabilistic constraint combination method, particularly since we are not inducing the structures themselves but choosing them *a priori*. However, as this is not the focus of the present paper I only summarize the assumptions here; details may be found in Wu (1992). Structures vary in both domain and specificity. A *modular ontology* divides representational primitives into a number of linguistically-, psychologically-, and neurologically-motivated modules. These include mental images, lexical semantics, lexicosyntactic constructions, and the conceptual system. Intermodular structures associate structures across modules. The notion of ontological modularity is a significant weakening of Fodor's (1983) process modularity, insofar as the same processes may span all modules. In the model proposed these processes are probabilistic.

Consider the types of information needed to correctly interpret *coast road*:

1. *Specific lexical signification*: The word *coast* used as a noun means a seacoast substantially more often than an unpowered movement.[2]
2. *Abstract semantic schemas and construction signification*: Prototypical spatial relationships between a one-dimensional entity (*road*) and an elongated two-dimensional space (*coast*) include parallel containment and linear order locative (i.e., destination). Nominal compounds are frequently used to express containment relationships, and somewhat less frequently to express linear order locative relationships.
3. *Intermediate conceptual schemas*: Most of the time when one thinks about roads in the context of seacoasts, one thinks not of generic roads but specifically of the subcategory of roads running along the coast (let's abbreviate that as *coastal road*).
4. *Specific conceptual schemas*: Living on the West Coast, *Highway 1* is a frequently used concept of the *coastal road* subcategory.

The proposed model represents all structures uniformly using standard unification grammar *typed feature-structures*,[3] which can also be thought of as frames, constraints, or templates. Figure 1 shows examples corresponding to the first two of the above knowledge types in (a) and (b).[4] The structured format and co-indexing mechanism (the superscripts) permit complex relational constraints to be represented (Shieber 1986). The uniformity, as we see below, facilitates constructing a consistent underlying probabilistic event space. Feature-structure syntax implicitly defines a partially-ordered hierarchical space (Shieber 1986). To eliminate redundancy the feature-structures are actually stored using MURAL, a terminological inheritance hierarchy in the style of KL-ONE.

[1] From the Brown corpus (Kučera and Francis 1967). In keeping with the healthy movement toward working on shared data, I have been using the same Brown corpus data as Warren's (1978) study of some 4,500 nominal compounds. My views on nominal compound patterns are discussed in Wu (1990); nominal compounds have a long history in linguistics (e.g., Lees 1963, 1970; Downing 1977; Levi 1978; McDonald 1982; Leonard 1984).

[2] There are intermediate lexicosyntactic signification patterns that do not appear in the examples here because of space limitations. For example, the construction *coast N* is often designates something related to a seacoast, as in *coast artillery, coast guard, coastland, coastline,* and *coast redwood.* Nominal compound constructions involving the unpowered movement sense tend to use *coaster* instead, as in *coaster wagon, coaster brake,* and *roller coaster.*

[3] With a couple of extensions that are non-essential for present purposes (Wu 1992).

[4] LM and TR stand for landmark and trajector, and denote image-schematic ground and figure roles (Talmy 1983, 1985; Lakoff 1987; Langacker 1987; Feldman *et al.* 1990).

Since hard constraints are often too rigid, researchers have sought to modify such frameworks with quantitative certainty or preference measures that yield "soft" relational constraints. From a purely qualitative point of view, it is as easy to obtain interacting relational constraints as simple feature constraints, for example by extending marker passing methods with weights. As mentioned above, however, the numbers in such models are not well grounded.

The relational constraint problem appears in different guises. In neural networks it is related to the *variable binding problem*. Broadly speaking, there are four approaches to the variable binding problem. One is to construct networks on the fly, instantiating and linking fragments as needed to accomodate new bound constants (e.g., Wermter 1989). This approach is related to weighted marker passing (Wu 1989), and again the problem is to give a statistical justification to the numbers. Since the representation in an instantiation scheme is necessarily structured rather than distributed, it is non-trivial to apply known methods for learning weights such as backpropagation. The other three approaches are statistically grounded, but have not been shown to learn generalizations over compositional structures very well. One strategy, as exemplified by $\mu$KLONE (Derthick 1990) or Hinton's (1981) and Touretzky's (1990) triples, is to store explicit binding or relational information into the network. Another approach is to expand the size of the network to allow all possible binding permutations, as with nonsaturated tensor product representations (Smolensky 1990). Finally, methods employing Hinton's (1990) notion of reduced descriptions recursively compress structures into fixed-width vector representations. These include Pollack's (1990) RAAM and Plate's (1991) Holographic Reduced Representations. It is not clear that any of these schemes capture structural generalizations, though some preliminary empirical investigations indicate that certain variants of RAAM do capture at least treelike regularities (Stolcke and Wu 1992). Simply storing compositional structures is not enough; the representation must allow processing generalizations over compositionally similar structures.

**A metarepresentational interpretation of probabilities** Probabilistic models are only statistical models if given an interpretation based on sampling. One probabilistic method of rating competing interpretation structures in a semantic network framework is proposed by Goldman and Charniak (1990a, 1990b; Charniak and Goldman 1988, 1989). The model employs Bayesian belief networks (Pearl 1988), including hypothesis nodes representing the binding of one structure to some role of another. The use of probability theory is a promising step since probabilities are customarily estimated by statistical sampling methods, thus grounding the numbers by giving them a derivational interpretation. However, Goldman does not suggest such an interpretation, and we cannot assume any sampling method without knowing, for example, whether probabilities are to represent objective real-world relative frequencies or subjective belief measures.[5] Similar comments may also turn out to apply to less explicitly probabilistic models such as Hobbs *et al.*'s (1988) weighted abduction model.

Except for lexical items, the structures we are dealing with are intermediate conceptual structures; consequently an interpretation based on sampling real-world physical events is not well-founded.[6] I propose a metalevel interpretation where a probability denotes how likely it is that some conceptual structure will be *useful* to the linguistic agent. These values are associated with feature-structures in the knowledge base using the PROB attribute as in figure 1. Unlike the usual AI interpretation of belief nets or probabilistic logics, probabilities are not directly available to the agent and do not represent estimates of real-world frequencies. This is consistent with Kahneman *et al.*'s (1982) finding that humans are not good at reasoning with probabilities.

The metalevel interpretation of probability is formulated in Wu (1992) using Russell and Wefald's (1991) limited rationality framework; space does not permit proper treatment here. Informally, parsing and semantic interpretation are viewed as a form of adaptive forward inference. (Actually I am concerned only with the non-attentional part of interpretation which I call *automatic inference* after the psychological distinction between *automatic* and *controlled* processes.) The interesting statistical subprocess is the compilation mechanism responsible for observing samples—i.e., input utterances and their eventual interpretations, arrived at by either supervisory training or functional context—and learning which conceptual structures most frequently turn out to be useful to infer given contextual cues (adaptation by more quickly "jumping to conclusions"). This mentalist interpretation reconciles probabilities, philosophically at least, with statistical sampling. Later I will discuss more practical possible estimation approaches.

**Conditioning on the input event** Parsing and interpretation are formulated in evidential terms. The input utterance[7] constitutes the conditioning event e. Figure 1 shows the input structure for *coast road* in (c). The desired output is the conceptual structure with the maximum conditional probability $P(q_i|e)$, a structure such as (d). In the current formulation the output structure must include as a subpart the entire

[5] See Hacking (1975); Weatherford (1982); Bacchus (1990).

[6] Unless one supposes physical brain-states can be sampled.

[7] Plus the context, if any; contextual priming is accomodated but not discussed here.

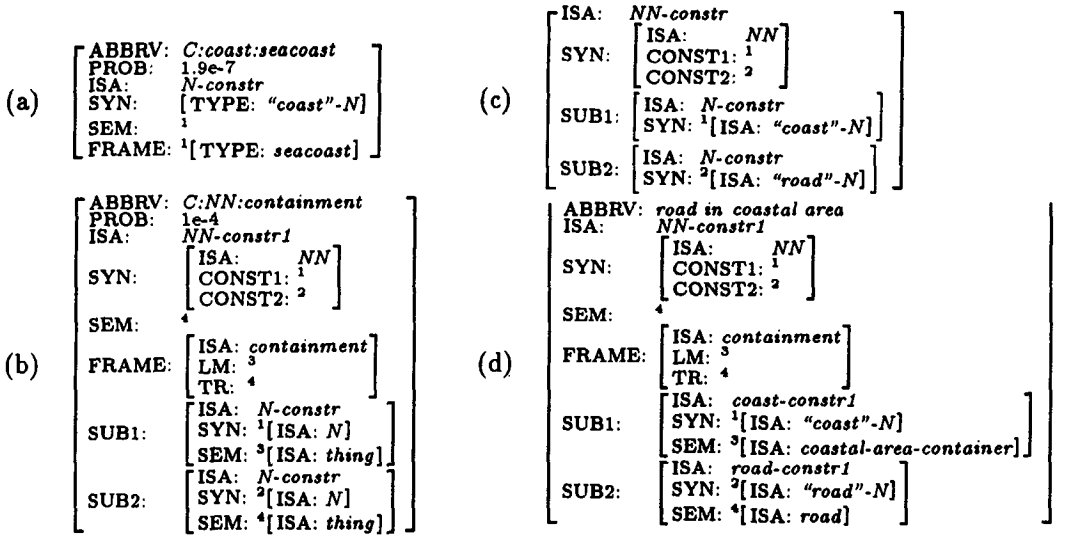


Figure 1: Feature-structures for (a) relational constraint 1 above, (b) relational constraint 2 above, (c) an input form, and (d) a full output structure (floor brackets indicate a *complete event*, defined below).

input structure, as well as a parse tree for the input string.

For present purposes I will assume that conditioning is performed by a marker-passing process or some similar hypothesis generator. That is, some coarse heuristic produces a "first cut" set of hypotheses as to the output structures considered most likely.[8] Conditioning is performed as a by-product, because only structures consistent with the input structure are hypothesized. That is, the hypothesis space is a subset of the conditional space. The task is then to compute the portion of the probability distribution over this conditional space, and to select the hypothesis with maximum probability.

## Model I: Maximum Entropy

The knowledge base contains probabilities for structures like (a) and (b) in figure 1. These structures are not full output structures but rather fragments that might be unified into a full output structure. Each such structure thus determines an abstract class of all the full output structures in which it is included.

We atomize the probability space as follows. Each possible full output structure $q_i$ corresponds to one of the set of exhaustive and disjoint events, and their probabilities $P_i$ must sum to unity. These are called *complete events*. Any set of hypotheses is therefore a set of complete events $\{q_i\}$. The probabilities of structures like (a) and (b) are marginal constraints specifying the sum probability over classes of $q_i$'s. These structures are *abstract events*.[9] What is stored, thus, is in fact only partial information about the distribution over the hypothesis space, because the marginals alone do not in general determine a unique distribution.

The maximum entropy principle (Jaynes 1979) is a canonical method that yields a unique completion of a partially constrained distribution. According to the principle, the distribution should be chosen to maximize the information-theoretic entropy measure

$$H = -\sum_{i=1}^{C} P_i \log P_i$$

This supplies the missing parts of the distribution in a least-informative manner. To solve the maximization problem Cheeseman's (1987) method can be generalized to the feature-structure space rather than the flat feature-vector space, as shown in Wu (1992). Applying Lagrange multipliers yields a system of constraints with the same number of unknowns and constraints, which can then be solved by a successive line minimization procedure.

This constitutes the core theory. More powerful generalizations can be expressed in the formulation than in flat feature space models. Those models can express conditional independence between features, but

[8] Abstractly, the hypothesis generator should produce all possible interpretations of the input along with all known constraints on the probability distribution over them. In fact, to do this for an interestingly large knowledge base would far exceed resource bounds. Instead hypothesis generation is assumed to produce only the most pertinent structures.

[9] In probabilistic terminology, complete and abstract events are *simple* and *compound* events. I avoid the terms here because simple events correspond to more complex feature-structures and compound events correspond to simpler structures.

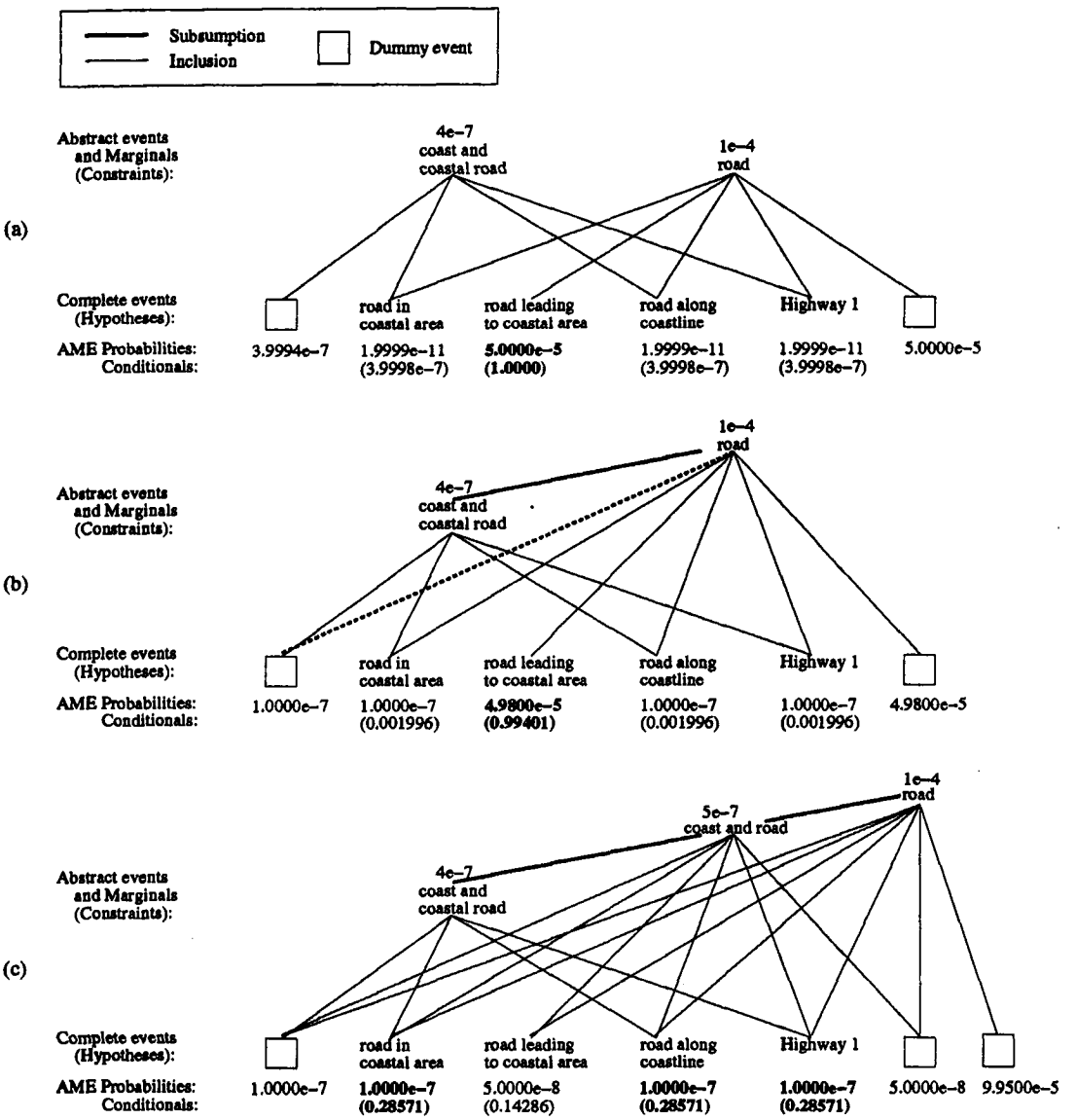


Figure 2: Approximate maximum entropy (AME), (a) with dummy events, (b) with proper subsumption, (c) with specific constraint.

only allow generalizations (marginal probabilities) over classes delineated by features. Featural independence reduces to the maximum-entropy principle as a special case; however, maximum entropy also allows generalization over compositionally similar classes. To the best of our knowledge, the application of general maximum entropy to a hierarchical compositional event space is new.

Several advantages result over a number of the statistical approaches mentioned earlier, including belief nets, $\mu$KLONE, and triples. The compositionally-structured event space eliminates explicit features (or hypothesis nodes) for variable bindings. Instead, the similarity between feature-structures with similar binding patterns is expressed in the subsumption lattice. In representations that employ explicit binding features, estimation of conditional probability matrices for is susceptible to inconsistencies because binding hypotheses interact combinatorically: each binding invalidates some subset of the other possible bindings. By leaving binding hypotheses implicit, the proposed model's representation can store marginal rather than conditional probabilities, which are relatively easy to keep consistent even with relational constraints. Similarly, binding hypotheses cause loops in belief nets that lead to highly interconnected constraints, making evaluation particularly expensive since there are no conditional independences to exploit.

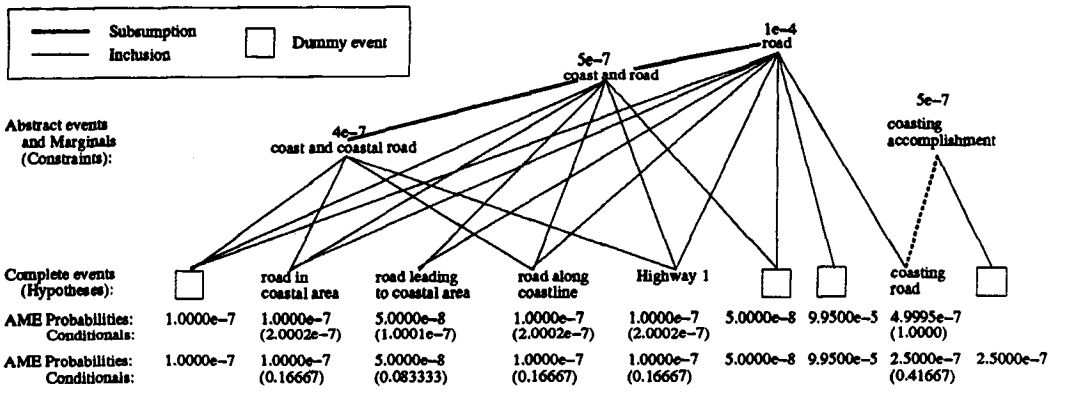


Figure 3: Results with an added hypothesis (a) without and (b) with specific constraint on *coasting accomplishment*.

The problem also manifests itself in the maximum entropy model because the full space of $q_i$'s is combinatoric, making full-fledged maximum entropy over the entire space infeasible. However, as discussed in the next section, it is possible to make *structural* approximations.

## Model II: Approximate Maximum Entropy

We would like to cut the space down by considering only part of the conditional space, i.e., the hypotheses and marginal constraints deemed pertinent by the heuristic hypothesis generator or marker passer. However, maximum entropy cannot simply be applied to the conditional space because the marginal constraints are defined over the entire space; indeed the marginals usually turn out to be internally inconsistent if one tries to interpret them over just the conditional space. (For instance, in the later example of figure 7 there is no consistent assignment of probabilities meeting the marginal constraints, without the "dummy" events to be proposed.) Thus an approximation method is needed that can be used over the conditional space.

To estimate the true maximum-entropy distribution for the full combinatoric space, I propose to use the same maximum-entropy mechanisms in a coarser space. We are only interested in ranking hypotheses within the conditional space. The essence of the approximation is to discard the details of how the event space is structured outside the immediate hypothesis space—what I will call the *complement space*. The complement space is entirely covered by a small number of "dummy" events that correspond to abstract events but are treated as if they were complete events. There are thus few enough events to be tractable. At the same time the dummy events make the marginals consistent, by providing nonzero subspaces for those events that have been counted into the marginals but are inconsistent with the hypothesis space.

Figure 2(a) shows the simplest approach one might take. Each node's label corresponds to the ABBRV of some abstract or complete feature-structure, and the arcs denote subsumption. A minimum of dummy events are used. One dummy event is needed for each marginal constraint considered pertinent, i.e., produced by the hypothesis generator (*coastal road* and *road*). Each dummy event represents all the complete events (full output structures) that are consistent with the corresponding constraint, but not with the conditional space. A single unshown *null event* represents all remaining events bringing the total probability to unity. Entropy can be maximized consistently over such a space giving the first row of probabilities at the bottom; the conditional distribution is obtained by normalizing over the hypothesis space as shown in parentheses. However, these numbers are unreasonable because of the crudeness of the structural approximation.

The approximate maximum entropy method (henceforth AME) prescribes two principles for constructing the hypothesis space to maintain approximation accuracy, described below. How closely full maximum entropy can be approximated depends on the categories whose marginals are constrained. Discrepancies in the approximation arise from the fact that the dummy events are treated as being disjoint even though they stand for event spaces that may overlap. In results to date the discrepancies have proved insignificant but larger experiments will be useful. .

The *proper subsumption* principle dictates that when one marginal constraint is superordinate to another in the original space, the dummy event for the subordinate in the approximate space should also be included in the superordinate. As shown in figure 2(b), this results in a correction of several orders of magnitude in the conditional probabilities.

The *most-specific constraint* principle enforces that the most specific applicable marginal constraints available from the database always be included. Figure 2(c) shows another large correction from including a more specific marginal that constrains the total amount of probability assignable to the hypotheses, causing much

Suppose $\mathcal{Q}$ is the set of complete (token) f-structures and $\mathcal{G}$ is the set of abstract (type) f-structures, and $\mathcal{F} \stackrel{\text{def}}{=} \mathcal{G} \cup \mathcal{Q}$. Let $\mathcal{H} = \{h_1, \ldots, h_i, \ldots, h_H\} \subset \mathcal{Q}$ be the candidate output structures produced by the hypothesis generator, and let $\mathcal{M} = \{m_1, \ldots, m_j, \ldots, m_M\} \subset \mathcal{G}$ be the pertinent abstract classes with associated marginal probabilities $P_{m_j} = P(m_j)$ from the constraint generator. Denote by $\sqsubset$ the partial ordering induced on $\mathcal{H} \cup \mathcal{M}$ by the subsumption lattice on f-structure space.

We define $\mathcal{D} = \{d_1, \ldots, d_j, \ldots, d_M\}$ as the set of dummy events. Let $\dot{\mathcal{F}} \stackrel{\text{def}}{=} \mathcal{H} \cup \mathcal{M} \cup \mathcal{D}$ be the *approximate event space*, and let $\dot{\mathcal{H}} \stackrel{\text{def}}{=} \mathcal{H} \cup \mathcal{D}$ be the *approximate hypothesis space*. Define the *approximate ordering relation* $\dot{\sqsubset}$ over $\dot{\mathcal{F}}$ as follows:

$$\begin{cases} a \dot{\sqsubset} b, & \text{if} \begin{cases} a \sqsubset b; a, b \in \mathcal{F} \\ a = m_j; b = d_j \\ a \sqsubset c; c = m_j; b = d_j \end{cases} \\ a \not\dot{\sqsubset} b, & \text{otherwise} \end{cases}$$

Let $\dot{P}_{m_j}$ be the marginal probability constraints on $\mathcal{F}$ and use $P_{m_j}$ as estimators for $\dot{P}_{m_j}$.

Assign $\hat{P}_{h_i}$ and $\hat{P}_{d_i}$ such that

$$\sum_{q \in \dot{\mathcal{H}}} \hat{P}_q = 1$$

while maximizing the entropy

$$E = -\sum_{q \in \dot{\mathcal{H}}} \hat{P}_q \log \hat{P}_q$$

subject to the marginal constraints

$$\sum_{q: q \in \dot{\mathcal{H}}, m_j \dot{\sqsubset} q} \hat{P}_q = \dot{P}_{m_j}$$

Figure 4: The AME (approximate maximum entropy) method.

of the probability weight to be shifted to the dummy events and switching the preferred hypotheses. Another example is shown in figure 3, where including the additional marginal on *coasting accomplishment* corrects the wrong preference for *coasting road*.[10] The numbers here are still incorrect as not all the pertinent constraints have been included yet. In particular, no lexicosyntactic constructions representing conventional linguistic usage contraints are present (without these we are simply predicting that *coasting accomplishment* is the more often-used concept.)

Specifics of the AME method are shown in figure 4. Formally, a second entropy-maximization problem is derived where the only detailed structure lies within the hypothesis space. Marginals from the original domain are used as estimators for marginals in the approximate space. As in Model I this yields a system of constraints as derived in figure 5. The numerical method used to solve the system is shown in figure 6.

We now consider a full case of how AME integrates constraints, using the introductory example. Suppose the set of constraints and hypotheses deemed pertinent are as shown in figure 7. We ignore the numbered boldface marginals for the basic case. The reader may verify that constraints (1)–(4) are qualitatively encoded by comparing the relative values of the marginals. For instance, to encode (2), the marginal on *C:NN:containment*—a noun-noun construction used to signify containment—is twice that of *C:NN:linear order locative*. For the basic case AME yields the conditional distribution in the uppermost row labelled "0:". The hypotheses *road in coastal area* and *road along coastline* are the winners with the highest probabilities, as may reasonably be expected.

Four additional runs are shown to demonstrate the effect of each constraint that is integrated into the conditional distribution. For each of constraints (1)–(4), the relevant abstract events are marked with an alternate marginal probability in boldface and labelled by the number of the constraint. Suppose constraint (1) weren't true, and *"coast"* were actually used more often to mean unpowered movement than a seacoast. Then switching the marginals on *"coast"* signifying *seacoast* and *coasting accomplishment* as shown produces the conditional distribution in row 1, where *coasting road* now dominates the hypotheses. The reader may similarly verify the effect of each of the other constraints. The examples were computed us-

[10] An *accomplishment* is a subkind of action that takes its name from Bach's (1986) work on aspect.

Define a new energy function $J$ to be minimized:

$$J \stackrel{\text{def}}{=} E + \sum_{j=1}^{M} \lambda_j (\dot{P}_{m_j} - \sum_{q:q\in\dot{\mathcal{H}},m_j\dot{\sqsubset}q} \hat{P}_q) = -\sum_{q\in\dot{\mathcal{H}}} \hat{P}_q \log \hat{P}_q + \sum_{j=1}^{M} \lambda_j (\dot{P}_{m_j} - \sum_{q:q\in\dot{\mathcal{H}},m_j\dot{\sqsubset}q} \hat{P}_q)$$

Observe that setting the gradients to zero gives the desired conditions:

$$\nabla_\lambda J = 0 \;\Rightarrow\; \frac{\partial J}{\partial \lambda_j} = 0; 1 \le j \le M \;\Rightarrow\; \text{expresses all marginal constraints}$$
$$\nabla_{\hat{\mathbf{P}}} J = 0 \;\Rightarrow\; \frac{\partial J}{\partial \hat{P}_q} = 0; q \in \dot{\mathcal{H}} \;\Rightarrow\; \text{maximizes entropy}$$

Since the partials with respect to $\hat{\mathbf{P}}$ are

$$\frac{\partial J}{\partial \hat{P}_q} = -\log \hat{P}_q - \sum_{j:m_j\dot{\sqsubset}q} \lambda_j$$

then at $\nabla_{\hat{\mathbf{P}}} J = 0$,

$$\log \hat{P}_q = -\sum_{j:m_j\dot{\sqsubset}q} \lambda_j$$

Defining $\omega_j \stackrel{\text{def}}{=} e^{-\lambda_j}$,

$$\hat{P}_q = \prod_{j:m_j\dot{\sqsubset}q} \omega_j$$

the original marginal constraints become

$$\dot{P}_{m_j} = \sum_{q:m_j\dot{\sqsubset}q} \prod_{k:m_k\dot{\sqsubset}q} \omega_k$$

which can be rewritten

$$\dot{P}_{m_j} - \sum_{q:m_j\dot{\sqsubset}q} \prod_{k:m_k\dot{\sqsubset}q} \omega_k = 0$$

to be solved using a numerical method.

Figure 5: Derivation of constraint system for AME.

ing a C implementation of AME with a symbolic user interface.

## Discussion

As mentioned, "improper" but practical probability estimation methods may suffice for interim applications. Lexical frequency counts over the Brown corpus and others are available (Francis and Kučera 1982) and parsed corpora will soon facilitate frequency counts for syntactic patterns as well. These counts may be taken as rough estimates of the frequency of an agent's use of the lexicosyntactic structures. The analogous procedure is not practical for semantic or conceptual structures, since fully interpreted corpora are not available. Warren's (1978) study contains frequency counts on manually-analyzed coarse semantic relation categories, but these must be massaged to fit more sophisticated AI ontologies.

Another potential use of large-corpora techniques has been suggested by Hearst (1991), who proposes an automated method for "coarse" disambiguation of noun homographs. Such a method, based on orthographic, lexical, and syntactic cues near the noun, may improve the relevance and accuracy rate of hypotheses.

Performance will depend heavily on which abstract events the investigator chooses to constrain the marginals for. In effect, the investigator decides the degree of *generalization*, because what maximum entropy does is to generalize the partial distributional information in the knowledge base to the rest of the event space. Choosing the abstract events is a kind of concept formation, which this work does not address, but toward this direction Wu (1991) proposes a theoretical distribution for modelling generalization from samples, related to discrete kernel estimation techniques. Choosing a set of marginal constraints can then be seen as a matter of best fit to the theoretical distribution.

The proposed model provides a probabilistic basis

1. Start with a constraint system $X \leftarrow \{\}$ and an estimated $\omega$ vector $\langle\rangle$ of length zero.
2. For each constraint equation,
   (a) Add the equation to $X$ and its corresponding $\omega_i$ term to $\langle\omega_1, \ldots, \omega_{i-1}, \omega_i\rangle$.
   (b) Repeat until $\langle\omega_1, \ldots, \omega_i\rangle$ settles, i.e., the change between iterations falls below some threshold:
      1. For each equation in $X$ constraining $\dot{P}_{m_j}$, solve for the corresponding $\omega_j$ assuming all other $\omega$ values have their current estimated values.

Figure 6: Numerical algorithm for solving the maximum-entropy constraint system.

for integrating relational constraints in parsing and semantic interpretation, using standard structural representations. Unlike most previous quantitative approaches, the probabilistic measures have a statistically grounded interpretation, thereby making it more plausible that the model can scale up to interestingly large domains. Moreover, the AME method addresses tractability and consistency issues that have been problematic for probabilistic models with relational constraints.

## Acknowledgements

Robert Wilensky, Jerome Feldman, and the members of the BAIR and $L_0$ groups contributed extensive valuable comments. Many thanks also to Graeme Hirst's and Geoff Hinton's groups.

## References

Bacchus, Fahiem 1990. *Representing and Reasoning with Probabilistic Knowledge: A Logical Approach to Probabilities.* MIT Press, Cambridge, MA.

Bach, Emmon 1986. The algebra of events. *Linguistics and Philosophy* 9:5–16.

Charniak, Eugene and Goldman, Robert 1988. A logic for semantic interpretation. In *Proceedings of the 26th Annual Conference of the Association for Computational Linguistics.*

Charniak, Eugene and Goldman, Robert 1989. A semantics for probabilistic quantifier-free first-order languages, with particular application to story understanding. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence.* 1074–1079.

Cheeseman, Peter 1987. A method of computing maximum entropy probability values for expert systems. In Smith, Ray C. and Erickson, Gary J., editors 1987, *Maximum-Entropy and Bayesian Spectral Analysis and Estimation Problems.* D. Reidel, Dordrecht, Holland. 229–240. Revised proceedings of the Third Maximum Entropy Workshop, Laramie, WY, 1983.

Church, Kenneth Ward and Hanks, Patrick 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics* 16(1):22–29.

Derthick, Mark 1990. Mundane reasoning by settling on a plausible model. *Artificial Intelligence* 46:107–157.

Downing, Pamela 1977. On the creation and use of English compound nouns. *Language* 53(4):810–842.

Feldman, Jerome A.; Lakoff, George; Stolcke, Andreas; and Weber, Susan Hollbach 1990. Miniature language acquisition: A touchstone for cognitive science. In *Program of the Twelfth Annual Conference of the Cognitive Science Society.* 686–693. Also available as technical report TR-90-009, International Computer Science Institute, Berkeley, CA.

Fodor, Jerry A. 1983. *The Modularity of Mind.* MIT Press, Cambridge, MA.

Francis, W. Nelson and Kučera, Henry 1982. *Frequency Analysis of English Usage: Lexicon and Grammar.* Houghton Mifflin, Boston. With the assistance of Andrew W. Mackie.

Fujisaki, T.; Jelinek, F.; Cocke, J.; Black, E.; and Nishino, T. 1991. A probabilistic parsing method for sentence disambiguation. In Tomita, Masaru, editor 1991, *Current Issues in Parsing Technology.* Kluwer, Boston. 139–152.

Goldman, Robert P. and Charniak, Eugene 1990a. Incremental construction of probabilistic models for language abduction: Work in progress. In *Working Notes from the Spring Symposium on Automated Abduction,* Stanford University, Stanford, CA. AAAI. 1–4.

Goldman, Robert P. and Charniak, Eugene 1990b. A probabilistic approach to text understanding. Technical Report CS-90-13, Brown Univ., Providence, RI.

Hacking, Ian 1975. *The Emergence of Probability.* Cambridge University Press, London.

Hearst, Marti A. 1991. Noun homograph disambiguation using local context in large text corpora. In *Seventh Annual Conference of the University of*

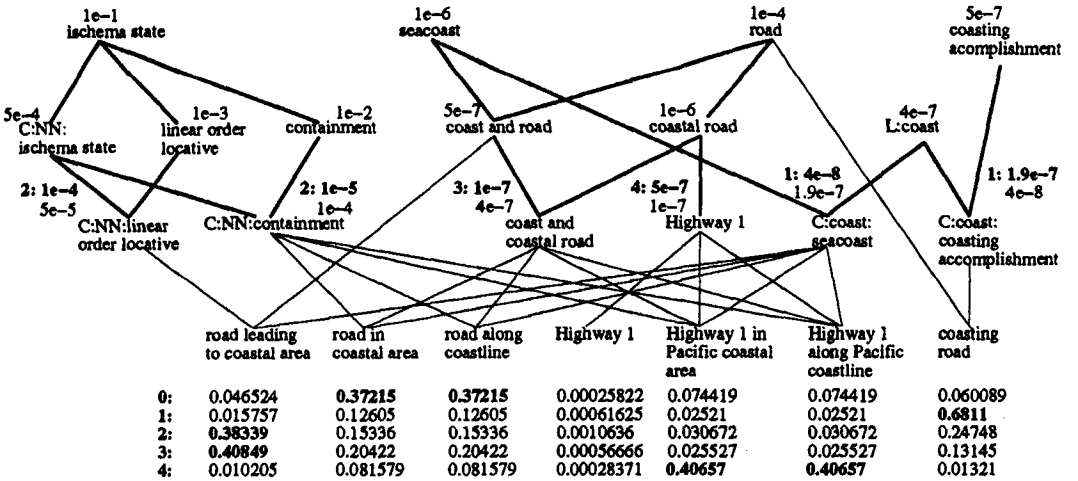


Figure 7: Base run for *coast road* plus four variations. The redundant inclusion arcs and dummy events are omitted and only conditional probabilities are shown.


*Waterloo Centre for the New OED and Text Research: Using Corpora*, Oxford. 1–22.

Hindle, Donald 1990. Noun classification from predicate-argument structures. In *Proceedings of the 28th Annual Conference of the Association for Computational Linguistics*, Pittsburgh, PA. 268–275.

Hinton, Geoffrey E. 1981. Implementing semantic networks in parallel hardware. In Hinton, Geoffrey E. and Anderson, John A., editors 1981, *Parallel Models of Associative Memory*. Lawrence Erlbaum Associates, Hillsdale, NJ. 161–188.

Hinton, Geoffrey E. 1990. Mapping part-whole hierarchies into connectionist networks. *Artificial Intelligence* 46:47–75.

Hobbs, Jerry R.; Stickel, Mark; Martin, Paul; and Edwards, Douglas 1988. Interpretation as abduction. In *Proceedings of the 26th Annual Conference of the Association for Computational Linguistics*, Buffalo, NY. 95–103.

Jaynes, E. T. 1979. Where do we stand on maximum entropy. In Levine, R. D. and Tribus, M., editors 1979, *The Maximum Entropy Formalism*. MIT Press, Cambridge, MA.

Kahneman, Daniel; Slovic, Paul; and Tversky, Amos, editors 1982. *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge University Press, Cambridge.

Kučera, Henry and Francis, W. Nelson 1967. *Computational Analysis of Present-Day American English*. Brown University Press, Providence, RI.

Lakoff, George 1987. *Women, Fire, and Dangerous Things*. University of Chicago Press, Chicago.

Langacker, Ronald W. 1987. *Foundations of Cognitive Grammar*, volume 1. Stanford University Press, Stanford, CA.

Lari, K. and Young, S. J. 1990. The estimation of stochastic context-free grammars using the inside-outside algorithm. *Computer Speech and Language* 4:35–56.

Lees, Robert B. 1963. *The Grammar of English Nominalizations*. Mouton, The Hague.

Lees, Robert B. 1970. Problems in the grammatical analysis of English nominal compounds. In Bierwisch, Manfred and Heidolph, Karl Erich, editors 1970, *Progress in Linguistics*. Mouton, The Hague. 174–186.

Leonard, Rosemary 1984. *The Interpretation of English Noun Sequences on the Computer*. North Holland, Amsterdam.

Levi, Judith N. 1978. *The Syntax and Semantics of Complex Nominals*. Academic Press, New York.

Magerman, David M. and Marcus, Mitchell P. 1990. Parsing a natural language using mutual information statistics. In *Proceedings of the Ninth National Conference on Artificial Intelligence*.

Marslen-Wilson, William and Tyler, Lorraine Komisarjevsky 1987. Against modularity. In Garfield, Jay L., editor 1987, *Modularity in Knowledge Representation and Natural-Language Understanding*. MIT Press, Cambridge, MA. 37–62.

McDonald, David B. 1982. Understanding noun compounds. Technical Report CMU-CS-82-102, Carnegie-Mellon Univ., Dept. of Comp. Sci., Pittsburgh, PA.

Pearl, Judea 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann, San Mateo, CA.

Plate, Tony 1991. Holographic reduced representations: Convolution algebra for compositional distributed representations. In *Proceedings of the Twelfth International Joint Conference on Artificial Intelligence*, Sydney, Australia.

Pollack, Jordan B. 1990. Recursive distributed representations. *Artificial Intelligence* 46:77–105.

Quillian, M. Ross 1969. The teachable language comprehender: A simulation program and theory of language. *Communications of the Association for Computing Machinery* 12(8):459–476.

Russell, Stuart J. and Wefald, Eric H. 1991. *Do the Right Thing: Studies in Limited Rationality.* MIT Press, Cambridge, MA.

Shieber, Stuart M. 1986. *An Introduction to Unification-Based Approaches to Grammar.* Center for the Study of Language and Information, Stanford, CA.

Smadja, Frank A. 1991. *Extracting Collocations From Text. An Application: Language Generation.* Ph.D. Dissertation, Columbia University, New York.

Smolensky, Paul 1990. Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence* 46:159–216.

Stolcke, Andreas and Wu, Dekai 1992. Tree matching with recursive distributed representations. In *AAAI-92 Workshop on Constraining Learning with Prior Knowledge*, San Jose, CA. 58–65. Also available as International Computer Science Institute report TR-92-025.

Talmy, Leonard 1983. Spatial orientation: Theory, research, and application. In Pick, Herbert and Acredolo, Linda, editors 1983, *How Language Structures Space.* Plenum Press, New York.

Talmy, Leonard 1985. Lexicalization patterns: Semantic structure in lexical forms. In Shopen, T., editor 1985, *Language Typology and Syntactic Description*, volume 3: Grammatical Categories and the Lexicon. Cambridge University Press, Cambridge.

Touretzky, David S. 1990. BoltzCONS: Dynamic symbol structures in a connectionist network. *Artificial Intelligence* 46:5–46.

Warren, Beatrice 1978. *Semantic Patterns of Noun-Noun Compounds.* Acta Universitatis Gothoburgensis, Gothenburg, Sweden.

Weatherford, Roy 1982. *Philosophical Foundations of Probability Theory.* Routledge & Kegan Paul, London.

Wermter, Stephan 1989. Integration of semantic and syntactic constraints for stuctural noun phrase disambiguation. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence.* 1486–1491.

Wilensky, Robert; Chin, David; Luria, Marc; Martin, James; Mayfield, James; and Wu, Dekai 1988. The Berkeley UNIX Consultant project. *Computational Linguistics* 14(4):35–84.

Wu, Dekai 1989. A probabilistic approach to marker propagation. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, Detroit, MI. Morgan Kaufmann. 574–580.

Wu, Dekai 1990. Probabilistic unification-based integration of syntactic and semantic preferences for nominal compounds. In *Proceedings of the Thirteenth International Conference on Computational Linguistics*, volume 2, Helsinki. 413–418.

Wu, Dekai 1991. A continuum of induction methods for learning probability distributions with generalization. In *Program of the Thirteenth Annual Conference of the Cognitive Science Society*, Chicago. Lawrence Erlbaum Associates.

Wu, Dekai 1992. *Automatic Inference: A Probabilistic Basis for Natural Language Interpretation.* Ph.D. Dissertation, University of California at Berkeley.