In *IJCAI-93, 13th International Joint Conference on Artificial Intelligence*, Chamberry, France: Aug. 93, 1290–1296.

Approximating Maximum-Entropy Ratings for Evidential Parsing and Semantic Interpretation

Dekai Wu*
Department of Computer Science
The Hong Kong University of Science & Technology
Clear Water Bay, Hong Kong
dekai@cs.ust.hk

Abstract

We consider the problem of assigning probabilistic ratings to hypotheses in a natural language interpretation system. To facilitate integrating syntactic, semantic, and conceptual constraints, we allow a fully compositional frame representation, which permits co-indexed syntactic constituents and/or semantic entities filling multiple roles. In addition the knowledge base contains probabilistic information encoded by marginal probabilities on frames. These probabilities are used to specify typicality of real-world scenarios on one hand, and conventionality of linguistic usage patterns on the other. Because the theoretical maximum-entropy solution is infeasible in the general case, we propose an approximate method. This method's strengths are (1) its ability to rate compositional structures, and (2) its flexibility with respect to the inputs chosen by the system it is embedded in. Arbitrary sets of hypotheses from the front-end processor can be accepted, as well as arbitrary subsets of constraints heuristically chosen from the long-term knowledge base.

1 Introduction

It has been generally accepted for some time that some linguistic constructs can only be properly disambiguated and interpreted by combining lexicosyntactic factors with semantic and conceptual knowledge. A prime example is the nominal compound construction, examples of which are *coast road*, *weekend guest*, *high-speed bus*, and *rubber baby buggy bumper*. To parse the latter correctly as [*rubber* [[*baby buggy*] *bumper*]] requires lexicosyntactic knowledge of the conventional collocation *baby buggy*, as well as conceptual knowledge that bumpers are typically constructed of rubber, but not buggies. Moreover, such constraints are not absolute, but soft and defeasible suggesting a numerical treatment (Wu 1990).

Probabilistic methods for language interpretation have recently come to the fore, spurred by developments in uncertain reasoning and statistical linguistics. Previous methods for combining these factors relied largely on *ad hoc* schemes requiring painstaking number tweaking, whereas to realize any hope of scalability, the calculus must be consistent in principle with some statistical sampling process. However, probabilistic methods are computationally expensive in domains with complex dependencies—like natural language—so efficient approximate techniques must be sought.

The following are the objectives behind the proposed method:

1. Integration of syntactic, semantic, and conceptual factors.
2. Consistent and statistically grounded evidence combination.
3. Flexible policy on hypothesis and constraint selection.

The latter is the most important feature distinguishing the proposed method from others. It means that the approximation technique does not impose strong restrictions on the probabilistic event space, with the following rationale. To keep computational expense reasonable, the rating procedure should not employ the full knowledge base, nor consider all possible interpretations. Instead, language analysis should employ various heuristics for hypothesizing good candidate interpretations and for retrieving pertinent constraints from the knowledge base. Thus the rating method should be capable of handling these arbitrary subpartitions of the hypothesis and constraint space.

The schema of this paper is as follows. The next section discusses related approaches. In section 3 we discuss why maximum-entropy ratings are difficult to obtain for natural language interpretation. Section 4 introduces an approximation technique to overcome this difficulty. Illustrative examples are given in section 5, followed by additional discussion in section 6.

2 Related Work

Related probabilistic interpretation methods fall into two main categories with respect to the type of concession made for computational feasibility. On one hand are methods that sacrifice the expressiveness of the language in which constraints

*Preparation of this paper was partially supported by the Natural Sciences and Engineering Research Council of Canada while the author was a postdoctoral fellow at the University of Toronto. 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 N00014-89-J-3205, the Sloan Foundation under grant 86-10-3, and the National Science Foundation under CDA-8722788.

are represented. On the other hand are methods that restrict the precedence ordering of dependencies between constraints.

Methods of the first category employ simple representation languages that do not permit expressing compositional constraints. Examples of this approach are n-gram models, hidden Markov models, and stochastic context-free grammars. These methods do not yield full conceptual representations, but can be used to rate alternative parses. For example, Fujisaki *et al.* (1991) rate potential bracketings of nominal compounds using a stochastic CFG. It is also possible to view microfeature-based PDP models that have probabilistic interpretations as members of this category.

Methods of the second category allow compositional constraints about roles and fillers to be predicated, but restrict the precedence of such constraints. The members of this group most relevant here (Goldman & Charniak 1990a, 1990b; Charniak & Goldman 1988, 1989; Burger & Connolly 1992) are based on Bayesian belief networks (Pearl 1988). Difficulties that arise with this group are discussed in a companion paper (Wu 1993a). Briefly, the precedence of dependencies is dictated by the topology of the net, according to the rule that any two nodes are conditionally independent given the instantiated values of all intervening nodes. Although nodes are permitted to express hypotheses about the binding of one variable to another, the state of a binding hypothesis eliminates many others. For example, if a role r is bound to some filler, no other hypothesis that binds r can hold. Under the conditional independence rule, either (1) all such interdependent hypotheses must be connected, creating loops which are expensive to evaluate, or (2) they must be clustered into single nodes as compound variables, with equivalent computational cost. The usual solution is to use *a priori* knowledge to generate as few binding hypotheses as possible, and to give some binding hypotheses precedence over others by appropriate choice of instance variables. However, as argued in the rest of this paper, the awkwardness of such a framework may be unnecessary.

3 Why the "Correct" Solution is Hard

Before proceeding to describe the proposed method, let us consider what it is about the natural language domain that makes getting the theoretically "correct" maximum-entropy solution different from other domains where statistical and uncertain reasoning techniques have proven so apt. (Maximum entropy (Jaynes 1979) is a principle for assigning probabilities to a partially constrained distribution without adding information.) First we ask what forms of information are needed to interpret nominal compounds, taking *coast road* as an example.[1] We conducted an informal survey that yielded as the most common interpretation a road in or along the coastal area. We also found that longtime residents of California automatically concretized this to Highway 1. Another interpretation that was volunteered, though less preferred, was a road amenable to coasting. In addition, all informants agreed that in a context like *Since the earthquake damaged the only Interstate to the coast, old Highway 17 will temporarily serve as the main coast road*, the interpretation of a road leading to the coast would be perfectly natural. To account for these preferences, some of the types of knowledge constraints needed are:

1. *Specific conceptual schemas*: The *Highway 1* concept is a specialization of *coastal road* that is often used when residing on the West Coast.[2]
2. *Intermediate conceptual schemas*: Usually when thinking about roads in the context of seacoasts, one thinks specifically of roads running along the coast (the subcategory we abbreviate as *coastal road*).
3. *Semantic schemas*: In our example we have a one-dimensional entity *road* and an elongated two-dimensional space *coast*. Prototypical abstract spatial relationships for this combination include *aligned containment* and *linear order locative* (i.e., destination).
4. *Specific constructions*:[3] By convention, the noun *coast* means a seacoast substantially more frequently than an unpowered movement.
5. *Intermediate constructions*: The construction *coast N* relates to a seacoast more often than unpowered movement, as in *coast artillery*, *coast guard*, *coastland*, *coastline*, and *coast redwood*. (In contrast, *coaster* tends to designate unpowered movement, as in *coaster wagon*, *coaster brake*, and *roller coaster*.)
6. *Abstract constructions*: Nominal compounds frequently express containment relationships, and somewhat less frequently, linear order locative relationships.

These factors involve a mix of feature constraints and role binding constraints and are conveniently notated using conventional frames or feature-structures, as shown in figures 1(a) and (b). (Details of our representation and ontology are given in Wu (1992, 1993b.) To see why role bindings make maximum-entropy estimation difficult, we may compare the structure of concept spaces with and without binding. In both cases, there is an implicit abstraction hierarchy governing the set of all tokens in the concept space. This can be made explicit by the device of *abstractive relations*. Without bindings, as in the simple binary feature example of figure 2(a), abstractive relations such as A and B indicate that the tokens 001 and 011 share a common abstract ancestral type $0x1$. (Note the distinction between the token 001 and its parent type 001.) Here any input evidence about a particular feature's value, say that the leftmost bit is 1, corresponds to conditioning by an internal node such as $1xx$. Marginal probabilities on internal nodes may or may not be known; for example we might have $P(0xx) = 0.6$ and $P(1xx) = 0.4$. The maximum-entropy distribution over the set of tokens $\{q_1, \ldots, q_C\}$ is computed by maximizing

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

[1] From the Brown corpus (Kučera & Francis 1967). For more discussion of nominal compounds, see Wu (1990) and the linguistics literature (e.g., Lees 1963, 1970; Downing 1977; Levi 1978; Warren 1978; McDonald 1982; Leonard 1984).

[2] We use *coastal road* to designate a conceptual category, as opposed to the novel lexical construct *"coast road"*.

[3] In the philosophy of Fillmore's (1988) Construction Grammar, a *construction* is a lexicosyntactic pattern paired with the semantic/conceptual representation it signifies (de Saussure 1966). Constructions come at various levels of specificity and an inheritance hierarchy is a convenient storage technique; see Wu (1992).

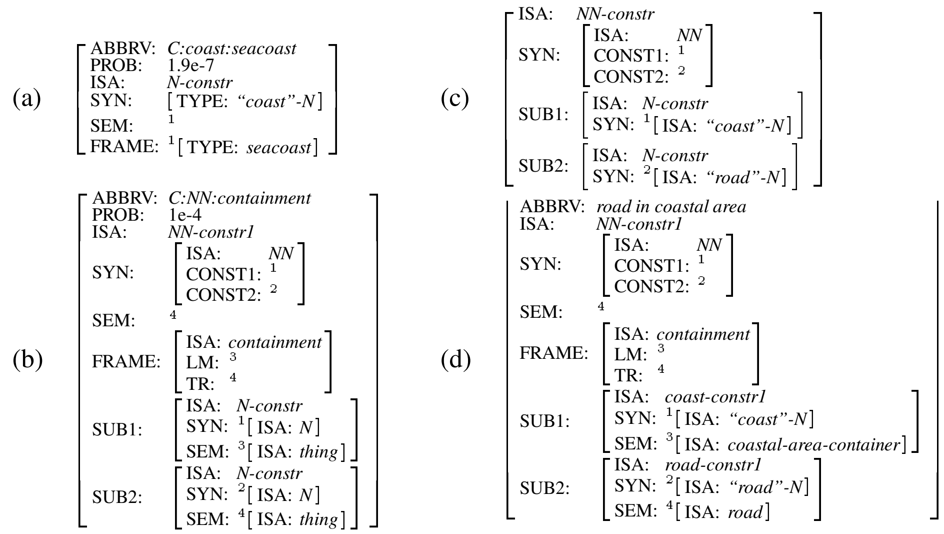


Figure 1: Feature-structures for (a) knowledge constraint 1 above, (b) knowledge constraint 6 above, (c) an input form, and (d) a full output structure (the floor brackets indicate a token as opposed to type).

subject to the marginal constraints, where $P_i \stackrel{\text{def}}{=} P(q_i)$. When the implicit abstraction hierarchy is a regular lattice such as this, it is easy to predicate conditional independence relationships between features that are deemed to be uncorrelated given the value of other chosen features, by omitting all information at a particular depth, say for $P(000), P(001), \ldots, P(111)$. This means certain marginal distributions are mutually orthogonal, and in this special case the maximum-entropy distribution can be computed simply by multiplying marginal probabilities. For example, if the only other things we know are $P(0x0) \neq P(1x0)$ and $P(00x) \neq P(10x)$, then the middle and rightmost features are conditionally independent given the value of the leftmost feature, so $P(000) = P(0x0)P(00x)$. In effect, efficient methods for probabilistic uncertain reasoning such as belief networks take advantage of this underlying assumption about the concept space.

Now consider the case with role bindings. In the abstraction hierarchy we must provide for specialization of the features that describe role types, in addition to specialization of concept features as in the no-bindings case. Moreover, we must allow for the commonality between feature-structures with two roles that are either bound or not, as performed in figure 2(b) by abstractive relations C and D. (For perspicuity feature-structures are depicted using directed graph notation.) The resulting abstraction hierarchy is a semilattice because any two feature-structure nodes have a unique least upper bound, but not necessarily a unique greatest lower bound. As before, the input evidence corresponds to conditioning on an internal type, having a form such as figure 1(c) which says the input phrase is *coast road*. Also, we may know marginal probabilities on some internal types, and the joint distribution is obtained by maximizing entropy. However, the distortion introduced into the abstraction lattice by the role bindings rules out orthogonal marginal distributions, except for extreme coincidences; there is no longer a single depth at which all information can be omitted to yield conditional independence. Thus we can no longer exploit conditional independence assumptions to make evaluation efficient. To the best of our knowledge, the application of maximum entropy in its general form (other than conditional independence) to this type of compositional "binding-distorted" semilattice has not previously been suggested.

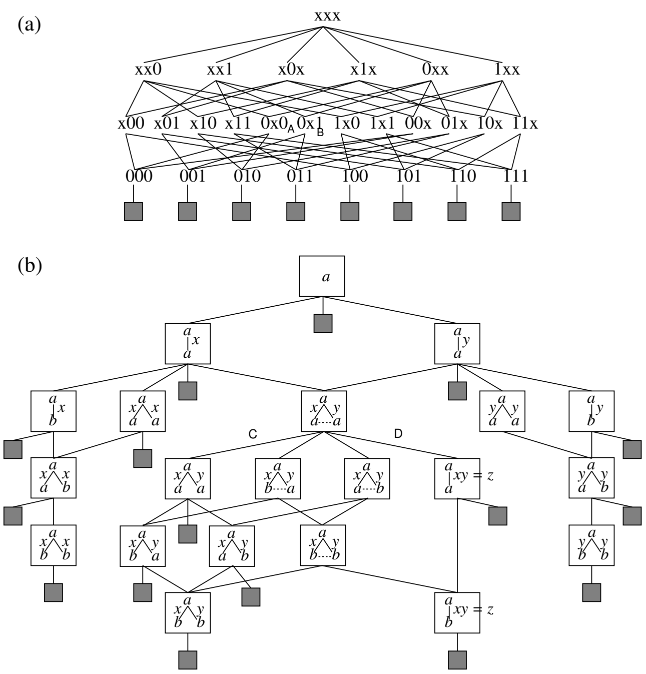


Figure 2: (a) Abstraction lattice for 3-bit feature vectors; (b) partial abstraction lattice for feature structures. Shaded boxes represent tokens with the same form as their immediate parent types. We assume b is a subtype of a, and z is a composite role equivalent to the conjuntion xy. A dashed line indicates an option to unify.

4 An Approximate Maximum-Entropy Technique

In section 2 we discussed two approaches that have been taken to feasible approximation. We propose a third approach, which is to relax the requirement of getting a strict conditional distribution. This proceeds from the assumption that language, though ambiguous, contains sufficiently many discriminating cues so that getting precise probabilistic ratings is computational overkill. Rather, it suffices to (1) generate hypothesized interpretations heuristically, (2) retrieve potentially relevant constraints from the knowledge base, and (3) compute an approximation to the conditional distribution (the first two steps can be done by a chart parser, marker passing, and/or other standard methods). The method presented here approaches the theoretical probabilistic solution as more hypotheses and constraints are given to it, but also becomes too expensive then. However, the method is quite practical given a reasonable number of heuristically generated hypotheses and constraints.

The essence of the method, called AME for Approximate Maximum Entropy, is to concentrate on rating only the hypotheses, and no other tokens; in contrast full-fledged maximum entropy assigns probabilities to all tokens. The approximation is accomplished by discarding the details of the abstraction lattice (or event space) outside the immediate hypothesis space. Instead, a minimum number of "dummy" tokens (or events) are introduced to cover the space, one for each marginal constraint. Each dummy event represents all the tokens that are consistent with the corresponding constraint, but not with the conditional space (the space subsumed by the input conditioning event type). In addition, a single *null* dummy event represents all remaining events. Dummy events are treated as tokens, even though each one takes the place of many real tokens. The standard maximum-entropy criterion is then applied to the new coarse space comprising the dummy tokens together with the hypothesis tokens. This problem is feasible since the space is much smaller; the only detailed structure lies within the hypothesis space. Discrepancies in the approximation arise from the fact that dummy events are treated as if they are disjoint, although in fact they represent event subspaces that may overlap.

More formally, suppose $\mathcal{Q}$ is the set of token f-structures and $\mathcal{G}$ is the set of 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 semilattice 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:

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

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

Then we wish to 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

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

which can be solved by an iterative method given in the companion paper (Wu 1993a).

This formulation has several advantages over a number of other probabilistic approaches including belief nets. First, it eliminates explicit features (or hypothesis nodes) for variable bindings. Instead, the similarity between feature-structures with similar binding patterns is expressed implicitly in the subsumption lattice. Thus the framework does not promote questionable precedence ordering judgements for binding hypotheses.

Second, the method robustly accepts arbitrary subsets of hypotheses and marginal probabilities, and does not require strictly consistent conditional probabilities. Aside from reducing the event space to a feasible size, dummy events serve a second purpose: they guarantee consistency of the marginal probabilities even when an "improper" subset of marginal constraints has been retrieved from the knowledge base. An example of this will be noted in the next section.

5 Examples

A better feel for the behavior of AME can be had from the following exercise. Figure 3(a) shows a simple example where the two dummy events are incorrectly related. Since *coast and coastal road* is subsumed by *road*, every token of the former type should be counted in the marginal probability for *road* as well, as shown in figure 3(b). The probabilities, shown both as joint and conditional (parenthesized) distributions, change substantially with proper subsumption.

It is also important to include the most specific applicable marginal constraints available from the knowledge base. The strong preference for *road leading to coastal area* in figure 3(b) arises because the hypothesis is left "floating" with no other marginal to constrain its probability. Another large correction occurs in figure 3(c) when a specific marginal on *coast and road* is included, which restricts the total amount of probability assignable to the hypotheses. This causes much of the probability weight to be shifted to the dummy events and switches the preferred hypotheses. Another simple example of this phenomenon is shown in figure 4 demonstrating the sensitivity of the *coasting road* hypothesis (a road amenable to coasting) to the presence of a constraining marginal on *coasting accomplishment* (coast in the unpowered movement sense).

Now consider again the range of knowledge needed to interpret *coast road*. We suppose that the subset of constraints

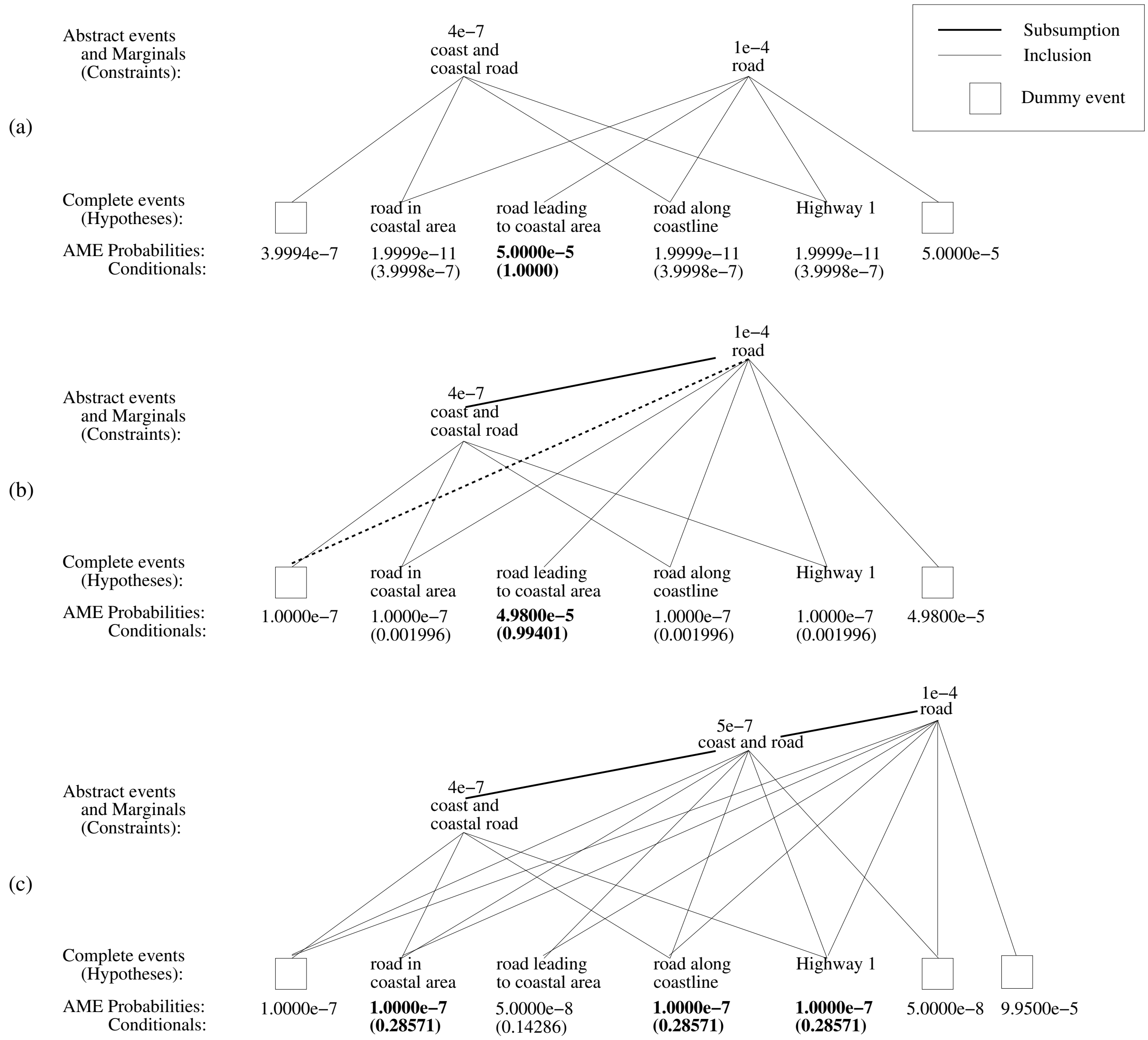


Figure 3: A comparison of dummy events (a) with improper subsumption, (b) with proper subsumption, (c) with specific constraint. The null event is not shown. Each internal node stands for a feature-structure *type* as in figures 1(a) and (b). Each leaf node stands for a hypothesis *token* as in figure 1(d).

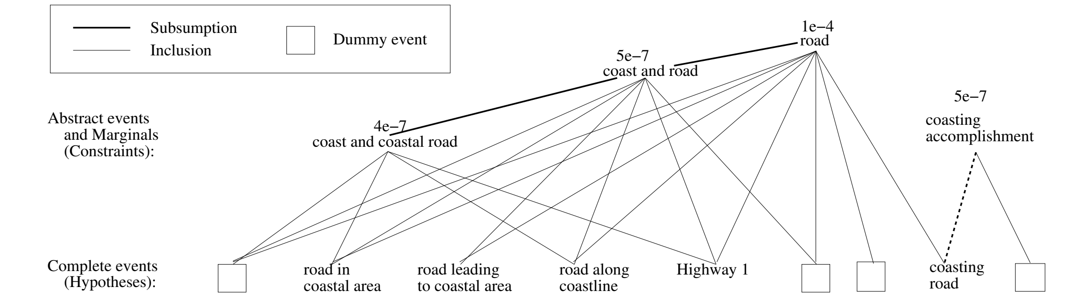
and hypotheses deemed pertinent are as shown in figure 5. The constraints correspond to the types of knowledge constraints enumerated in section 3. Knowledge constraint 1 is encoded by the relative magnitudes of $P($*coastal road*$)$ and $P($*Highway 1*$)$. Similarly, constraint 2 is encoded by $P($*coast and road*$)$ and $P($*coast and coastal road*$)$; constraint 4 is encoded by $P($*L:coast*$)$, $P($*C:coast:seacoast*$)$, and $P($*C:coast:coasting accomplishment*$)$; and constraint 6 is encoded by $P($*C:NN:linear order locative*$)$ and $P($*C:NN:containment*$)$, which says that a noun-noun construction is used to signify containment twice as often as linear order locatives. Constraints 3 and 5, as well as other intermediate lexicosyntactic signification patterns, are omitted for the sake of brevity.

In the basic case, AME yields the conditional distribution shown in the uppermost row labelled "**0:**". The hypotheses *road in coastal area* and *road along coastline* are the winners with the highest probability, as might be expected. (The probabilities for these two hypotheses will always be the same here as we have omitted their distinguishing features.)

The next run is labelled "**1:**". Suppose constraint 4 weren't true, and *"coast"* were actually used more often to mean unpowered movement than a seacoast. To model this, the marginals on *"coast"* signifying *seacoast* and *coasting accomplishment* are switched as shown by the alternate marginals labelled "**1:**". In the resulting conditional distribution, *coasting road* now dominates the hypotheses.

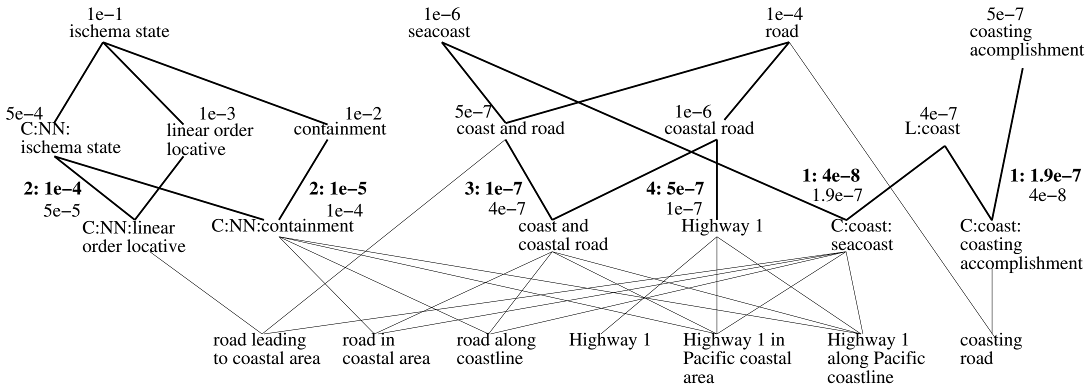
The marginals labelled "**2:**" model another alternate case where constraint 6 is reversed such that noun-noun constructions signify linear order locatives more often than containment. In this case the preferred interpretation switches to *road leading to coastal area*.

In "**3:**" we relax constraint 2 so that one does not think specifically of roads running along the seacoast quite as fre-



		(dummy)	road in coastal area	road leading to coastal area	road along coastline	Highway 1	(dummy)	(dummy)	coasting road	(dummy)
(a)	AME Probabilities:	1.0000e−7	1.0000e−7	5.0000e−8	1.0000e−7	1.0000e−7	5.0000e−8	9.9500e−5	4.9995e−7	
	Conditionals:		(2.0002e−7)	(1.0001e−7)	(2.0002e−7)	(2.0002e−7)			(1.0000)	
(b)	AME Probabilities:	1.0000e−7	1.0000e−7	5.0000e−8	1.0000e−7	1.0000e−7	5.0000e−8	9.9500e−5	2.5000e−7	2.5000e−7
	Conditionals:		(0.16667)	(0.083333)	(0.16667)	(0.16667)			(0.41667)	

Figure 4: Results with an added *coasting road* hypothesis (a) without and (b) with specific constraint on *coasting accomplishment*.



	road leading to coastal area	road in coastal area	road along coastline	Highway 1	Highway 1 in Pacific coastal area	Highway 1 along Pacific coastline	coasting road
0:	0.046524	**0.37215**	**0.37215**	0.00025822	0.074419	0.074419	0.060089
1:	0.015757	0.12605	0.12605	0.00061625	0.02521	0.02521	**0.6811**
2:	**0.38339**	0.15336	0.15336	0.0010636	0.030672	0.030672	0.24748
3:	**0.40849**	0.20422	0.20422	0.00056666	0.025527	0.025527	0.13145
4:	0.010205	0.081579	0.081579	0.00028371	**0.40657**	**0.40657**	0.01321

Figure 5: Base run for *coast road* plus four variations. The redundant inclusion arcs and dummy events are omitted and only conditional probabilities are shown.

quently (when thinking about roads in the context of seacoasts). Again, the result is that the preferred interpretation switches to *road leading to coastal area*.

Lastly, in "**4:**" we model constraint 1 for West Coast residents by increasing the magnitude of the $P(\textit{Highway 1})$ marginal, resulting in a corresponding change in the dominating hypotheses. All examples were computed using a C implementation of AME.

The preceding section mentioned that dummy events also serve to guarantee consistency of marginal probabilities. When only a subpartition of the knowledge base is considered, the marginals may be inconsistent due to the omission of some of the probability weight. For instance, in figure 5,

$$P(\textit{coast and coastal road}) + P(\textit{Highway 1}) < P(\textit{coastal road})$$

when they should be equal. The dummy events provide nonzero subspaces to absorb the remaining probability for *coastal road* that goes to other subcategories that have been neglected in the approximation.

6 Discussion

In practice it will be difficult to obtain the marginal probabilities on many frames. However, note that our guiding principle is that the method accepts whatever statistics *are* available and improves upon the interpreter's discrimatory power to the extent possible. Pragmatic, if theoretically unsatisfactory, probability estimation methods may well suffice for interim applications. Lexical frequency counts over the Brown corpus and others are available (Francis & Kučera 1982) and parsed corpora will soon facilitate frequency counts for syntactic patterns as well. These counts may be used as estimators for the marginals of lexicosyntactic structures. The analogous procedure is impractical for semantic or conceptual structures, since fully interpreted corpora are not available. Warren (1978) gives frequency counts on manually-analyzed, coarse semantic relation categories in her study of nominal compounds, but these must be massaged to fit more realistic 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 that could be used as a front-end hypothesis generator. 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 to a large degree on which frames the investigator chooses to place marginal constraints on; this model is as susceptible to the ontology problem as any other.

In effect, the investigator decides the degree of *generalization*, because what maximum entropy does is to generalize the partial distributional constraints in the knowledge base to the rest of the event space. Choosing the frames is a kind of concept formation, which this work does not aim to address, although toward this direction Wu (1991) proposes a discrete kernel estimation model of generalization from samples. Choosing a set of marginal constraints can then be seen as a matter of best fit to the theoretical kernel distribution.

Finally, it is worth noting that although we developed AME specifically to address our language interpretation problem, we believe the method is of sufficient generality as to find applications in other uncertain domains where the constraints are highly interconnected.

7 Conclusion

The high degree of interconnectedness between lexicosyntactic, semantic, and conceptual factors in natural language interpretation suggests quantitative integration of constraints. Evaluating hypotheses on a probabilistic basis rather than an *ad hoc* weighting scheme yields a model that will more plausibly scale up to interestingly large domains. The proposed technique has several advantages compared with other probabilistic methods. First, it evaluates structured hypotheses subject to compositional constraints, without additional binding precedence assumptions. It is also robust and highly flexible in that arbitrary heuristically-generated subsets of hypotheses and constraints are accepted. These two properties make the method straightforward to incorporate as a filtering/rating stage within existing theories and architectures.

Acknowledgements

I am indebted to Robert Wilensky, Jerome Feldman, and the members of the BAIR and L_0 groups for many valuable discussions, as well as Graeme Hirst, Geoff Hinton, and their respective groups.

References

BURGER, JOHN D. & DENNIS CONNOLLY. 1992. Probabilistic resolution of anaphoric reference. In *AAAI Fall Symposium on Probabilistic NLP*, Cambridge, MA. Proceedings to appear as AAAI technical report.

CHARNIAK, EUGENE & ROBERT GOLDMAN. 1988. A logic for semantic interpretation. In *Proceedings of the 26th Annual Conference of the Association for Computational Linguistics*, 87–94.

CHARNIAK, EUGENE & ROBERT GOLDMAN. 1989. A semantics for probabilistic quantifier-free first-order languages, with particular application to story understanding. In *Proceedings of IJCAI-89, Eleventh International Joint Conference on Artificial Intelligence*, 1074–1079.

DE SAUSSURE, FERDINAND. 1966. *Course in general linguistics*. New York: McGraw-Hill. Translated by Wade Baskin from the French edition of 1915. Reprinted from the 1959 edition (New York: The Philosophical Library).

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

FILLMORE, CHARLES J., 1988. On grammatical constructions. Unpublished draft, University of California at Berkeley.

FRANCIS, W. NELSON & HENRY KUČERA. 1982. *Frequency analysis of English usage: Lexicon and grammar*. Boston: Houghton Mifflin. With the assistance of Andrew W. Mackie.

FUJISAKI, T., F. JELINEK, J. COCKE, E. BLACK, & T. NISHINO. 1991. A probabilistic parsing method for sentence disambiguation. In *Current issues in parsing technology*, ed. by Masaru Tomita, 139–152. Boston: Kluwer.

GOLDMAN, ROBERT P. & EUGENE CHARNIAK. 1990a. Incremental construction of probabilistic models for language abduction: Work in progress. In *Working Notes from the Spring Symposium on Automated Abduction*, 1–4, Stanford University, Stanford, CA. AAAI.

GOLDMAN, ROBERT P. & EUGENE CHARNIAK. 1990b. A probabilistic approach to text understanding. Technical Report CS-90-13, Brown Univ., Providence, RI.

HEARST, MARTI A. 1991. Noun homograph disambiguation using local context in large text corpora. In *Seventh Annual Conference of the University of Waterloo Centre for the New OED and Text Research: Using Corpora*, 1–22, Oxford.

JAYNES, E. T. 1979. Where do we stand on maximum entropy. In *The maximum entropy formalism*, ed. by R. D. Levine & M. Tribus. Cambridge, MA: MIT Press.

KUČERA, HENRY & W. NELSON FRANCIS. 1967. *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

LEES, ROBERT B. 1963. *The grammar of English nominalizations*. The Hague: Mouton.

LEES, ROBERT B. 1970. Problems in the grammatical analysis of English nominal compounds. In *Progress in linguistics*, ed. by Manfred Bierwisch & Karl Erich Heidolph, 174–186. The Hague: Mouton.

LEONARD, ROSEMARY. 1984. *The interpretation of English noun sequences on the computer*. Amsterdam: North Holland.

LEVI, JUDITH N. 1978. *The syntax and semantics of complex nominals*. New York: Academic Press.

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*. San Mateo, CA: Morgan Kaufmann.

WARREN, BEATRICE. 1978. *Semantic patterns of noun-noun compounds*. Gothenburg, Sweden: Acta Universitatis Gothoburgensis.

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, 413–418, Helsinki.

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*, 949–953, Chicago. Lawrence Erlbaum Associates.

WU, DEKAI, 1992. *Automatic inference: A probabilistic basis for natural language interpretation*. University of California at Berkeley dissertation. Available as UC Berkeley Computer Science Division Technical Report UCB/CSD 92/692.

WU, DEKAI. 1993a. Estimating probability distributions over hypotheses with variable unification. In *Proceedings of AAAI-93, Eleventh National Conference on Artificial Intelligence*, Washington, D.C. To appear.

WU, DEKAI. 1993b. An image-schematic system of thematic roles. In *Proceedings of PACLING-93, First Conference of the Pacific Association for Computational Linguistics*, Vancouver. To appear.