A Semantics for Probabilistic Quantifier-Free First-Order Languages, with Particular Application to Story Understanding

Eugene Charniak and Robert Goldman*
Dept. of Computer Science, Brown University
Box 1910,
Providence, RI 02912

Abstract

We present a semantics for interpreting probabilistic statements expressed in a first-order quantifier-free language. We show how this semantics places constraints on the probabilities which can be associated with such statements. We then consider its use in the area of story understanding. We show that for at least simple models of stories (equivalent to the script/plan models) there arc ways to specify reasonably good probabilities. Lastly, we show that while the semantics dictates seemingly implausibly low prior probabilities for equality statements, once they are conditioned by an assumption of spatio-temporal locality of observation the probabilities become "reasonable."

1 Introduction

In this paper we present a semantics for quantifier-free first-order formulas as used in probabilistic statements. Quite often when using probabilities one wants to talk about the probability of a proposition being true. In texts, such as Pearl's [1988], the propositions are taken to be random variables, i.e. functions whose values are either 1 or 0, taken to mean true and false. It is assumed no further interpretation is needed. Nilsson's "probabilistic logic" [1986] gives more interpretation, but it is restricted to propositional logic, and as we will see, does not solve the problems which motivated this paper. The rest of this introduction will lay out what these problems are.

We arc interested in the use of probability theory to help with problems in natural language understanding (NLU) [Charniak and Goldman, 1988, Goldman and Charniak, 1988]. For the purposes of this paper, we simplify the problem by considering only written, expository text describing events and objects in the real world. Modal verbs, such as "want" or "will" are not allowed.[1] This allows us to view the language user as a transducer.

*This work has been supported in part by the National Science Foundation under grants 1ST 8416034 and 1ST 8515005 and Office of Naval Research under grant N00014-79-C-0529.

[1] Because of the limitations of our semantics, we also exclude statements about groups of objects.

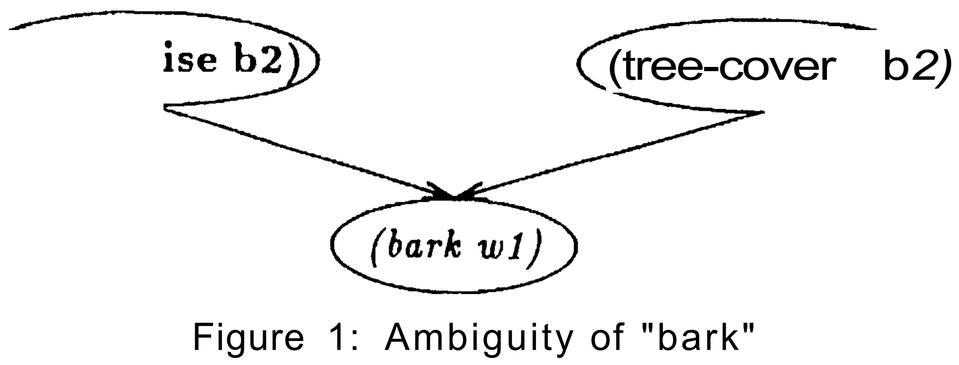

Figure 1: Ambiguity of "bark"

The language user observes some thing (event or object), and translates this thing into language. Our task is to reason from the language to the intentions of the language user and thence to the thing described.

This paper will not attempt to justify the use of probability theory for this task other than to say that language comprehension is naturally viewed as an "abduction" problem [Hobbs *et a*/., 1988], and probabilities seem a good way to represent the uncertainty which arises in abduction.

To take a particular NLU problem, consider word-sense disambiguation, but now in a probabilistic light. Figure 1 shows a Bayesian network designed to capture (part of) the situation when an author uses an ambiguous word like "bark." We have chosen to display our probability distributions using Bayesian networks because of the convenient way they summarize dependencies. However, nothing in our semantics depends on this choice. The nodes in a Bayesian network correspond to random variables, and the arcs indicate *direct* probabilistic influences. We have adopted here a convention, to be used throughout the paper, of using bold face for entities in the world, and predicates on them, and italics for words and predicates on them. So (dog-noise b2) says that the entity b2 is the noise that a dog makes, while *(bark wl)* says that *wl* is a token of the English word "bark." Also, we consistently use a "lisp-like" syntax for logical expressions.

In this diagram, *wl* is an instance of the word "bark," and b2 is a token representing the denotation of *wl*. Looking at, the top-leftmost node, its connection to the bottom node is designed to capture an influence on the decision to use *wl*, a token of the word "bark." In this case the influence is that an author is likely to use the

word "bark" if the object she wishes to refer to is of the type dog-noise. If the entity she wanted to talk about were a radish, she obviously would not have used the word "bark."

Ideally we would have here a probabilistic description of word-choice in English, but the nice thing about probabilistic models is that even very incomplete models can do some good, and here we have reduced the problem of word choice to that of matching the kind of object to the words. At any rate, given this formalism we calculate the probability that the word means, say, dog-noise using Bayes' Theorem.

$$\mathrm{P}((\textbf{dog-noise b2}) \mid (\textit{bark w1})) = \mathrm{P}((\textbf{dog-noise b2}))\ \frac{P((\textit{bark w1})\mid(\textbf{dog-noise b2}))}{P((\textit{bark w1}))}$$

A number like the probability of using the word "bark" given that the entity is the corresponding noise is easy to estimate. It is certainly high, say .9. But the other probabilities required here are harder. What, for example, is the prior of (dog-noise b2)? Typically we think of terms in our language as adhering to entities in the world, like a sound emanating from my backyard last night. If so, then since I thought that it was most likely a dog barking, I might say that the probability is .7. On the other hand, b2 is an arbitrary symbol, created by my language-comprehension system to denote whatever the writer was referring to. What is the probability that an "arbitrary symbol" denotes a bark? This must be astronomically small, assuming we can understand the notion at all. Or again, given that b2 is arbitrary, perhaps we should interpret the formula (dog-noise b2) as the skolemized version of the formula exists(x)(dog-noise x). Interpreted in this light the probability is 1, since, of course, barking sounds do exist.

As we have seen here, the problem is not really assigning the probability, per se, but rather deciding what the formula (dog-noise b2) *means.* Nor is (dog-noise b2) the only kind of formula we will have problems with. In a sentence like "Janet killed the boy with some poison." there is case ambiguity in that the word "with" can indicate that the "poison" is in the instrumental case instr, or the accompaniment case ace. That is, did Janet use the poison, or just take it along for the ride (as in "Janet went to the movies with Bill.")? Here we need the prior probability of a formula like (instr k1) = p1. All the same problems arise, and more.

2 The Model

2.1 The language

The syntax of our language is a restriction of the language of first-order predicate calculus. We have the customary constants, functions, predicates and connectives. However, we do not allow quantifiers or variables.

Having said this, it is important to emphasize that we are *not* providing a logical calculus. Our calculus is probability theory.

2.2 Semantics

1. We define two disjoint sets of primitive events, and probability distributions over them:
   (a) The set of all words, *W.*
   (b) The set of all individual things, *T.* 'Thing' is defined as per the isa hierarchy: events, objects, persons, concepts, etc.
2. We define the overall sample space, $\Omega = \{W \cup T\}$.
3. Constants represent the outcomes of trials. For example, in Figure 1, *w1* and b2 are random variables with values from *W* and T, respectively.
4. Functions of arity *n* are functions $\Omega^n \rightarrow \Omega$. For example, rope-of is a function which maps hanging events to the ropes used in them (if there is one) and is arbitrary otherwise. ("Arbitrary" here simply means that there is no correlation between the type of the argument of the function and the type of its value. If c4 is a potato, its rope might be anything from a graduation ceremony to a light bulb.)
5. A predicate P^n is a function $\Omega^n \rightarrow \{0, 1\}$. The predication (dog-noise b2) in Figure 1 (where b2 is the referent of *w1*) denotes the proposition that the word "bark" denotes a dog-noise.
6. The Boolean operators, or, and, and not allow us to compose predicates, in the customary way. Formally boolean operators are functions from predicates, or pairs of predicates, to predicates.

2.3 Where do the numbers come from?

For our simple axiomatization of the domain of story comprehension, we need a specific set of probabilities. Our axiomatization uses a conventional frame-based knowledge representation language with an isa-hierarchy. Slots, sub-acts, and roles in frames are represented by functions (e.g., the patient of a get action, g1, is represented as (patient g1)). The relations of interest between entities are represented by equality statements. For example, in order to represent that a going action (g2) is part of a plan to go shopping at a supermarket (plan1), we write (go-step plan1) = g2. The only predications we require specify the types of objects (e.g., (rope r1)), or syntactic relations between words (e.g., *(object w1 w2)).* Entities in *T* are described by using words in *W* which denote an object of the correct type. We assume that each open class word in *W* describes an entity in *T* as in Figure 1, where *w1* describes b2.

2.3.1 Priors

We require prior probabilities for propositions that an entity is of a given type. In this section we give a principled way of getting these priors from our isa-hierarchy and show that this method will provide consistent probabilities.

Our isa-hierarchy can be taken as a Bayesian network. Pearl [1988] shows that it is possible to assign a consistent probability distribution to any Bayesian network. We can direct the edges either from the leaves of the isa-hierarchy, or from the root. Given such a network, we can efficiently compute a prior for any type.

Note that each entity does not have to be equiprobable. That would imply that raw frequency in the world is the only factor in deciding whether something will be discussed. This is obviously not true. For example, one

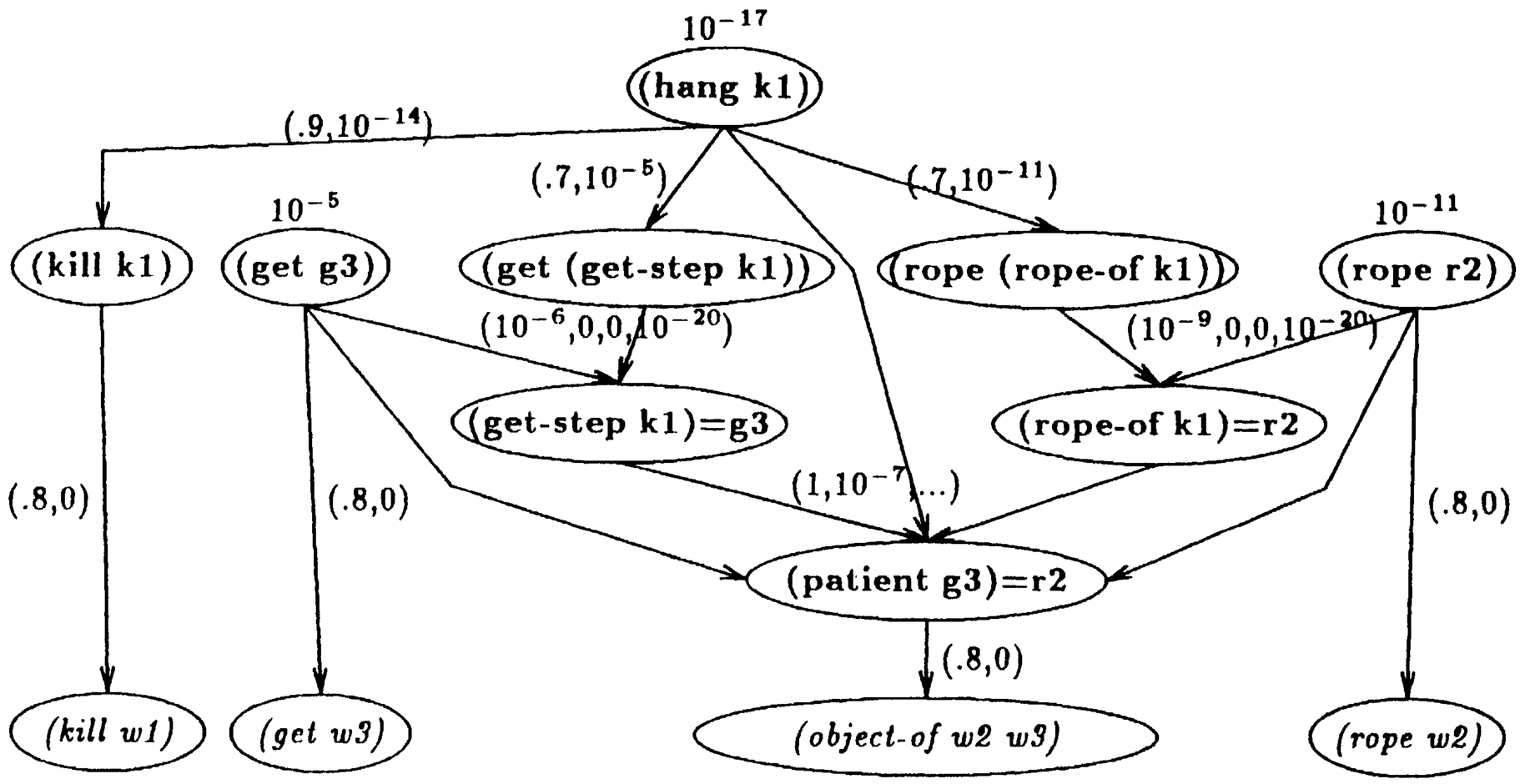

Figure 2: The Bayesian network for "Jack got a rope. He killed himself."

could imagine a universe of discourse, U, containing a large number of inanimate objects, but only 2 people, Jack and Jill. Certainly the probability of a person being discussed will be higher than $\frac{2}{|U|}$. Furthermore, if we wish to discuss domains with infinite subclasses (e.g., the integers), as well as finite ones (U.S. Presidents), to assume a flat distribution over the sample space would entail that the probability of talking about George Bush would be zero. Although weighting the isa-hierarchy for 'interestingness' or other qualities, poses no theoretical difficulty, it does make computation of the probability of equality statements more difficult.

2.3.2 Conditional probabilities

For story understanding in our formalization, we need conditional probabilities of three sorts. We need the probability of an equality relation between two objects of a given type. We need the probability of a particular word choice, given that thing it denotes is of a given type. Finally, we need the probability of a syntactic relation existing between two words, given that a particular relation exists between the things denoted by the two words.

- Probabilities of equality statements: We require the probability of equality statements, *x* — *y*, conditional upon *x* and *y* being objects of the same type, *t.* Given that we know the size of T, and a prior for *t(x),* and assuming *T* is finite and all of its elements equiprobable, *P(x* = *y*/t.(x), *t(y))* is seen to be

$$\frac{1}{\mathrm{P}(t(z)) \times |T|}$$

  P(word|denotation): In principle, these probabilities could be computed from the lexicon, and could absorb information about relative frequency of different words.

- Syntactic relations: We have to provide the probability of syntactic relations, given some relationship between the entities denoted by two words. For example, in Figure 2, we need *V((object-of w2 w3)* | *(patient g3)* = *r2),* where g3 is the action referred to by *w3*, an instance of the verb 'went'; and r2 is the entity referred to by , *w2,* an instance of the noun 'rope.' I.e., we need the probability of an author expressing a patient relation between two entities by means of the direct-object relation between the words which denote these entities.[2]

3 Getting the model to produce reasonable results

We have defined our model and shown that it gives us some guidance in assigning probabilities we need. In this section we show that the guidance is not sufficient. We give an example which shows that the model we have outlined so far does not see stories as coherent wholes. We show that this is because the model does not take into account enough conditioning information, and show how to fix it.

Let us consider the example "Jack got a rope. He killed himself." Intuitively we would say that the probability that he did it by hanging was quite high. Certainly greater than .1.

Actually there is a complication in these last two probabilities. Really the conditional probability is the product of the probabilities we have outlined above *times the probability that a particular object (or relation) would be realized in the sentence at all.* When this is factored in we get a very low number, but it will be large relative to the probability that the author would use the word (or syntactic relation) given that the proper facts in the world did not exist. This is all we need.

Figure 2 depicts a fragment of a Bayesian network for understanding this sentence. The nodes at the bottom of the diagram correspond to the words in the sentence and the relations between them. The nodes above represent information about the referents of these words, and their interrelations.

The numbers above root nodes are the prior probabilities of these nodes. For example, the 10^{-11} above (rope r2) indicates the prior probability of an arbitrary entity being a rope. The numbers on the arc connecting a child c with one parent *p* to that parent are the P(c | p),P(c | -p). Examining the link between (rope r2) and *(rope w2),* we see that the probability of using the word 'rope', given that the referent of that word is a rope, is 0.8. Links connecting a child to two parents are annotated with the four conditional probabilities, starting with P(c |true, true). For example, the probability of r2 filling the rope-of slot of k1, given that both the (rope-of kl) and r2 are ropes, is 10^{-9}, if one is a rope and one is not, the probability is 0, and if neither is a rope, the probability is $\frac{1}{|T|}$. For the sake of readability, we have simply given the probability of (patient g3) = r2 for the cases when all five of its parents are true, and when (get g3) and (rope r2) are true.

Put in words, the network expresses the facts that a hanging is also a killing, and that the slots in the hanging frame, rope-of and get-step, must be filled by ropes and getting events respectively, with the latter being the event in which (rope-of kl) is obtained. The equality statements express the idea that the "get" and "rope" mentioned are, in fact, the ones which fill the appropriate slots. It follows from these facts that r2 must be the thing fetched in g3. This is captured in the connections to the node (patient g3) = r2. Since the network as given does not contain the information about Jack being both the person who does the killing, and who obtains the rope, the probability on the equality of get-step and g3 has been modified to reflect this information. The numbers used are calculated as discussed earlier, with $|T| = 10^{20}$, 10^9 ropes, 10^9 people, 10^{15} get events, 10^3 hangings, and about 10^6 killings.

When one evaluates this network[3] one gets a probability of hang very close to 10^{-3}. That is, the information about getting a rope has made no difference, since the probability of hang given kill is about 10^{-3}. The problem turns out to be the probabilities assigned to the equality statements.

To see this, it is necessary to get some feel for the flow of probabilities in this network (rope r2), (get g3), and (kill kl) have probability one, (hang kl) is about $10\sim^3$, but the equality nodes for get-step and rope-of have very low probabilities because their posteriors given type equality are $\frac{1}{|ropes|}$ and $\frac{1}{|gets|}$ which are very small indeed. Changing our belief in (patient g3) = r2 to 1 raises the probability of hanging, but not by much. The reason is that the current belief in get-step and rope-of equalities are *so* low that it is more likely that the rope was the patient of get "simply by accident," rather than seeing it as a consequence of the get-step and rope-of equalities. Thus the low probabilities on these two equality statements are the reason for this counter-intuitive result.

The solution is clear from an analysis of why the rope-of and get-step posterior probabilities are so low. Looking at the first one, it is asking the question, picking an arbitrary hanging event, and an arbitrary rope, what is the probability that the rope so chosen will be the one used for the hanging event? Obviously it is quite low, since we might be picking a rope which exists thousands of miles away from the hanging, or which existed a few hundred years ago (if we are including historical objects and events in Ω.) So *given this meaning for the number,* the number we gave is in the right ballpark. But there is other information which we did not bring to bear. Obviously in a story like "Jack got a rope. He killed himself." it is simply not likely that the rope and the killing are separated by thousands of miles, or hundreds of years. In other words, stories, and for that matter, observations in the world, are typically constrained by temporal and spatial locality. Our semantics, with random experiments over Ω, has no such constraint. We need to include it.

There are two ways to go. One would be to change the semantics to include some recognition of spatial and temporal locality of objects and events. The other is to keep the same semantics and simply include the assumption of spatio-temporal locality as a conditioning event. We will do the latter. Later we will see why changing the semantics is probably a bad idea.

Figure 3 shows the rope-of equality statement from Figure 2, but now with it conditioned on spatio-temporal locality.[4] Such an stl predication would also appear as a condition on the get-step equality statement. Figure 3 gives the probability for rope-of being a particular rope, assuming locality. Note the difference in probabilities. Before it was 10^{-9}, since only one of the 10^9 ropes would be the right one. Now, however, with spatio-temporal locality, we are asking a very different question. Given that a hanging, and a rope, are found in a small part of space-time, what is the probability that the rope was used in the hanging? Obviously this depends on the size of the "part of space-time," and a more sophisticated analysis should make stl take this into account. However, for current purposes, suppose we envision it as a city block. Now, rather than 10^9 ropes we are talking about 10, or perhaps 100. Figure 3 adopts 100, so the probability that any one of these is the rope in question is 10^{2}. Similarly, the get-step equality changes from 10^{-6} to "0^{-2}. Naturally, the probability of hang now goes up precipitously, to .3.

Now if stl statements were always true our analysis would be quite simple. But they are not. In stories neighboring sentences, or even different clauses in the same sentence, can be about different parts of space and time. Normal perception, it is true, sticks to local things, but if we watch TV we can get widely dispersed images as well. Thus we need some theory of under what cir-

[3]Actually we evaluated a singly-connected version using the algorithm of Pearl and Kim. [Pearl, 1988]

[4]The statement (stl kl r2) indicates that kl and r2 are in the same locality.

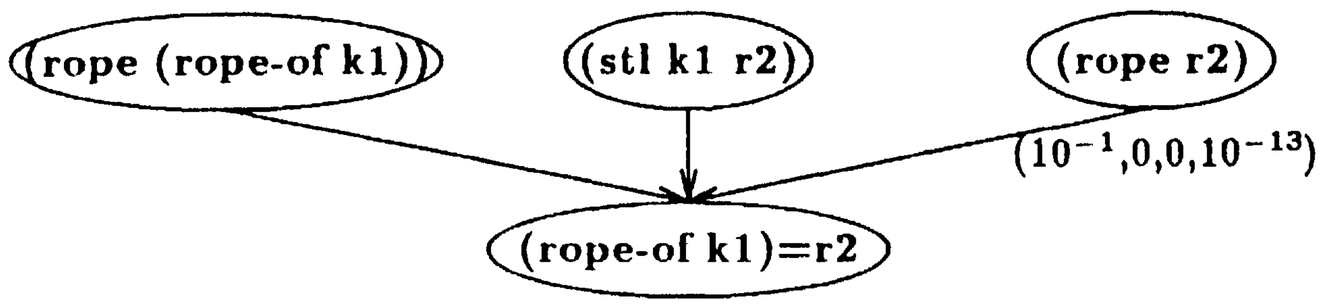

Figure 3: Network with spatio-temporal locality

cumstances stl statements are true. Building stl into the semantics would, presumably, mean building this theory into it as well, and since this promises to be a substantive theory, it seems a bad idea to include it in the semantical definitions.

Interestingly enough, some parts of an stl theory are already in place, albeit not under this name. Within the AI/Natural Language Understanding community there is a sizable body of work on "discourse structure." [Grosz and Sidner, 1986, Webber, 1987] To take a typical example (this one from, [Allen, 1987] which provides a good overview)

1 Jack and Sue went to a hardware store to buy a new lawnmower
2 since their old one had been stolen.
3 Sue had seen the men who took it
4 and had chased them down the street,
5 but they'd driven away in a truck.
6 After looking in the store, they realilzed that they couldn't afford a new one

Note that lines 1 and 6 are about the same part of space-time, as are lines 2-5, but there is no commonality between the two. Discourse structure theorists would say that there are two discourse segments in this example, a major segment consisting of 1 and 6, and a sub-segment consisting of 2-5. It is obvious in this example that this analysis, and the one necessary to determine spatio-temporal locality, exactly overlap. Furthermore, when one looks at the clues that discourse theorists suggest for determining this structure, (change of time, no referents for pronouns, certain key phrases, such as "by the way") it is easy to see how they would be equally useful in determining the truth of a spatio-temporal locality hypothesis. We have done some preliminary work on exploring more discourse-oriented conditioning events than simple spatio-temporal locality, see [Charniak and Goldman, 1989].

4 Probabilistic logics

A reasonable question to ask is why we don't use a probabilistic logic like that of Nilsson [l98f>] or Bundy [1988]. There are two reasons.

First, Nilsson's logic (Bundy's approach is similar) allows the user to specify the probabilities of statements and uses this information to bound a sample space of 'possible worlds' or truth-assignments. It is not possible to use conditional probabilities in this framework, since they specify ratios of probabilities of statements, rather than probabilities of statements in isolation. This is appropriate for Nilsson's problem which is to explain probabilistic entailment, but we are more concerned with conventional probabilistic inference, than with probabilistic analogs to material implication.

Second, Nilsson's Probabilistic Logic is only a propositional language. However, as the formulas of our language contain neither quantifiers nor variables, we could try to deal with them as propositional constants. If we do this, however, we sacrifice information about the relation between propositions. For example, suppose that we have a probability distribution such that (rope r1) is .5, and in the probabilistic logic we assign probabilities to the possible worlds to make the probability of (rope r1) .5 as well. Now let us ask, what is the probability of, say, (rope r19). In the absence of any other information, for us this must be .5 as well. Probabilistic logic makes no such commitment. It is a propositional logic, so the probability of the two propositions (rope r1) and (rope r19) can be varied independently. Thus, our semantics is more restrictive. Furthermore, this extra restriction has the effect of placing tight bounds on the probabilities we can assign to basic type predications like (rope r1). Since it will specify that any outcome of a basic experiment is a rope 1/2 the time, a probability of .5 commits us to the belief that half the entities we deal with will be ropes; not a very plausible assumption. Bundy's Incidence Calculus is a full first-order language, but he does not provide any guidance in interpreting the constants of the language, so the Incidence Calculus is no more restrictive than Probabilistic Logic.

After completing the work described in this paper, we discovered the work of Bacchus [1988], which is in many ways similar. His logic, Lp, is similar to our language in that the distributions he uses are over the domain of discourse, rather than over interpretations. Like our language, Lp makes it possible to refer to random variables. His language differs from ours in two ways. First of all, it is intended to support theorem-proving, whereas ours is designed to give a clear semantics to statements about random variables. Second, Bacchus' language is designed to support reasoning about probabilistic judgments. Statements about the probability of given events can be expressed in Lp, whereas in our language they are metalogical.

5 Future work

We are currently engaged in further exploration of this probabilistic approach to NLU. We are in the process of

writing a program which will take English language input and produce Bayesian networks like those presented in this paper. We are examining a number of possible approaches to evaluating such diagrams. At the same time, we are trying to determine which conditions, like stl, to apply to our networks to get the proper distributions.

6 Conclusion

We have presented a semantics for interpreting probabilistic statements expressed in a first-order quantifier-free language. We have shown how this semantics constrains the probabilities which can be associated with the propositions. Finally, we saw that while the semantics dictates very low prior probabilities for many of the statements we needed, once they are adequately conditioned, in particular with spatio-temporal locality, the probabilities become more "reasonable." We suggested that our notion of spatio-temporal locality, and the notion of discourse segment found in current AI NLU work are a', least very close, and may be identifiable. In our estimation this possibility sheds some interesting light on the notion of discourse segments, since it allows for their computation in a probabilistic way. Those familiar with the work in the area will be aware of how hard it has proven to give deterministic, non-circular rules about when such segments are to be created, and what can be determined from their creation.

7 Acknowledgments

We would like to thank the reviewers, Kate Sanders, and Mark Johnson for their comments.

References

[Allen, 1987] James Allen. *Natural Language Understanding.* Benjamin/Cummings Publishing Company, Menlo Park, California, 1987.

[Bacchus, 1988] Fahiem Bacchus. Statistically founded degrees of belief, pages 59-66, 1988.

[Bundy, 1988] Alan Bundy. Incidence calculus: A mechanism for probabilistic reasoning. In John F. Leinrner and Laveen F. Kanal, editors, *Uncertainty in Artificial Intelligence 2,* pages 177-184. North-Holland, 1988.

[Charniak and Goldman, 1988] Eugene Charniak and Robert P. Goldman. A logic for semantic interpretation. In *Proceedings of the Annual Meeting of the ACL,* 1988.

[Charniak and Goldman, 1989] Eugene Charniak and Robert P. Goldman. Plan recognition in stories and in life, forthcoming, 1989.

[Goldman and Charniak, 1988] Robert P. Goldman and Eugene Charniak. A probabilistic ATMS for plan recognition. In *Proceedings of the Plan-recognition workshop,* 1988.

[Grosz and Sidner, 1986] Barbara J. Grosz and Candace Sidner. Attention, intention and the structure of discourse. *Computational Linguistics,* 12, 1986.

[Hobbs *et al.*, 1988] Jerry R. Hobbs, Mark Stickel, Paul Martin, and Douglas Edwards. Interpretation as abduction. In *Proceedings of the 26th Annual Meeting of the ACL,* pages 95-103, 1988.

[Nilsson, 1986] Nils Nilsson. Probabilistic logic. *Artificial Intelligence,* 28:71-88, 1986.

[Pearl, 1988] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann Publishers, Inc., 95 First Street, Los Altos, CA 94022, 1988.

[Webber, 1987] Bonnie Lynn Webber. The interpretation of tense in discourse. In *Proceedings of the 25th Annual Meeting of the ACL,* 1987.