Probabilistic Resolution of Anaphoric Reference

John D. Burger & Dennis Connolly

The MITRE Corporation

{ john / decon } @linus.mitre.org

Abstract

This paper describes the use of a Bayesian network to resolve anaphora by probabilistically combining linguistic evidence. By adopting a Bayesian approach, we are able to combine diverse evidence in a principled way, extend current understanding of linguistic phenomena by quantifying relationships empirically, and better model the non-deterministic role of linguistic evidence in resolution of anaphora. We also briefly discuss our current research into the use of learning techniques to automatically construct Bayesian networks from data.

1 Introduction

Considerable research has been performed in the area of anaphoric reference resolution. A number of theories have been proposed for identifying the referent of an anaphor and performing the related task of identifying discourse focus. These theories apply various sources of linguistic evidence to the identification task. No single theory, however, considers all of the available evidence. Furthermore, little work has been done combining these theories,[1] and consequently the linguistic evidence that they entail. In addition, the individual theories are often characterized by an intuitive determination of preference criteria, making it difficult to evaluate competing theories and to combine their results.

This paper describes the use of Bayesian techniques to both combine sources of linguistic evidence in a principled way and derive individual contributions empirically from large text corpora. A Bayesian network is used to compute the probabilities of competing reference hypotheses from available linguistic evidence.[2] By representing the reference resolution problem with a Bayesian network, we gain the following advantages:

- Diverse evidence is combined in a principled way.
- Statistics used in this probabilistic technique can be gleaned from tagged and parsed text corpora.[3]
- Empirically derived contributions can improve upon and generalize the incomplete understanding of phenomena provided by current theories.
- Probabilistic treatment more accurately model the "non-deterministic" way in which linguistic evidence supports reference hypotheses.

The paper is organized as follows. In the following section, we briefly review Bayesian networks. In Section 3, we describe some of the sources of evidence that play a role in resolving anaphoric reference. In Section 4, we describe a Bayesian network that we believe is a fairly accurate characterization of the way in which various linguistic phenomena contribute to resolving pronominal and definite anaphoric reference and the dependencies between these phenomena. Section 5 provides an example to illustrate use of this network in resolving anaphora. In Sections 6 and 7 we discuss some of the issues encountered in our approach and how we are addressing these with our current learning research.

2 Bayesian networks

A Bayesian network is a graph in which propositions are represented as nodes and conditional probabilities between propositions are represented as links between corresponding nodes (see Pearl [11] and Charniak [5]). Each node is a multi-valued random variable representing the current belief (probability) of each possible value that the proposition can take. Associated with each link is a link matrix containing the conditional probabilities $p(x \mid y)$ for all possible values of the two variables connected by the link. Certain nodes represent directly

[1] But see Rich and LuperFoy [13] and LuperFoy [10].

[2] The use of Bayesian networks for semantic interpretation is described in Charniak and Goldman [6], while Burger and Davis [4] apply these methods to syntactic attachment disambiguation.

[3] There are already some moderately large corpora that have been parsed by hand, and automatic and semi-automatic analysis techniques are being developed [9].

observed evidence that is introduced into the network, while the other nodes represent hypotheses supported by this evidence. Typically, one of these is the hypothesis whose value is of interest. Beliefs are updated via a propagation algorithm that computes new probabilities at each node from the probabilities of its neighbors and corresponding link matrices whenever new evidence is supplied to the network.

By explicitly representing conditional dependence relationships between variables using links, the probabilities of individual propositions may be computed efficiently and locally by referring to only those variables that are immediately connected. Furthermore, only those conditional probabilities associated with links must be computed and stored rather than the complete joint distribution of all propositions.

The critical issue in designing Bayesian networks is determining the dependencies (and therefore links) between variables and structuring the network such that computations can be performed efficiently. This means creating a simple topology in which the propagation algorithm can take advantage of the conditional independence relationships to perform simple local computations. Achieving this often involves introducing additional "hidden" variables (intermediate hypotheses) into the network that render the other variables conditionally independent given the hidden variable.

3 Sources of evidence

Many sources of evidence have been identified for helping to resolve anaphoric reference. This section briefly describes some of the sources of evidence that have been brought to bear on the reference resolution problem, along with some of the theories that attempt to explain how this evidence can be understood with respect to linguistic phenomena.

c-command - *C-command* (see Reinhart [12]) is a syntactic constraint that places restrictions upon the forms that coreferring noun phrases can take, depending upon their relative location with respect to a parse tree. One object is c-commanded by another if it is a sibling or descendant of a sibling of the other in the parse tree. The *c-command* constraint includes restrictions such as the requirement that reflexive pronouns be c-commanded by the NP that they corefer with.

semantic agreement - This constraint states that coreferring NPs must be descriptions consistent with the same real world entity. Evidence for this includes agreement in gender, person, and count. In addition, we make use of a semantic representation, e.g., a KL-ONE-style term subsumption hierarchy (see Brachman and Schmolze [2]).[4] Given such a representation of the meanings of referring expressions, one can take as evidence the subsumption relationship among the representations. In addition, a weaker semantic similarity relationship between knowledge representation entities may be employed. Furthermore, impoverished lexical knowledge may be augmented with morphological knowledge and verb compliment evidence.

discourse focus - An important source of evidence for anaphor resolution is the related notion of *discourse focus*. This is defined in many ways, but usually refers to the entities that are most likely to be referred to at any given point in the discourse and that roughly correspond to what that part of discourse is about. We make use of a model of focus similar to that described in Sider and Burger [15], which is related to that of Grosz and Sidner ([8, 16]). Such a model, of course, only begs the question, requiring evidence in its own right.

discourse structure - A common approach to tracking discourse focus is to explicitly represent the structure of the discourse (see Grosz and Sidner [8]). Typically, this corresponds to a tree structure that partitions the text into nested segments. Within this tree structure, the closer a discourse entity is to an anaphor, the more it is in focus with respect to that anaphor. In task-oriented discourses, discourse structure usually corresponds to the structure of the underlying task. This approach usually requires domain knowledge associated with limited tasks. In the more general case in which a more impoverished world model is available, the focus space is represented as a stack that is manipulated via weaker methods such as cue phrases and failure driven transitions (e.g., a change in focus is indicated when the current focus cannot corefer with the anaphor on semantic grounds).

recency - A source of anaphoric reference evidence that is closely related to focus is *recency*. In its simplest form, this evidence states that more recent linguistic entities are more likely to corefer with an anaphor than less recent entities. Clearly, this evidence interacts strongly with the *task structure* approach described above. This evidence, however, is trivially computed, more widely applicable, and may be seen as a heuristic approximation to the more complicated task structure that actually underlies a discourse. This approximation exhibits the greatest utility for pronominal anaphora.

centering - *Centering* is an approach to focus that attempts to explain the local relationships between utterances that contribute to discourse coherence (see Grosz, *et al.* [8], as well as Brennan, *et al.* [3]). Its primary elements are *centers*, which are discourse entities that tie the current utterance to its neighbors. For the current utterance U_n, the *backward center*, $C_b(U_n)$,

[4] For details on the particular knowledge representation formalism employed here, see Bayer and Vilain [1].

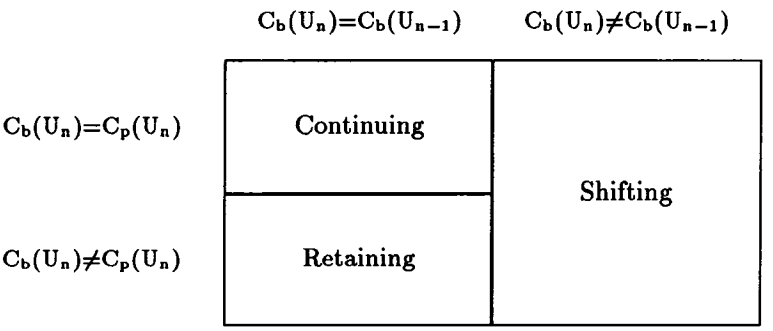

Figure 1: Transitions among adjacent utterances, defined in terms of centers

is the discourse entity that is the focus of that utterance. *Forward centers*, $C_f(U_n)$, are entities that may become the focus of subsequent utterances. The most likely of these to become the subsequent focus is the *preferred center*, $C_p(U_n)$. Typically, this preference is associated with an ordering on grammatical role (e.g., prefer subject over direct object, etc.) The relationships among $C_f(U_n)$, $C_b(U_n)$ and $C_b(U_{n-1})$ (the backward center of the previous utterance) define certain kinds of *transitions* between adjacent utterances, as shown in Figure 1. These transitions essentially capture the fact that discourse is coherent, and that changes in focus occur only in certain, well-defined ways. The most likely transition is *continuing*, which occurs when the focus remains the same. The next most likely transition is *retaining*, which signals an intention to shift focus. Finally, *shifting* indicates that an actual change of focus has occurred. A further constraint is that if one of the *forward centers* of an utterance is a pronoun, then the *backward center* should also be a pronoun.

miscellaneous - Some of the elements of specific theories such as *centering* have been proposed independently of such theories and may warrant separate treatment. Examples include the variants of the restriction that focus be pronominal when other pronouns appear in the sentence and preferences for particular grammatical or thematic roles (e.g., subject or theme) as focus.

4 Anaphoric reference network

This section describes our attempt to organize several of the sources of linguistic evidence described above into a Bayesian network for probabilistically predicting anaphoric reference. This task is considerably more complex than one might expect. The primary issue is identifying the conditional dependencies between the various theories and evidence sources. This is difficult because many theories address overlapping linguistic phenomena or rely on overlapping linguistic evidence. Furthermore, many seemingly distinct surface linguistic phenomena reflect underlying deep phenomena that are related.

The Bayesian network for resolving anaphoric reference is shown in Figure 2. This network has been constructed by hand, after a careful analysis of the relevant linguistic theory. The conditional probabilities associated with the links are estimated from relative frequencies derived from tagged text corpora.[5] The grey nodes are where evidence is introduced into the network. The node labeled **corefer(x, y)** represents the boolean hypothesis that **x** and **y** corefer, where **y** is an anaphoric NP and **x** is a candidate referent. For each anaphor to be resolved and each candidate to be evaluated, **y** and **x** are bound and the network is re-evaluated using the corresponding evidence. The instantiation of **x** with the highest probability is chosen to resolve the anaphor **y**. While comparison of separate boolean hypotheses ignores possible interactions between the hypotheses, this is a simplification that is necessary since the number of hypotheses is essentially unbounded. The remainder of this section discusses some highlights of the Bayesian network.

The portion of the Bayesian network identified in the figure as subnet A is the network sub-structure that is collectively responsible for *centering* phenomena. This portion of the network is described in the following paragraphs.

[5] We are currently in the process of tagging corpora. Pending completion of this process, we approximate conditional probabilities using subjective estimates provided by human experts. While both Pearl [11] and Charniak [5] argue that such an approach is justified, we feel that it is important to use empirically derived estimates in a domain as complex as linguistics and have adopted this measure only to assess the gross viability of our approach while we are collecting statistics. The tagging process will be partially automated using NLP components including a partial parser currently under development. These components will also provide values for evidence nodes when this network is actually used to resolve references.

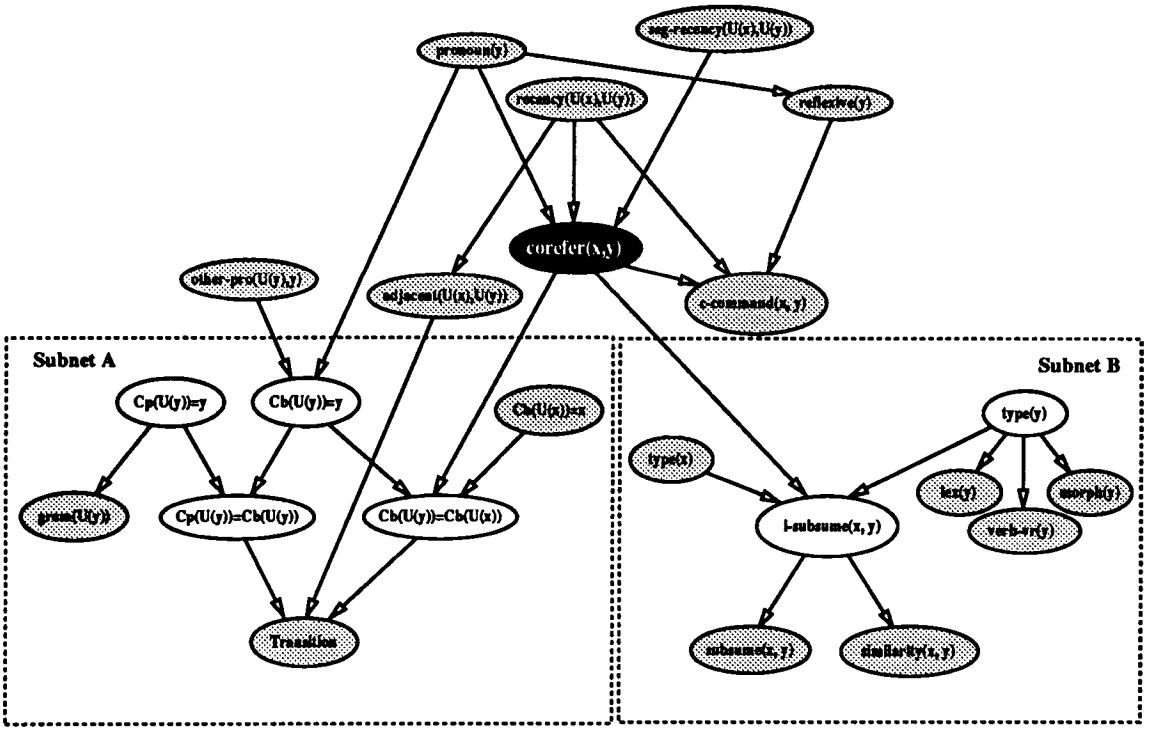

Figure 2: Bayesian Network for Anaphoric Reference.

The variables labeled $C_p(U(y))=y$, $C_b(U(y))=y$, and $C_b(U(x))=x$ correspond to the *centers*. $C_p(U(y))=y$ is the boolean proposition that the NP to which **y** is bound is the *preferred center* of the utterance **U(y)** in which it appears. $C_b(U(y))=y$ is the proposition that this NP is the *backward center* of that utterance (i.e., the *current backward center*). $C_b(U(x))=x$ is the proposition that the NP to which **x** is bound is the *backward center* of the utterance in which it appears (i.e., the *previous backward center*).

The variable **transition** corresponds to the centering transitions with values of **continuing**, **retaining**, and **shifting**. These are defined in terms of the variables $C_p(U(y))=C_b(U(y))$ and $C_b(U(y))=C_b(U(x))$, as described earlier. The former proposition is true simply when $C_p(U(y))=y$ and $C_b(U(y))=y$ are both true. The latter is true when $C_b(U(y))=y$ and $C_b(U(x))=x$ are true, *and* **x** and **y** corefer. These relationships are actually represented via joint distributions between the variables. The relative likelihoods among the various transitions (e.g., continuing is preferred over shifting) is represented by the a priori probabilities for **transition**.

The variables **pronoun(y)** and **other-pro(U(y),y)** represent the propositions that the the NP bound to **y** is a pronoun and that another NP in the same utterance is pronominal, respectively. The joint distribution between these variables and $C_b(U(y))=y$ represents the constraint that the *backward center* be pronominal if other pronouns are present.

The variable **gram(U(y))** represents the grammatical case assignments of the NPs in the utterance in which **y** appears. The distribution in the link to $C_p(U(y))=y$ captures the ordering imposed upon the *forward centers* by grammatic role (e.g., prefer subject).

The variable **adjacent((U(x),U(y))** states simply that the utterances in which **x** and **y** appear are adjacent (i.e., **recency(U(x),U(y))** = 1). The link to **transition** reflects the fact that centering only makes sense for adjacent sentences.

Finally, it should be noted that the distribution for $C_b(U(x))=x$, the *previous backward center*, is actually derived by using the distribution that was computed in $C_b(U(y))=y$ when **y** was bound to the NP that **x** is currently bound to. This amounts to treating $C_b(U(x))=x$ as an evidence node.

Collectively, this sub-network provides evidence for **corefer(x, y)**, by favoring assignments to **x** and **y** that are consistent with transitions predicted by the centering model.

Subnet B is responsible for capturing *semantic agreement* evidence. Subsumption and semantic similarity with respect to a known lexical hierarchy are sources of evidence that are clearly not independent. This rela-

tionship is accounted for in our network via the intermediate variable **i-subsume** (ideal subsumption). The intuition is that for two entities to corefer, they must be related via a subsumption relationship in some world in which the references are members. This relationship is captured by subsumption in an ideal hierarchy that is only approximated by the hierarchy available during linguistic processing. Both subsumption and semantic similarity with respect to this hierarchy provide evidence for the actual relationship in the ideal hierarchy. Knowledge of the ideal relationship would completely determine the distributions of the subsumption and similarity relationships observed, thus rendering these variables conditionally independent. Additionally, these relationships are defined in terms of the types of the entities in question. The actual types are also unknown, but are probabilistically determined by the available lexicon, morphological evidence, and evidence from verb complement restrictions. Sources of evidence such as syntactic gender and count agreement are a special case of ideal subsumption in which the morphological evidence predicts type via syntactic features. Since the subsumption relationship is completely dependent on the types of both arguments, **type(x)** and **type(y)** are not conditionally independent, and the joint distribution of these values with respect to **i-subsume** must be explicitly represented.

Other evidence considered in this network includes recency, c-command, and segment recency. Recency evidence is represented by the variable **recency(U(x),U(y))**. Its effect on **corefer(x, y)** is represented by its joint distribution with **pronoun(y)**, reflecting the higher relevance of recency when the anaphor is pronominal. C-command evidence is represented by the joint conditional distribution of the variable **c-command(x,y)** with respect to the variables **corefer(x, y)**, **reflexive(y)**, and **recency(U(x),U(y))**. **c-command(x,y)** is the proposition that the NP to which **x** is bound c-commands the NP to which **y** is bound. **reflexive(y)** is the proposition that the NP to which **y** is bound is a reflexive pronoun. **recency(U(x),U(y))** appears in this distribution because c-command is not independent of recency. Segment recency captures the notion of global focus. The variable **seg-recency(U(x),U(y))** represents how far down the segment stack the utterance containing **x** is.

5 Example

To illustrate use of the Bayesian network described above to resolve anaphora, consider the fragment of text[6] shown in Figure 3. The NP, *the company's*, shown highlighted in the third paragraph of this text fragment,

Philip Morris Cos. is launching a massive corporate advertising campaign that will put the tobacco giant's name in TV commercials for the first time since the early 1950s, when it stopped advertising its namesake cigarette brand on television.

⋮

The company is expected to spend about $30 million a year on its two-year corporate campaign, created by WPP Group's Ogilvy & Mather unit in New York.

⋮

Philip Morris, which became the U.S.'s largest food company last year with its $12.9 billion acquisition of Kraft Inc., seems determined to evolve beyond its roots in Marlboro country. **The company's** *research suggests that its name recognition among most consumers remains unusually low, although its array of brands—including Maxwell House coffee, Jell-O, Cheez Whiz, and Miller beer—blanket supermarket shelves.*

Figure 3: Example Fragment of Text.

is a definite anaphoric reference. Other NPs in the fragment that are candidates for corefering with this anaphor include *a massive corporate advertising campaign*, *WPP Group's Ogilvy & Mather unit*, *Philip Morris*, *Kraft Inc.*, and *Marlboro country*. Our network assigns the following belief values to these candidates,

a massive corporate advertising campaign	0.003
WPP Group's Ogilvy & Mather unit	0.180
Philip Morris	0.890
Kraft Inc.	0.690
Marlboro country	0.025

The highest belief is attributed to *Philip Morris*, which is, therefore, selected to resolve the anaphor. In order to understand how these values are assigned, we must consider the evidence supporting the various candidates.

A major source of evidence is semantic agreement. This is represented by strong conditional probabilities on the link between the variables **corefer(x, y)** and **i-subsume** in the network. This evidence source, alone, would result in low belief values for *a massive corporate advertising campaign*, and *Marlboro country*.

Of the the remaining candidates, recency provides the greatest support for *Philip Morris* and *Kraft Inc.*. *Philip Morris*, however, receives greater support from the centering portion of the network. The reason for this is that *Philip Morris* and *the company's* are both subjects of the sentences in which they appear. As a

[6] This example is taken from the Wall Street Journal [7].

result, *the company's* is likely to be the preferred center and *Philip Morris* the backward center, of their respective sentences. (In the first sentence of a paragraph, we chose the preferred center as the backward center.) With these assignments, the a priori preference for *continuing* supports making *the company's* the backward center of its sentence and requiring that *the company's* corefer with *Philip Morris*. As a result of these considerations, *Philip Morris* achieves the high degree of belief shown above.

6 Learning Bayesian networks

The successful application of Bayesian networks to a problem relies heavily on correctly identifying the conditional dependence relationships between propositions. As suggested in this paper, doing this by hand is a difficult knowledge engineering task. Identification of these dependencies is a painstaking process involving carefully analyzing individual theories and characterizing the relationships and potential interactions between them.

Using Bayesian networks to model linguistic theory such as anaphora is particularly difficult, in part due to the inherent complexity, but more importantly because linguistic theory is an area of current research. Precisely because linguistic theory is still being developed, it does not easily lend itself to the knowledge engineering required to identify dependencies. Some theories reflect a better understanding than others, but none yet provide a complete understanding of the phenomena that they attempt to explain and, as stated earlier, no single theory explains all phenomena. This reliance on evolving theories to inform the knowledge engineering process not only renders this process exceedingly difficult, but casts doubt upon the correctness of the resulting model.

To further confound efforts to model these phenomena with Bayesian networks, some of the important components of these theories, components that must be modeled by variables in the Bayesian network to account for important dependencies, are artifacts of the linguistic theory that are not necessarily directly observable. Discourse focus, for example, is useful in explaining a number of observed phenomena but is itself an abstraction. This introduces the additional problem of estimating conditional probabilities for these unobserved events.

As a result of these considerations, we have been investigating machine learning techniques for automatically constructing such networks. In our current approach, we use mutual information clustering to induce network topology and have derived a learning algorithm analogous to neural network backpropagation [14] to determine the conditional probability weights. The latter is described in greater detail in the Appendix. By using learning to derive the network empirically, we hope to reduce knowledge engineering, while improving and generalizing our understanding of linguistic phenomena and their interactions. Finally, as suggested above, even when a Bayesian network is derived by hand as was described in Section 4, some of the variables may still be truly hidden variables in the sense that they reflect non-observable phenomena, and consequently, the corresponding conditional probabilities cannot be directly estimated. In this situation, the backpropagation learning algorithm can be used to assign the probabilities.

7 Discussion

This paper has described the use of Bayesian networks to resolve anaphora by probabilistically combining linguistic evidence. By adopting this approach, we are able to combine diverse evidence in a principled way, extend current understanding of linguistic phenomena with dependency models partly derived from data, and better model non-deterministic phenomena. While illustrating the central ideas, and hopefully the usefulness, of such an approach, this work exposes additional issues that we hope to address in future work.

The successful application of Bayesian networks to a problem relies heavily on correctly identifying the correct conditional dependence relationships between propositions. As suggested in the previous section, this is a difficult proposition when applied to evidence from linguistic theories such as those dealing with anaphora. One cannot treat separate theories as "black boxes" whose results are simply combined probabilistically, because the phenomena (both surface and deep) that these theories entail either directly overlap or strongly interact. Furthermore, identification of these interactions is a painstaking process involving carefully analyzing individual theories and characterizing the relationships between them. Even when this process is successful, we cannot guarantee correctness, since the individual theories are in fact only theories. Some theories reflect a better understanding than others, but none yet provide a complete understanding of the phenomena that they explain.

As we suggest, we believe that the most promising avenue for future research is to investigate machine learning techniques for automatically deriving the topology of Bayesian networks. While the painstaking process of constructing such a network by hand has been a useful exercise that both suggests the utility of Bayesian techniques for anaphoric reference, and provides a better understanding of the issues involved, we expect this to be the last such exercise that we undertake. By automatically deriving networks empirically, we hope to more accurately model the dependencies between phe-

nomena that are not fully understood, as well as provide an even better model of the individual phenomena themselves and further generalize existing theories. In addition to motivating such an undertaking, the hand-derived network described in this paper identifies many of the evidence sources required for this effort and may help to interpret the results of this learning.

Appendix: Backpropagation in Bayesian networks

In order to determine the correct conditional probabilities or weights in the Bayesian network, we compute the partial derivatives of an error function described below with respect to these weights and perform gradient descent. There are two cases of interest. In the first case, we are concerned with the weights between the root of the tree, Z, and an immediate child. Since these weights directly influence the belief at node Z where the error is measured, this computation is straightforward. In the second case we are concerned with the weights between a "hidden" variable (i.e. a variable other than the root) and one of its children. Since these weights only indirectly influence the error through intermediate nodes, the derivatives cannot be computed directly. This credit assignment problem is similar to the problem of training hidden units in neural networks, and indeed our approach to learning conditional probabilities is analogous to that used in the backpropagation approach [14].

We define the error for a tree-structured Bayesian network rooted at the variable Z as the sum-squared error:

$$E = \sum_z [BEL(z) - T(z)]^2,$$

where z ranges over the values of Z, T is the correct probability for z provided by an oracle, and $BEL(z)$ is the probability attributed to z by the Bayesian network using the belief update equations described in Pearl [11]:

$$BEL(z) = \alpha \lambda(z) \pi(z)$$

$$\lambda(z) = \prod_j \lambda_{Y_j}(z)$$

$$\lambda_Y(z) = \sum_y \lambda(y) p(y \mid z)$$

Drawing inspiration from the derivation of the backpropagation algorithm in neural networks, we derive expressions for the partial derivatives for the two cases described above that prescribe a recursive update procedure for hidden variables:

Case I:

$$\Delta_{yz} = \frac{\partial E}{\partial p(y \mid z)} = 2\,[BEL(z) - T(z)]\, BEL(z) \frac{\lambda(y)}{\lambda_Y(z)}$$

Case II:

$$\Delta_{xy} = \frac{\partial E}{\partial p(x \mid y)} = \left[\sum_z \Delta_{yz} p(y \mid z) \right] \frac{\lambda(x)}{\lambda_X(y)}$$

References

[1] Samuel Bayer and Marc Vilain. The relation-based knowledge representation of KING KONG. In *Working Notes of the AAAI Spring Symposium on Implemented Knowledge Representation Systems*, Stanford, CA, March 1991.

[2] Ronald J. Brachman and James G. Schmolze. An overview of the KL-ONE knowledge representation system. *Cognitive Science*, 9:171–216, April–June 1985.

[3] Susan E. Brennan, Marilyn W. Friedman, and Carl J. Pollard. A centering approach to pronouns. In *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics*, Stanford, 1987.

[4] John D. Burger and Anthony R. Davis. Probabilistic ambiguity resolution in KING KONG. In *Proceedings of the AAAI Workshop on Knowledge-Based Construction of Probabilistic and Decision Models*, Anaheim, CA, July 1991.

[5] Eugene Charniak. Bayesian networks without tears. *AI Magazine*, 12(4):50–63, 1991.

[6] Eugene Charniak and Robert Goldman. 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*, pages 1074–1079, 1989.

[7] Alix M. Freedman. Philip Morris to launch image ads. *Wall Street Journal*, November 1, 1992.

[8] Barbara J. Grosz and Candace L. Sidner. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12:175–204, 1986.

[9] Donald Hindle. Noun classification from predicate argument structures. In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, pages 268–275, 1990.

[10] Susann LuperFoy. *Discourse Pegs: A Computational Analysis of Context-Dependent Referring Expressions.* PhD thesis, University of Texas at Austin, 1991.

[11] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann, Los Altos, CA, 1988.

[12] Tanya Reinhart. *The Syntactic Domain of Anaphora.* PhD thesis, MIT, Cambridge, MA, 1976.

[13] Elaine A. Rich and Susann LuperFoy. An architecture for anaphora resolution. In *Proceedings of the Second Conference on Applied Natural Language Processing,* Austin, 1988.

[14] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagation. In David E. Rumelhart and James L. McClelland, editors, *Parallel Distributed Processing* Vol. 1. MIT Press, 1986.

[15] Judith Schaffer Sider and John D. Burger. Intention structure and extended responses in a portable natural language interface. *User Modeling and User-Adapted Interaction,* 2(1–2), 1992.

[16] Candace L. Sidner. The use of focus as a tool for the disambiguation of definite noun phrases. In *TINLAP 2: Theoretical Issues in Natural Language Processing.* ACM and ACL, 1978.