Abduction for Discourse Interpretation: A Probabilistic Framework

Ekaterina Ovchinnikova
USC/ISI
4676 Admiralty Way
Marina del Rey, CA 90292
katya@isi.edu

Andrew S. Gordon
USC/ICT
12015 Waterfront Drive
Los Angeles, CA 90094-2536
gordon@ict.usc.edu

Jerry Hobbs
USC/ISI
4676 Admiralty Way
Marina del Rey, CA 90292
hobbs@isi.edu

Abstract

Abduction allows us to model interpretation of discourse as the explanation of observables, given additional knowledge about the world. In an abductive framework, many explanations can be constructed for the same observation, requiring an approach to estimate the likelihood of these alternative explanations. We show that, for discourse interpretation, weighted abduction has advantages over alternative approaches to estimating the likelihood of hypotheses. However, weighted abduction has no probabilistic interpretation, which makes the estimation and learning of weights difficult. To address this, we propose a formal probabilistic abductive framework that captures the advantages weighted abduction when applied to discourse interpretation.

1 Introduction

In this paper, we explore discourse interpretation based on a mode of inference called *abduction*, or inference to the best explanation. Abduction-based discourse processing was studied intensively in the 1980s and 1990s (Charniak and Goldman, 1989; Hobbs et al., 1993). This framework is appealing because it is a realization of the observation that we understand new material by linking it with what we already know. It instantiates in discourse understanding the more general principle that we understand our environment by coming up with the best explanation for the observables in the environment. Hobbs et al. (1993) show that abductive proofs can be efficiently exploited for a whole range of natural language pragmatics problems, such as word sense disambiguation, anaphora and metonymy resolution, interpretation of noun compounds and prepositional phrases, and detection of discourse relations. As applied to discourse interpretation, abduction was shown to have advantages over deduction, a more classical mode of inference (Ovchinnikova, 2012). One serious advantage concerns treatment of incomplete knowledge. In the cases when it is impossible to provide it with all the knowledge which is relevant for interpretation of a particular piece of text, deductive reasoners fail to find a prove. Instead of a deterministic yes/no proof abduction provides a way of measuring in how far the input formula was proven and which of its parts could not be proven.

In the early 90s, research on abduction-based discourse processing resulted in good theoretical work and in interesting small-scale systems, but it faced three difficulties: 1) parsers were slow and not accurate enough, so that inference had no place to start, 2) inference processes were neither efficient nor accurate enough, 3) there was no large knowledge base designed for discourse processing applications. In the last two decades, the first of these difficulties has been addressed by progress in statistical parsing, e.g. (McClosky et al., 2006; Huang, 2008; Bos, 2011). Recently, efficient reasoning techniques were developed that overcome the second difficulty (Inoue and Inui, 2011; Inoue et al., 2012b). Finally, it has been shown that there exists sufficient knowledge about the world – at a level of precision that enables its translation into formal logic – available in a variety of resources (Ovchinnikova et al., 2011; Ovchinnikova, 2012). These advances have recently been capitalized upon in several large-scale applications of abduction to discourse processing tasks (Inoue and Inui, 2011; Ovchinnikova et al., 2011; Ovchinnikova, 2012; Inoue et al., 2012a).

In an abductive framework, often many explanations can be provided for the same observation. In order to find the best solution for our pragmatic problem, we need to be able to choose the best, i.e. the most probable, explanation. Several ap-

proaches were proposed for estimating the likelihood of alternative abductive explanations: cost-based abduction (Charniak and Shimony, 1990), weighted abduction (Hobbs et al., 1993), abduction based on Bayesian Networks (Pearl, 1988; Charniak and Goldman, 1989; Raghavan and Mooney, 2010), abduction based on Markov Logic Networks (Kate and Mooney, 2009).

In this paper, we show that weighted abduction employing a cost propagation mechanism (see Section 3) and favoring low-cost explanations has certain features relevant for discourse processing that other approaches do not have (see Section 4). The main such feature is the approach to unification, i.e. associating two entities with each other, so that their common properties only need to be proved or assumed once (see Section 2). Weighted abduction favors explanations with the maximum number of unifications. Thus, it favors those explanations that link parts of observations together and supports discourse coherence, which is crucial for discourse interpretation.

There is not yet any work on linking weights in weighted abduction to probabilities, which makes the estimation and learning of the weights difficult. In this paper, we show that the original cost propagation mechanism in weighted abduction as informally introduced in (Hobbs et al., 1993) cannot be interpreted in terms of probabilities. However, we can still capture features of weighted abduction desirable for discourse processing in a formal probabilistic framework based on Bayesian Networks. As a result, we obtain a theoretically sound probabilistic abductive framework favoring explanations relevant for discourse interpretation.

2 Abduction

Abduction is inference to the best explanation. Formally, logical abduction is defined as follows:
Given: Background knowledge B, observations O, where both B and O are sets of first-order logical formulas,
Find: A hypothesis H such that $H \cup B \models O, H \cup B \not\models \bot$, where H is a set of first-order logical formulas.

Observation O is usually a conjunction of existentially quantified propositions (Charniak and Goldman, 1989; Hobbs et al., 1993; Raghavan and Mooney, 2010):

$$\exists x_1, ..., x_k, ..., y_1, ..., y_l (q_1(x_1, ..., x_k) \wedge ... \wedge q_n(y_1, ..., y_l)).$$

We extend the notion of observation by allowing inequalities ($x \neq y$) as conjuncts. Sometimes inequalities follow from the natural language syntax. For example, if we read *There is a cat on the mat. Another cat is on the table*, we immediately know that there are two different cats mentioned. This text can be logically represented as follows:

$$\exists x_1, x_2, y_1, y_2 (cat(x_1) \wedge on(x_1, y_1) \wedge mat(y_1) \wedge cat(x_2) \wedge on(x_2, y_2) \wedge table(y_2) \wedge x_1 \neq x_2)).$$

Background knowledge B is a set of first-order logic formulas. In order to keep the inference process computationally tractable, B is often restricted to a set of Horn clauses (Charniak and Shimony, 1990; Hobbs et al., 1993; Kate and Mooney, 2009; Raghavan and Mooney, 2010). Thus, each background axiom has the form

$$P_1 \wedge ... \wedge P_n \rightarrow Q,$$

where all variables on the left-hand side are universally quantified with the widest possible scope and all variables occurring on the right-hand side only are existentially quantified. We weaken this restriction allowing multiple literals on the right-hand side of the background axioms because of the importance of the context and compositionality for discourse interpretation. For example, in order to express the fact that a testing process can be called "dry run", we use the following axiom:

$$\forall x, y, e, z, u (process(x) \wedge of(x, e) \wedge test(e, z, u) \rightarrow dry(x) \wedge run(x)).$$

Breaking this axiom into two different axioms (one implying that the process is dry and the other implying that it is a run) will result in loosing the binding of the arguments of *dry* and *run*.

We allow inequalities ($x \neq y$) as conjuncts in the background axioms. Inequalities can be used to represent incompatibility. For example, the axiom below represents the fact that the arguments of the relation *parent_of* refer to different objects:

$$\forall x, y (parent\_of(x, y) \rightarrow x \neq y).$$

The two main inference operations in abduction are backchaining and unification. *Backchaining* is the introduction of new assumptions given an observation and background knowledge. For example, given $O = q(A)$ and $B = \{\forall x(p(x) \rightarrow q(x))\}$, there are two candidate hypotheses: $H_1 =$

$q(A)$ and $H_2 = p(A)$. We say that $p(A)$ *explains* $q(A)$ in H_2. If an atomic proposition is included in a hypothesis (*hypothesized*) and not explained, then it is *assumed*, e.g., $q(A)$ is assumed in H_1.

Unification is merging of propositions with the same predicate name by assuming that their arguments are same.[1] For example, $O = \exists x, y(p(x) \wedge p(y) \wedge q(y))$. Given this observation, the propositions $p(x)$ and $p(y)$ are unifiable. Thus, there is a hypothesis $H = \exists x(p(x) \wedge q(y) \wedge x = y)$.

Both operations (backchaining and unification) can be applied as many times as possible to generate a possibly infinite set of hypotheses. The generation of the set of hypotheses $\mathcal{H}$ initialized as an empty set can be formalized as follows.

Backchaining

$$\frac{\bigwedge_{i=1}^{n} P_n \rightarrow \bigwedge_{j=1}^{m} Q_j \in B \text{ and } O \wedge H \models \bigwedge_{j=1}^{m} Q_j \text{ and } O \wedge H \wedge \bigwedge_{i=1}^{n} P_n \not\models \bot, \text{ where } H \in \mathcal{H}}{\mathcal{H} := \mathcal{H} \cup \{H \wedge \bigwedge_{i=1}^{n} P_n\}}$$

Unification

$$\frac{O \wedge H \models p(X) \wedge p(Y) \text{ and } O \wedge H \wedge X = Y \not\models \bot, \text{ where } H \in \mathcal{H}}{\mathcal{H} := \mathcal{H} \cup \{H \wedge X = Y\}}$$

3 Estimating Hypothesis Likelihood

Often many hypotheses can be constructed for the same observation. In order to find the best solution for our pragmatic problem, we need to choose the best, i.e. the most probable, hypothesis. Several approaches were proposed for estimating the likelihood of alternative abductive explanations.

Charniak and Shimony (1990) propose cost-based abduction. In this framework, the likelihood of a hypothesis depends on the probability of the assumed atomic propositions to be true.

Another popular approach to abduction is based on Bayesian Networks (Pearl, 1988; Charniak and Goldman, 1989; Raghavan and Mooney, 2010). In this framework, abductive explanations are represented by a directed graph constituting a Bayesian net, such that the nodes of the graph correspond to atomic predications and the edges connect explanations with the predications they explain. Each node has an associated conditional probability $P(A|B)$, where B is an explanation of A. Given the constructed Bayesian net, the best abductive hypothesis is selected using standard methods, which assign values to the unobserved nodes in the network that maximize the posterior probability of the joint assignment given the observations.

One more approach developed by (Kate and Mooney, 2009) is based on Markov Logic Networks (MLNs) (Richardson and Domingos, 2006). In this approach, a weight is assigned to each background axiom that reflects the strength of a constraint it imposes on the set of possible worlds. The higher the weight, the lower the probability of a world that violates the axiom. An MLN can be viewed as a set of templates for constructing Markov networks. Originally, MLNs employ deductive reasoning. Kate and Mooney (2009) adapt MLNs for abductive inferences by introducing reverse implications for every axiom in the knowledge base and adding mutual exclusivity constraints on the transformed axioms.

Finally, weighted abduction (Hobbs et al., 1993) proposes a cost propagation mechanism for selecting best hypotheses. In this framework, each atomic observation is assigned a positive real-valued cost. Atomic antecedents in the background axioms are assigned positive real-valued weights. If an axiom $\alpha = P \rightarrow Q$ is applied then the cost of each newly introduced literal p in P is equal to the sum of the costs of the literals in Q multiplied by the weight of p in α. For example, given the axiom $\forall x(p(x)^{0.9} \wedge s(y)^{0.1} \rightarrow q(x))$ and the observation $q(A)^{\$10}$, the literal $p(A)$ costs $\$10 \times 0.9 = \$9$ and the literal $s(y)$ costs $\$10 \times 0.1 = \$1$. When two literals are unified, the result of their unification is assigned the minimum of their costs. For example, given the observation $p(x)^{\$10} \wedge p(y)^{\$20}$ there is a hypothesis $x = y^{\$10}$. The cost of the hypothesis is equal to the sum of the costs of the assumptions. Each unification reduces the overall cost of the hypothesis, while an application of an axiom can increase or decrease the overall cost depending on whether its total weight is less or greater than 1. There is not yet any work on interpreting the weighted abduction cost propagation in terms of probabilities. Therefore the minimal cost hypothesis does not necessarily correspond to the most probable one.

All mentioned approaches to estimating the likelihood of abductive hypotheses have a common problem. The problem is that they all imply certain assumptions that cannot be proved or disproved practically because of the absence of the gold standard (collection of correct proof graphs)

[1] Note that the abduction unification mechanism is different from how unification is usually understood in computer science and logic, because it allows us to assume equalities of constants.

that is obviously very difficult to obtain. Cost-based abduction implies that the likelihood of a hypothesis depends on the joint likelihood of the assumptions only and that the assumptions are mutually independent. Abduction based on Bayesian Networks implies that the truth of the literals depends on their direct explanations only. MNL-based abduction implies that the probability of a background axiom to hold does not depend on the observation. All mentioned framework imply that unifications always hold.

In order to successfully apply abductive inference to pragmatic tasks, we should formulate the underlying independence assumptions with a good understanding of our domain of interest (in our case, it is discourse interpretation) and design a probabilistic framework correspondingly.

4 Abduction for Discourse Processing

Weighted abduction has three features, missing in other abduction-based frameworks, that are especially relevant for discourse processing. In this section, we discuss these features.

Unification The first feature is related to the unification inference. Weighed abduction prefers hypotheses with the maximum number of unifications. Therefore, it favors those explanations that link parts of observations together and thus support discourse coherence.

Suppose we want to construct an interpretation for the sentence *John composed a sonata*. The verb *compose* has two readings, 1) the "put together" reading (e.g., *The party composed a committee*, and 2) the "create art" reading. Suppose there are the following axioms:

1) $put\_together(e, x_1, x_2) \wedge collection(x_2) \rightarrow compose(e, x_1, x_2)$
2) $create\_art(e, x_1, x_2) \wedge work\_of\_art(x_2) \rightarrow compose(e, x_1, x_2)$
3) $sonata(x) \rightarrow work\_of\_art(x)$

Axioms (1) and (2) correspond to the two readings of *compose*. Axiom (3) states that a sonata is a work of art. Weighted abduction favors Axiom (2) over (1) for the observed sentence, because unification of *sonata* resulting from the application of Axioms 2 and 3 with the observable *sonata* reveals the implicit discourse redundancy and supports linking the meanings of *compose* and *sonata*.

As mentioned above, weighted abduction implies unconditional unification. In the discourse interpretation context, unification is one of the principal methods by which coreference is resolved. A naive approach to coreference in an inference-based framework is to unify propositions having the same predicate names unless it implies logical contradictions (Hobbs et al., 1993; Bos, 2011). However, in situations when knowledge necessary for establishing contradictions is missing, the naive procedure results in overmerging. For example, given $O = \exists x, y(animal(x) \wedge animal(y))$, we do not want to assume that x equals y when $dog(x) \wedge cat(y)$ are observed. For *John runs and Bill runs*, with the observations $O = \exists x, y(John(x) \wedge run(x) \wedge Bill(y) \wedge run(y))$, we do not want to assume that John and Bill are the same individual just because they are both running. If we had complete knowledge about incompatibility (*dog* and *cat* are disjoint, people have unique first names), the overmerging problem might not occur because of logical contradictions. However, it is not plausible to assume that we would have an exhaustive knowledge base. A proposal to introduce weighted unification is described in (Inoue et al., 2012a), where unification costs depend on the semantic relation (synonymy vs. antonymy), modality and polarity, and shared properties of the unified literals.

Observations costs The second feature concerns the unequal treatment of atomic observations depending on their initial cost. Hobbs et al. (1993) mention that costs reflect the demand for propositions to be proved. Those propositions that are most likely to be linked referentially to other parts of the discourse are expensive to assume. This idea is illustrated by an example provided in (Blythe et al., 2011). Suppose there are two sentences.

The smart man is tall.
The tall man is smart.

The logical representation for each of them is $\exists x(smart(x) \wedge tall(x) \wedge man(x))$. But certain syntactic features attached to propositions (e.g., definite article) influence the probability of the propositions to be explained or assumed. In the first sentence we want to prove $smart(x)$ to anchor the sentence referentially. Then $tall(x)$ is new information to be assumed. Blythe et al. (2011) suggest having a high cost on $smart(x)$ to force the proof procedure to find this referential anchor. The cost on $tall(x)$ will be low, to allow it to be assumed without expending effort in trying to locate that fact in background knowledge. For

the second sentence, the case is the reverse.

Suppose we know that educated people are smart and big people are tall, and furthermore that John is educated and Bill is big and both of them are men. This knowledge is formalized as follows:

$\forall x(educated(x) \rightarrow smart(x))$
$\forall x(big(x) \rightarrow tall(x))$
$educated(John), big(Bill), man(John),$
$man(Bill)$

In weighted abduction, the best interpretation for the first sentence is that the smart man is John, because he is educated, and the cost for assuming he is tall is paid. The interpretation to avoid is one that says x is Bill; he is tall because he is big, and the cost of assuming he is smart is paid. Weighted abduction with its differential costs on observables favors the first and disfavors the second.

Weighted conjuncts in the antecedents The third feature of weighted abduction is related to the weights of the conjuncts in the antecedents of the background axioms. Hobbs et al. (1993) say that the weights correspond to the "semantic contribution" each conjunct makes to its consequent and discuss the following example:

$\forall x(car(x) \wedge$ *no-top*$(x) \rightarrow convertible(x))$

Hobbs et al. (1993) assume that *car* contributes more to *convertible* than *no-top*, therefore the former should have a higher weight forcing its explanation. Thus, given a convertible mentioned in text, we will probably intend to link it to some other mentioning of a car rather than to a mentioning of an object with no top.

5 Graph Representation of Hypotheses

In this section, we introduce a formalization allowing us to estimate probabilities of abductive hypotheses in Section 6. We follow (Charniak and Shimony, 1990) and represent the set of all possible hypotheses as an AND/OR directed acyclic graph (AODAG).

Definition 1 *An AODAG is a 3-tuple* $< G, l, o >$*, where:*

1. *G is a directed acyclic graph, $G = (V, E)$.*
2. *l is a function from V to {AND, OR}, called the label. A node labeled AND is called an AND node, etc.*
3. *$o \subseteq V$ is a set of observed nodes.*

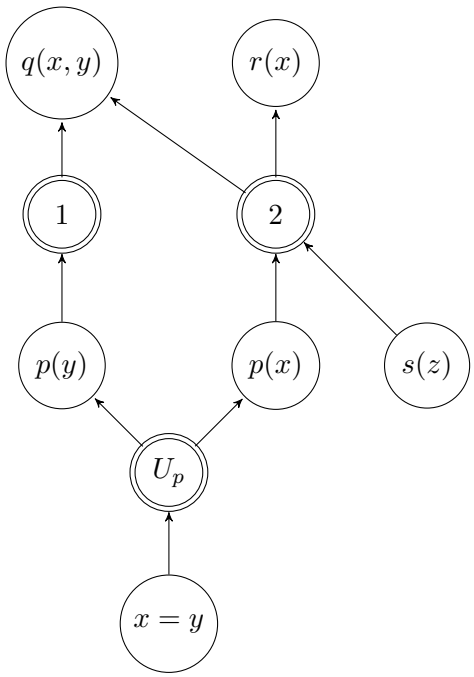

Figure 1: AODAG for the running example.

Consider an observation $O = \exists x, y(q(x, y) \wedge r(x))$ and the background knowledge B:

1) $\forall y(p(y) \rightarrow \exists x(q(x, y)))$
2) $\forall x, z(p(x) \wedge s(z) \rightarrow \exists y(q(x, y) \wedge r(x)))$

The AODAG in Fig. 1 is constructed by applying backchaining and unification to observation O. The nodes marked with a double circle represent inference operations: backchaining using Axioms 1 ("1" node) and 2 ("2" node) as well as unification ("U_p" node). Note that all operation nodes are AND nodes. All literal nodes are OR nodes. The notation $u \searrow v$ is used to say that u is an immediate parent of v. In our example, node "1" is a parent of $q(x, y)$ or $1 \searrow q(x, y)$.

Definition 2 *A truth assignment for an AODAG is a function f from V to $\{T, F\}$. A truth assignment is a model if the following conditions hold:*

1. *If $v \in o$ then $f(v) = T$.*
2. *If $v \notin o$ and v is an AND node then one of the following statements hold:*
   (a) *$f(v) = F$ and $\exists u \searrow v : f(u) = F$.*
   (b) *$f(v) = T$ and $\forall u \searrow v : f(u) = T$.*
3. *If $v \notin o$ and v is an OR node then one of the following statements hold:*
   (a) *$f(v) = T$ and $\exists u \searrow v : f(u) = T$.*

(b) $f(v) \in \{T, F\}$ *and* $\forall u \searrow v : f(u) = F$.

4. *If* $\exists v_1, .., v_n$ *such that for all* $i \in \{1, .., n\} : v_i$ *is* $x_i = x_{i+1}$ *and* $\exists v_0$ *equal to* $x_1 \neq x_{n+1}$ *then* $f(v_0) \wedge f(v_1) \wedge ... \wedge f(v_n) = F$.

Condition 1 in Definition 2 ensures that observables are true in every model. Condition 2 ensures that an operation node is true if the result of this operation is true. Otherwise, an operation node is false. Condition 3 ensures that a literal node is true if one of its explanations is true. Otherwise, it can be either true or false. We rely on the "open world" assumption, i.e., we do not assume that the knowledge base contains all possible facts about the world. Thus, assumptions can be made without explanations. Condition 4 rules out inconsistencies that result from an equality and an inequality of the same variables. It rules out truth assignments that assign T to both equality chains $x_1 = x_2 ... = x_{n+1}$ and an inequality $x_1 \neq x_{n+1}$.

It is easy to see that the set of hypotheses corresponds to the set of models of the AODAG. Given Definition 2, the truth assignment $M = \{(q(x, y), T), (r(x), T), (1, T), (p(y), T), (2, F), (p(x), F), (s(z), F), (U, F), (x = y, F)\}$ is a model of the example AODAG. It corresponds to the hypothesis $p(y) \wedge r(x)$. The nodes in a model that are assigned the truth value T and have no parents with the truth value T are called *assumptions* in this model. If $u \searrow v$ and both u and v are assigned the truth value T in a model, then u *explains* v in this model. For example, $r(x)$ is an assumption in the model M above, whereas $q(x, y)$ is explained by Axiom 1 in M.

6 Probabilities and Independence Assumptions

Now we are ready to estimate the likelihood of abductive hypotheses relevant for discourse interpretation. Let us associate a random variable from the set $\{X_1, ..., X_n\}$ with each of v nodes in an AODAG. The variables X_i ($i \in \{1, ..., n\}$) take values from the set $\{T, F\}$. If $f(v_i) = T$ then $X_i = T$; otherwise $X_i = F$. The joint probability distribution of the set $\{X_1, ..., X_n\}$ is as follows:

$$P(X_1, ..., X_n) = \prod_{i=1}^{n} P(X_i|\pi_i), \qquad (1)$$

where X_i is conditioned on π_i that denotes all other variables from the set $\{X_1, ..., X_n\}$ on which X_i depends. The question is how to define π_i for each X_i. In order to do it, we need to make independence assumptions.

As discussed in Section 4, the cost propagation mechanism in weighted abduction results in the following model preferences:

1. Other things being equal, a model that results from application of more reliable axioms is favored.

2. Other things being equal, a model that contains more true unification nodes is favored.

3. Other things being equal, a model that explains referential observables is favored.

Let us formulate independence assumptions reflecting the above model preferences. We can use the local Markov property: each variable is conditionally independent of its non-descendants given its immediate parent variables. But we also need a special account for unifications, because any true unification raises the likelihood of the corresponding model.

One option is to say that every axiom node in an AODAG also depends on its parent unification nodes. For example, nodes 1 and 2 in the example AODAG depend on the node U_p. However, given more observables there could be more unifications resulting from axiom applications. For example, if we add observable $s(t)$ then the application of Axiom 2 can result in one more unification ($t = z$). Given a set of golden AODAG models, one can compute all possible unifications resulting from a particular axiom. Alternatively, we can say that it does not matter unifications of which literals result from an axiom; the only thing that matters is how many unifications are there. In order to implement this second option, we introduce one more type of random variables associated with an AODAG: $numbU_v$ is associated with each axiom node v. It takes values from the set $\mathbb{N}$ and stands for the number of true unifications that are parents of v.

In order to account for referentiality, we introduce another type of random variables Ref_v associated with each literal node v in an AODAG. It takes values from the set $\{T, F\}$. If v is a referential observable or it has a referential observable as its child, then $Ref_v = T$; otherwise $Ref_v = F$. Each axiom application depends on whether its immediate children are referential or not.

We associate random variables X_{node_name} with each node of our example AODAG. In addition,

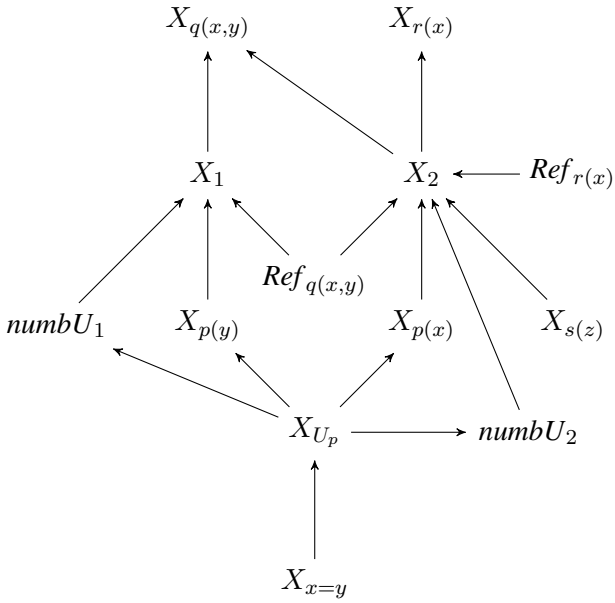

Figure 2: Bayesian network for the running example AODAG.

we introduce random variables $numbU_1$, $numbU_2$, $Ref_{q(x,y)}$ and $Ref_{r(x)}$. Fig. 2 shows the corresponding Bayesian network for the example AODAG that has the following joint probability distribution:

$$P(X_{x=y}) * P(X_{U_p}|X_{x=y}) * P(X_{p(y)}|X_{U_p}) * P(X_{p(x)}|X_{U_p}) * P(X_{s(z)}) * P(numbU_{p(y)}|X_{U_p}) * P(X_1|X_{p(y)}, numbU_{p(y)}, Ref_{q(x,y)}) * P(X_2|X_{p(x)}, X_{s(z)}, numbU_{p(x),s(z)}, Ref_{q(x,y)}, Ref_{r(x)}) * P(X_{q(x,y)}|X_1, X_2) * P(X_{r(x)}|X_2) * P(numbU_{p(x),s(z)}|X_{U_p}) * P(Ref_{q(x,y)}) * P(Ref_{r(x)})$$

Now we can estimate the probability of all abductive hypotheses or compute the best hypothesis a using standard method for computing Most Probable Explanation (Pearl, 1988) that maximizes the posterior probability of the joint assignment given the observations (values of variables $X_{q(x,y)}$, $X_{r(x)}$, $Ref_{q(x,y)}$, $Ref_{r(x)}$ in our example). If conditional probability tables need to be learned, we can use standard algorithms: Expectation Maximization (Dempster et al., 1977; Langseth and Bangsø, 2001; Ramoni and Sebastiani, 2001) and Markov Chain Monte Carlo methods (Liao and Ji, 2009).

7 Linking Costs and Weights in Weighted Abduction to Probabilities

Section 6 gives us a probabilistic approach to abduction that preserves the relevant discourse interpretation features of weighted abduction, so now we want to see what are the relationships between weights and probabilities across these two frameworks. Consider our running example again. Suppose $cost(q(x,y)) = c_1$, $cost(r(x)) = c_2$, weight of $p(y)$ in Axiom 1 is w_1, and weights of $p(x)$ and $s(z)$ in Axiom 2 are w_2 and w_3 correspondingly. There are 5 hypotheses for the given observation. According to the cost propagation scheme, the hypotheses are assigned the following costs.

$H_1 = q(x,y) \wedge r(x)$
$cost(H_1) = c_1 + c_2$
$H_2 = p(y) \wedge r(x)$
$cost(H_2) = w_1 * c_1 + c_2$
$H_3 = p(x) \wedge s(z)$
$cost(H_3) = w_2 * (c_1 + c_2) + w_3 * (c_1 + c_2)$
$H_4 = p(y) \wedge p(x) \wedge s(z)$
$cost(H_4) = w_1 * c_1 + w_2 * (c_1 + c_2) + w_3 * (c_1 + c_2)$
$H_5 = p(y) \wedge p(x) \wedge s(z) \wedge y = x$
$cost(H_5) = min(w_1 * c_1, w_2 * (c_1 + c_2)) + w_3 * (c_1 + c_2)$

The corresponding AODAG has 5 models:

$M_1 = \{(q(x,y),T), (r(x),T), (1,F), (p(y),F), (2,F), (p(x),F), (s(z),F), (U,F), (x=y,F)\}$
$M_2 = \{(q(x,y),T), (r(x),T), (1,T), (p(y),T), (2,F), (p(x),F), (s(z),F), (U,F), (x=y,F)\}$
$M_3 = \{(q(x,y),T), (r(x),T), (1,F), (p(y),F), (2,T), (p(x),T), (s(z),T), (U,F), (x=y,F)\}$
$M_4 = \{(q(x,y),T), (r(x),T), (1,T), (p(y),T), (2,T), (p(x),T), (s(z),T), (U,F), (x=y,F)\}$
$M_5 = \{(q(x,y),T), (r(x),T), (1,T), (p(y),T), (2,T), (p(x),T), (s(z),T), (U,T), (x=y,T)\}$

Our goal is to find function g such that

$$\forall i \in \{1,..,5\} : cost(H_i) = g(P(M_i)). \quad (2)$$

The hypothesis cost is a sum of the assumption costs (e.g., $cost(H_1) = c_1 + c_2$). Can we derive costs of atomic literals from the probabilities of these literals to be assumed? The smaller the cost, the bigger the probability that the literal is assumed. The event when no axioms are applied to the literal node v is denoted by $Assume(v)$. If we set g to the negative logarithm, then summing costs will be equal to multiplying probabilities:

$$cost(v) = -log(P(Assume(v))). \quad (3)$$

Model M_1 refers to the event when no axioms are applied: $Assume(q(x,y)) \cap Assume(r(x))$. Obviously, the events $Assume(q(x,y))$ and

$Assume(r(x))$ are not independent, because Axiom 2 is applicable to both $q(x, y)$ and $r(x)$. Therefore we get the following contradiction:

$$cost(H_1) = cost(q(x, y)) + cost(r(x)) = -log(P(Assume(q(x, y)) * P(Assume(r(x))) \neq -log(P(Assume(q(x, y) \cap Assume(r(x))) = -log(P(M_1)).$$

We cannot link the sum of costs of atomic literals to the product of the probabilities of these literals to be assumed, because the assumption events are not independent. Therefore we have to reject Eq. 3. Suppose we selected c_1 and c_2 so that

$$c_1 + c_2 = -log(P(Assume(q(x, y)) \cap Assume(r(x))) = -log(P(M_1)).$$

Can we then link axiom weights to probabilities? Model M_3 refers to the situation when only Axiom 2 is applied. It has the following probability[2]:

$$P(M_3) = P(X_1 = F \cap X_2 = T \cap X_{p(y)} = F \cap X_{p(x)} = T \cap X_{s(z)} = T \cap X_{U_p} = F | X_{q(x,y)} = T, X_{r(x)} = T).$$

Since $cost(H_3) = (w_2 + w_3) * (c_1 + c_2)$, we can try to link $w_2 + w_3$ to the probability of Axiom 2 to be applied. But in order to compute $P(M_3)$ the value of $cost(H_3)$ is also required to accommodate the probability of Axiom 1 not to be applied. Thus, instead of one axiom weight for each axiom α we need to have a table of conditional weights depending on all other axioms that can be applied in combination with α. This is not the case in weighted abduction.

The discussion above shows that we need conditional probabilities that cannot be linked to atomic literal costs and weights, because variables assigned to the atomic literal nodes are *not independent*. The question remains open if it is possible to tune weights and costs so that least cost hypotheses in weighted abduction correspond to the most pragmatically relevant (and the most probable) explanations. This is an empirical question and the answer to it depends on a particular application.

The fact that costs and weights in weighted abduction cannot be linked to probabilities does not make the framework inapplicable to discourse interpretation or any other task. One can see costs and weights as being parameters that need to be tuned in a practical setting. Inoue and Inui (2011) show that it is possible to represent weighted abduction as a linear constraint optimization problem and learn costs and weights in a large-margin learning procedure (Inoue et al., 2012b) including unification cost learning (Inoue et al., 2012a).

However, the problem remains how to set prior values for costs and weights before starting the learning. Furthermore, it is impossible to interpret learned values, which results in the choice of the best hypothesis being unpredictable.

[2]For simplicity, we ignore the referential variables.

8 Conclusion

Abduction allows us to model interpretation of discourse as the explanation of observables given knowledge about the world. In an abductive framework, many explanations can be constructed for the same observation. Therefore, an approach to estimating the likelihood of the alternative explanations is required.

In this paper, we showed that the cost propagation mechanism in weighted abduction has advantages over alternative approaches when applied to discourse interpretation. However, costs and weights in weighted abduction have no probabilistic interpretation, which makes their estimation and learning difficult. We proposed a formal framework for computing likelihood of abductive hypotheses with an account of variable inequalities and probabilistic unification. We discussed independence assumptions relevant for discourse processing. We showed that the cost propagation mechanism cannot be interpreted in terms of probabilities, but that features of weighted abduction relevant for discourse interpretation can be still captured in a probabilistic framework.

Future work concerns implementation of the probabilistic abductive framework proposed in Section 6 and its comparison with weighted abduction as tested on specific discourse processing tasks, such as recognizing textual entailment or coreference resolution; see (Ovchinnikova et al., 2011; Inoue et al., 2013) and (Inoue et al., 2012a) for applications of abduction to these tasks.

Acknowledgments

We thank Chris Wienberg for his valuable comments. This research was supported by ONR grant N00014-13-1-0286.

References

J. Blythe, J. R. Hobbs, P. Domingos, R. J. Kate, and R. J. Mooney. 2011. Implementing weighted abduction in markov logic. In *Proc. of IWCS'11*, pages 55–64, Oxford, England.

J. Bos. 2011. A survey of computational semantics: Representation, inference and knowledge in wide-coverage text understanding. *Language and Linguistics Compass*, 5(6):336–366.

E. Charniak and R. P. Goldman. 1989. A Semantics for Probabilistic Quantifier-Free First-Order Languages, with Particular Application to Story Understanding. In N. S. Sridharan, editor, *IJCAI'89*, pages 1074–1079. Morgan Kaufmann.

E. Charniak and S. E. Shimony. 1990. Probabilistic semantics for cost-based abduction. In *Proc. of the 8th National Conference on AI*, pages 106–111.

A. P. Dempster, N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, 39:1–38.

J. R. Hobbs, M. Stickel, D. Appelt, and P. Martin. 1993. Interpretation as abduction. *Artificial Intelligence*, 63:69–142.

L. Huang. 2008. Forest reranking: Discriminative parsing with non-local features. In *Proc. of ACL'08*, pages 586–594.

N. Inoue and K. Inui. 2011. ILP-Based Reasoning for Weighted Abduction. In *Proc. of AAAI Workshop on Plan, Activity and Intent Recognition*.

N. Inoue, E. Ovchinnikova, K. Inui, and J. R. Hobbs. 2012a. Coreference Resolution with ILP-based Weighted Abduction. In *Proc. of COLING'12*, pages 1291–1308.

N. Inoue, K. Yamamoto, Y. Watanabe, N. Okazaki, and K. Inui. 2012b. Online large-margin weight learning for first-order logic-based abduction. In *Proc. of the 15th Information-Based Induction Sciences Workshop*, pages 143–150.

N. Inoue, E. Ovchinnikova, K. Inui, and J. R. Hobbs. 2013. Weighted abduction for discourse processing based on integer linear programming. In *Plan, Activity, and Intent Recognition*.

R.J. Kate and R. J. Mooney. 2009. Probabilistic abduction using markov logic networks. In *Proc. of PAIR'09*, Pasadena, CA.

H. Langseth and O. Bangsø. 2001. Parameter learning in object-oriented bayesian networks. *Ann. Math. Artif. Intell.*, 32(1-4):221–243.

W. Liao and Q. Ji. 2009. Learning bayesian network parameters under incomplete data with domain knowledge. *Pattern Recognition*, 42(11):3046–3056.

D. McClosky, E. Charniak, and M. Johnson. 2006. Effective self-training for parsing. In *Proc. of HLT-NAACL'06*.

E. Ovchinnikova, N. Montazeri, T. Alexandrov, J. R Hobbs, M. McCord, and R. Mulkar-Mehta. 2011. Abductive Reasoning with a Large Knowledge Base for Discourse Processing. In *Proc. of IWCS'11*, pages 225–234, Oxford, UK.

E. Ovchinnikova. 2012. *Integration of World Knowledge for Natural Language Understanding*. Atlantis Press, Springer.

J. Pearl. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.

S. Raghavan and R. Mooney. 2010. Bayesian abductive logic programs. In *Proc. of Star-AI'10*, pages 82–87, Atlanta, GA.

M. Ramoni and P. Sebastiani. 2001. Robust learning with missing data. *Machine Learning*, 45(2):147–170.

M. Richardson and P. Domingos. 2006. Markov logic networks. *Machine Learning*, 62(1-2):107–136.