Abduction in Natural Language Understanding

Jerry R. Hobbs
Artificial Intelligence Center
SRI International

1 Language and Knowledge

We are able to understand language so well because we know so much. When we read the sentence

(1) John drove down the street in a car.

we know immediately that the driving and hence John are in the car and that the street isn't. We attach the prepositional phrase to the verb "drove" rather than to the noun "street". This is not syntactic knowledge, because in the syntactically similar sentence

(2) John drove down a street in Chicago.

it is the street that is in Chicago.

Therefore, a large part of the study of language should be an investigation of the question of how we use our knowledge of the world to understand discourse. This question has been examined primarily by researchers in the field of artificial intelligence (AI), in part because they have been interested in linking language with actual behavior in specific situations, which has led them to an attempt to represent and reason about fairly complex world knowledge.

In this chapter I describe how a particular kind of reasoning, called ABDUCTION, provides a framework for addressing a broad range of problems that are posed in discourse and that require world knowledge in their solutions. I first defend first-order logic as a mode of representation for the information conveyed by sentences and the knowledge we bring to the discourses we interpret, but with one caveat: Reasoning must be defeasible. I discuss several ways that defeasible inference has been formalized in AI, and introduce abduction as one of those methods. Then in successive sections I show

- how various problems in LOCAL PRAGMATICS, such as reference resolution, metonymy, interpreting compound nominals, and word sense disambiguation can be solved via abduction;
- how this processing can be embedded in a process for recognizing the structure of discourse; and
- how these can all be integrated with the recognition of the speaker's plan.

I close with a discussion of the relation of this framework to Relevance Theory and of some of the principal outstanding research issues.

2 Logic as the Language of Thought

A very large body of work in AI begins with the assumptions that information and knowledge should be represented in first-order logic and that reasoning is theorem-proving. On the face of it, this seems implausible as a model for people. It certainly doesn't seem as if we are using logic when we are thinking, and if we are, why are so many of our thoughts and actions so illogical? In fact, there are psychological experiments that purport to show that people do not use logic in thinking about a problem (e.g., Wason and Johnson-Laird 1972).

I believe that the claim that logic is the language of thought comes to less than one might think, however, and that thus it is more controversial than it ought to be. It is the claim that a broad range of cognitive processes are amenable to a high-level description in which six key features are present. The first three of these features characterize propositional logic and the next two first-order logic. I will express them in terms of "concepts", but one can just as easily substitute propositions, neural elements, or a number of other terms.

- Conjunction: There is an additive effect ($P \wedge Q$) of two distinct concepts (P and Q) being activated at the same time.

- Modus Ponens: The activation of one concept (P) triggers the activation of another concept (Q) because of the existence of some structural relation between them ($P \supset Q$).

- Recognition of Obvious Contradictions: The recognition of contradictions in general is undecidable, but we have no trouble with the easy ones, for example, that cats aren't dogs.

- Predicate-Argument Relations: Concepts can be related to other concepts in several different ways. For example, we can distinguish between a dog biting a man ($bite(D, M)$) and a man biting a dog ($bite(M, D)$).

- Universal Instantiation (or Variable Binding): We can keep separate our knowledge of general (universal) principles ("All men are mortal") and our knowledge of their instatiations for particular individuals ("Socrates is a man" and "Socrates is mortal").

Any plausible proposal for a language of thought must have at least these features, and once you have these features you have first-order logic.

Note that in this list there are no complex rules for double negations or for contrapositives (if P implies Q then not Q implies not P). In fact, most of the psychological experiments purporting to show that people don't use logic really show that they don't use the contrapositive rule or that they don't handle double negations well. If the tasks in those experiments were recast into problems involving the use of modus ponens, no one would think to do the experiments because it is obvious that people would have no trouble with the task.

As an aside, let me mention that many researchers in linguistics and in knowledge representation make use of **higher-order** logic. It is straightforward, through various kinds of reification, to recast these logics into first-order logic, and in view of the resulting simplification in characterizing the reasoning process, there are very good reasons to do so (Hobbs 1985a).

There is one further property we need of the logic if we are to use it for representing and reasoning about commonsense world knowledge—defeasibility or nonmonotonicity.

3 Nonmonotonic Logic

The logic of mathematics is monotonic, in that once we know the truth value of a statement, nothing else we learn can change it. Virtually all commonsense knowledge beyond mathematics is uncertain or defeasible. Whatever general principles we have are usually only true most of the time or true with high probability or true unless we discover evidence to the contrary. It is almost always possible that we may have to change what we believed to be the truth value of a statement upon gaining more information. Almost all commonsense knowledge should be tagged with "insofar as I have been able to determine with my limited access to the facts and my limited resources for reasoning." The logic of commonsense knowledge must be nonmonotonic.

The development of nonmonotonic logics has been a major focus in AI research (Ginsberg 1987). One early attempt involved "negation as failure" (Hewitt 1972); we assume that not P is true if we fail to prove that P. Another early nonmonotonic logic (McDermott and Doyle 1980) had rules of the form "If P is true and Q is consistent with everything else we know, then take Q to be true."

Probably the most thoroughly investigated nonmonotonic logic was that developed by McCarthy (1980). He introduced ABNORMALITY CONDITIONS which the reasoner then minimized. For example, the general fact that birds fly is expressed

(3) $(\forall x) bird(x) \wedge \neg ab_1(x) \supset fly(x)$

That is, if x is a bird and not abnormal in a way specific to this rule, then x flies. Further axioms might spell out the exceptions:

(4) $(\forall x) penguin(x) \supset ab_1(x)$

That is, penguins are abnormal in the way specific to the "birds fly" rule.

Then to draw conclusions we minimize, in some fashion, those things we take to be abnormal. If all we know about Tweety is that he is a bird, then we assume he is not abnormal, and thus we conclude he can fly. If we subsequently learn that Tweety is a penguin, we retract the assumption that he is not abnormal in that way.

A problem arises with this approach when we have many axioms with different abnormality conditions. There may be many ways to minimize the abnormalities, each leading to different conclusions. This is illustrated by an example that is known as the NIXON DIAMOND (Reiter and Criscuolo 1981). Suppose we know that generally Quakers are pacifists. We can write this as

(5) $(\forall\, x)Quaker(x) \wedge \neg ab_2(x) \supset pacifist(x)$

Suppose we also know that Republicans are generally not pacifists.

(6) $(\forall\, x)Republican(x) \wedge \neg ab_3(x) \supset \neg pacifist(x)$

Then what do we conclude when we learn that Nixon is both a Quaker and a Republican? Assuming both abnormality conditions results in a contradiction. If we take ab_2 to be false, we conclude Nixon is a pacifist. If we take ab_3 to be false, we conclude Nixon is not a pacifist. How do we choose between the two possibilities? Researchers have made various suggestions for how to think about this problem (e.g., Shoham 1987). In general, some scheme is needed for choosing among the possible combinations of assumptions.

In recent years there has been considerable interest in AI in the reasoning process known as abduction, or inference to the best explanation. As it is normally conceived in AI, it can be viewed as one variety of nonmonotonic logic.

4 Abduction

The simplest way to explain abduction is by comparing it with two words it rhymes with—deduction and induction. In deduction, from P and $P \supset Q$, we conclude Q. In induction, from P and Q, or more likely a number of instances of P and Q together with other considerations, we conclude $P \supset Q$. Abduction is the third possibility. From an observable Q and a general principle $P \supset Q$, we conclude that P must be the underlying reason that Q is true. We assume P because it explains Q.

Of course, there may be many such possible P's, some contradictory with others, and therefore any method of abduction must include a method for evaluating and choosing among alternatives. At a first cut, suppose in trying to explain Q we know $P \wedge R \supset Q$ and we know R. Then R provides partial evidence that Q is true, making the assumption of P more reasonable. In addition, if we are seeking to explain two things, Q_1 and Q_2, then it is reasonable to favor assuming a P that explains both of them rather than a different explanation for each.

The conclusions we draw in this way are only assumptions and may have to be retracted later if we acquire new, contradictory information. That is, this method of reasoning is nonmonotonic.

Abduction has a history. Prior to the late seventeenth century science was viewed as deductive, at least in the ideal. It was felt that, on the model of Euclidean geometry, one should begin with propositions that were self-evident and deduce whatever consequences one could from them. The modern view of scientific theories, probably best expressed by Lakatos (1970), is quite different. One tries to construct abstract theories from which observable events can be deduced or predicted. There is no need for the abstract theories to be self-evident, and they usually are not. It is only necessary for them to predict as broad a range as possible of the observable data and for them to be "elegant", whatever that means. Thus, the modern view is that science is fundamentally abductive. We seek hidden principles or causes from which we can deduce the observable evidence.

This view of science, and hence the notion of abduction, can be seen first, insofar as I am aware, in some passages in Newton's *Principia* (1934 [1686]). At the end of *Principia*, in a justification for not seeking the cause of gravity, he says, "And to us it is enough that gravity does really exist, and act according to the laws which we have explained, and abundantly serves to account for all the motions of the celestial bodies, and of our sea." (Newton 1934:547) The justification for gravity (P) and its laws ($P \supset Q$) is not in their self-evidential nature but in what they account for (Q).

In the eighteenth century, the German philosopher Christian Wolff (1963 [1728]) shows, to my knowledge, the earliest **explicit** awareness of the importance of abductive reasoning. He presents almost the standard Euclidean account of certain knowledge, but with an important provision in his recognition of the inevitability and importance of hypotheses:

> Philosophy must use hypotheses insofar as they pave the way to the discovery of certain truth. For in a philosophical hypothesis certain things which are not firmly established are assumed because they provide a reason for things which are observed to occur. Now if we can also deduce other things which are not observed to occur, then we have the opportunity to either observe or experimentally detect things which otherwise we might not have noticed. In this way we become more certain as to whether or not anything contrary to experience follows from the hypothesis. If we deduce things which are contrary to experience, then the hypothesis is false. If the deductions agree with experience, then the probability of the hypothesis is increased. And thus the way is paved for the discovery of certain truth. (Wolff 1963:67)

He also recognizes the principle of parsimony: "If one cannot necessarily deduce from a hypothesis the things for which it is assumed, then the hypothesis is spurious." (Wolff 1963:68) However, he views hypotheses as only provisional, awaiting deductive proof.

The term "abduction" was first used by C. S. Pierce (e.g., 1955). His definition of it is as follows:

(7) The surprising fact, Q, is observed;
But if P were true, Q would be a matter of course,
Hence, there is reason to suspect that P is true. (Pierce 1955:151)

(He actually used A and C for P and Q.) Pierce says that "in pure abduction, it can never be justifiable to accept the hypothesis otherwise than as an interrogation", and that "the whole question of what one out of a number of possible hypotheses ought to be entertained becomes purely a question of economy." That is, there must be an evaluation scheme for choosing among possible abductive inferences.

The earliest formulation of abduction in artificial intelligence was by Morgan (1971). He showed how a complete set of truth-preserving rules for generating theorems could be turned into a complete set of falsehood-preserving rules for generating hypotheses.

The first use of abduction in an AI application was by Pople (1973), in the context of medical diagnosis. He gave the formulation of abduction sketched above and showed how it can be implemented in a theorem-proving framework. Literals (or propositions) that are "abandoned by deduction in the sense that they fail to have successor nodes"

(Pople 1973:150) are taken as the candidate hypotheses. That is, one tries to prove the symptoms and signs exhibited and the parts of a potential proof that cannot be proven are the candidate hypotheses. Those hypotheses are best that account for the most data, and in service of this principle, he introduced factoring or synthesis, which attempts to unify goal literals. Hypotheses where this is used are favored. That is, that explanation is best that minimizes the number of causes.

Work on abduction in artificial intelligence was revived in the 1980s at several sites. Reggia and his colleagues (e.g., Reggia et al., 1983; Reggia 1985) formulated abductive inference in terms of parsimonious covering theory. Charniak and McDermott (1985) presented the basic pattern of abduction and then discussed many of the issues involved in trying to decide among alternative hypotheses on probabilistic grounds. Cox and Pietrzykowski (1986) present a formulation in a theorem-proving framework that is very similar to Pople's, though apparently independent. It is especially valuable in that it considers abduction abstractly, as a mechanism with a variety of possible applications, and not just as a handmaiden to diagnosis.

Josephson and Josephson (1994) provide a comprehensive treatment of abduction, its philosophical background, its computational properties, and its utilization in AI applications.

I have indicated that the practice of science is fundamentally abductive. The extension of abduction to ordinary cognitive tasks is very much in line with the popular view in cognitive science that people going about in the world trying to understand it are scientists in the small. This view can be extended to natural language understanding—interpreting discourse is coming up with the best explanation for what is said.

The first appeal to something like abduction that I am aware of in natural language understanding was by Grice (1967, 1989), when he introduced the notion of CONVERSATIONAL IMPLICATURE to handle examples like the following:

(8) A: How is John doing on his new job at the bank?
B: Quite well. He likes his colleagues and he hasn't embezzled any money yet.

Grice argues that in order to see this as coherent, we must assume, or draw as a conversational implicature, that both A and B know that John is dishonest. Although he does not say so, an implicature can be viewed as an abductive move for the sake of achieving the best interpretation.

Lewis (1979) introduces the notion of ACCOMMODATION in conversation to explain the phenomenon that occurs when you "say something that requires a missing presupposition, and straightaway that presupposition springs into existence, making what you said acceptable after all." The hearer accommodates the speaker.

Thomason (1990) argued that Grice's conversational implicatures are based on Lewis's rule of accommodation. We might say that implicature is a procedural characterization of something that, at the functional or interactional level, appears as accommodation. Implicature is the way we do accommodation.

In the middle 1980s researchers at several sites began to apply abduction to natural language understanding (Norvig 1983, 1987; Wilensky 1983; Wilensky et al. 1988; Charniak and Goldman 1988, 1989; Hobbs et al. 1988; Hobbs et al. 1993). At least in the last case the recognition that implicature was a use of abduction was a key observation in the development of the framework.

Norvig, Wilensky, and their associates proposed an operation called CONCRETION, one of many that take place in the processing of a text. It is a "kind of inference in which a more specific interpretation of an utterance is made than can be sustained on a strictly logical basis" (Wilensky et al. 1988:50). Thus, "to use a pencil" generally means to write with a pencil, even though one could use a pencil for many other purposes.

Charniak and his associates also developed an abductive approach to interpretation. Charniak (1986) expressed the fundamental insight: "A standard platitude is that understanding something is relating it to what one already knows. ... One extreme example would be to prove that what one is told must be true on the basis of what one already knows. ... We want to prove what one is told *given certain assumptions*." (Charniak 1986:585)

Charniak and Goldman developed an interpretation procedure that incrementally built a belief network (Pearl 1988), where the links between the nodes, representing influences between events, were determined from axioms expressing world knowledge. They felt that one could make not unreasonable estimates of the required probabilities, giving a principled semantics to the numbers. The networks were then evaluated and ambiguities were resolved by looking for the highest resultant probabilities.

Stickel invented a method called WEIGHTED ABDUCTION (Stickel 1988; Hobbs et al. 1993) that builds the evaluation criteria into the proof process. Briefly, propositions to be proved are given an assumption cost—what you will have to pay to assume them. When we backchain over a rule of the form $P \supset Q$, the cost is passed back from Q to P, according to a weight associated with P. Generally, P will cost more to assume than Q, so that short proofs are favored over long ones. But if partial evidence is found, for example, if $P \wedge R \supset Q$ and we can prove R, then it will cost less to assume P than to assume Q, and we get a more specific interpretation. In addition, if we need to prove Q_1 and Q_2 and P implies both, then it will cost less to assume P than to assume Q_1 and Q_2. This feature of the method allows us to exploit the implicit redundancy inherent in natural language discourse.

Weighted abduction suggests a simple way to incorporate the uncertainty of knowledge into the axioms expressing the knowledge. Propositions can be assumed at a cost. Therefore, we can have propositions whose only role is to be assumed and to levy a cost. For example, let's return to the rule that birds fly. We can express it with the axiom

(9) $(\forall x)[bird(x) \wedge etc_1(x) \supset fly(x)]$

That is, if x is a bird and some other unspecified conditions hold for x ($etc_1(x)$), then x flies. The predicate etc_1 encodes the unspecified conditions. There will never be a way to prove it; it can only be assumed at cost. The cost of etc_1 will depend inversely on the certainty of the rule that birds fly. It will cost to use this rule, but the lowest-cost proof of everything we are trying to explain may nevertheless involve this rule and hence the inference that birds fly. We know that penguins don't fly:

(10) $(\forall x)[penguin(x) \supset \neg fly(x)]$

If we know Tweety is a penguin, we know he doesn't fly. Thus, to assume etc_1 is true of Tweety would lead to a contradiction, so we don't. The relation between the *etc* predicates and the abnormality predicates of McCarthy's nonmonotonic logic is obvious: etc_1 is just $\neg ab_1$.

The framework of "Interpretation as Abduction" (IA) (Hobbs et al. 1993) follows directly from this method of abductive inference, and it is the IA framework that is presented in the remainder of this chapter. Whereas in Norvig and Wilensky's work, abduction or concretion was one process among many involved in natural language understanding, in the IA framework abduction is the whole story. Whereas in Charniak and Goldman's work, specific procedures involving abduction are implemented to solve specific interpretation problems, in the IA framework there is only one procedure—abduction—that is used to explain or prove the logical form of the text, and the solutions to specific interpretation problems fall out as byproducts of this process.

It should be pointed out that in addition to what is presented below there have been a number of other researchers who have used abduction for various natural language understanding problems, including Nagao (1989) for resolving syntactic ambiguity, Dasigi (1988) for resolving lexical ambiguity, Rayner (1993) for asking questions of a database, Ng and Mooney (1990) and Lascarides and Oberlander (1992) for recognizing discourse structure, McRoy and Hirst (1991) for making repairs in presupposition errors, Appelt and Pollack (1990) for recognizing the speaker's plan, and Harabagiu and Moldovan (1998) for general text understanding using WordNet as a knowledge base.

5 Interpretation as Abduction

In the IA framework we can describe very concisely what it is to interpret a sentence:

(11) Prove the logical form of the sentence,
together with the selectional constraints that predicates impose on
their arguments,
allowing for coercions,
Merging redundancies where possible,
Making assumptions where necessary.

By the first line we mean "prove, or derive in the logical sense, from the predicate calculus axioms in the knowledge base, the logical form that has been produced by syntactic analysis and semantic translation of the sentence."

In a discourse situation, the speaker and hearer both have their sets of private beliefs, and there is a large overlapping set of mutual beliefs. An utterance lives on the boundary between mutual belief and the speaker's private beliefs. It is a bid to extend the area of mutual belief to include some private beliefs of the speaker's. It is anchored referentially in mutual belief, and when we succeed in proving the logical form and the constraints, we are recognizing this referential anchor. This is the given information, the definite, the presupposed. Where it is necessary to make assumptions, the information comes from the speaker's private beliefs, and hence is the new information, the indefinite, the asserted. Merging redundancies is a way of getting a minimal, and hence a best, interpretation.

Merging redundancies and minimizing the assumptions result naturally from the method of weighted abduction.

6 Abduction and Local Pragmatics

Local pragmatics encompasses those problems that are posed within the scope of individual sentences, even though their solution will generally require greater context and world knowledge. Included under this label are the resolution of coreference, resolving syntactic and lexical ambiguity, interpreting metonymy and metaphor, and finding specific meanings for vague predicates such as in the compound nominal.

Consider a simple example that contains three of these problems.

(12) The Boston office called.

This sentence poses the problems of resolving the reference of "the Boston office", expanding the metonymy to "[Some person at] the Boston office called", and determining the implicit relation between Boston and the office. Let us put these problems aside for the moment, however, and interpret the sentence according to the IA characterization. We must prove abductively the logical form of the sentence together with the constraint "call" imposes on its agent, allowing for a coercion. That is, we must prove abductively that there is a calling event by a person who may or may not be the same as the explicit subject of the sentence, but it is at least related to it, or coercible from it, and that there is an office bearing some unspecified relation to Boston.

The sentence can be interpreted with respect to a knowledge base of mutual knowledge that contains the following facts and rules, expressed as axioms:

There is a city of Boston.
There is an office in Boston.
John is a person who works for the office.
The "in" relation can be represented by a compound nominal.
An organization can be coerced into a person who works for it.

Given these rules, the proof of all of the logical form is straightforward except for the existence of the calling event. Hence, we assume that; it is the new information conveyed by the sentence.

Now notice that the three local pragmatics problems have been solved as a by-product. We have resolved "the Boston office" to the specific office we know about. We have determined the implicit relation in the compound nominal to be "in". And we have expanded the metonymy to "John, who works for the Boston office, called."

For an illustration of the resolution of lexical ambiguity, consider an example from Hirst (1987):

(13) The plane taxied to the terminal.

The words “plane”, “taxied”, and “terminal” are all ambiguous.

Suppose the knowledge base consists of axioms with the following content:

> An airplane is a plane.
> A wood smoother is a plane.
> For an airplane to move on the ground is for it to taxi.
> For a person to ride in a cab is for him or her to taxi.
> An airport terminal is a terminal.
> A computer terminal is a terminal.
> An airport has airplanes and an airport terminal.

To prove the logical form of the sentence, we need to prove abductively the existence of a plane, a terminal and a taxi-ing event. The minimal proof of will involve assuming the existence of an airport, deriving from that an airplane, and thus the plane, and an airport terminal, and thus the terminal, assuming that a plane is moving on the ground, and recognizing the identity of the airplane at the airport with the one in that reading of “taxi”.

Another possible interpretation would be one in which we assumed that a wood smoother, a ride in a cab, and a computer terminal all existed. It is because weighted abduction favors merging redundancies that the correct interpretation is the one chosen. That interpretation allows us to minimize the assumptions we make.

7 Recognizing Discourse Structure

Syntax can be incorporated into this framework (Hobbs 1998) by encoding the rules of Pollard and Sag’s (1994) Head-driven Phrase Structure Grammar in axioms. The axioms involve predications asserting that strings of words describe entities or situations. Parsing a sentence is then proving that there is a situation that the sentence describes. This proof bottoms out in the logical form of the sentence, and proving this is the process of interpretation described in the previous section. We have recast the process of interpreting a sentence from the problem of proving the logical form into the problem of proving the string of words is a grammatical, interpretable sentence, where “interpretable” means we can prove the logical form.

When two segments of discourse are adjacent, that very adjacency conveys information. Each segment, insofar as it is coherent, conveys information about a situation or eventuality, and the adjacency of the segments conveys the suggestion that the two situations are related in some fashion, or are parts of larger units that are related. Part of what it is to understand a discourse is to discover what that relation is.

Overwhelmingly, the relations that obtain between discourse segments are based on causal, similarity, or figure-ground relations between the situations they convey. We can thus define a number of COHERENCE RELATIONS in terms of the relations between the situations. This will not be explored further here, but it is described in greater detail in Kehler (this volume). Here it will be indicated how this aspect of discourse structure can be built into the abduction framework.

The two rules defining coherent discourse structure are as follows:

A grammatical, interpretable sentence is a coherent segment of discourse.

If two coherent segments of discourse are concatenated and there is a coherence relation between the situations they describe, then the concatenation is a coherent segment of discourse, and the situation it describes is determined by the coherence relation.

That is, when we combine two coherent segments of discourse with a coherence relation we get a coherent segment of discourse. By applying this successively to a stretch of discourse, we get a tree-like structure for the whole discourse. Different structures result from different choices in ordering the concatenation operations.

Now interpreting a text is a matter of proving that it is a coherent segment of discourse conveying some situation.

Consider an example. Explanation is a coherence relation, and a first approximation of a definition of the Explanation relation would be that the eventuality described by the second segment causes the eventuality described by the first. That is, if what is described by the second segment could cause what is described by the first segment, then there is a coherence relation between the segments.

Consider a variation on a classic example of pronoun resolution difficulties from Winograd (1972):

(14) The police prohibited the women from demonstrating.
They feared violence.

How do we know "they" in the second sentence refers to the police and not to the women?

As in Section 6, we will not attack the coreference problem directly, but we will proceed to interpret the text by abduction. To interpret the text is to prove abductively that the string of words comprising the whole text is a coherent segment of discourse describing some situation. This involves proving that each sentence is a segment, by proving they are grammatical, interpretable sentences, and proving there is a coherence relation between them. The proof that they are sentences would bottom out in the logical forms of the sentences, thus requiring us to prove abductively those logical forms.

One way to prove there is a coherence relation between the sentences is to prove there is an Explanation relation between them by showing there is a causal relation between the eventualities they describe.

Thus, we must prove abductively the existence of the police, their prohibition of the demonstrating by the women, the fearing by someone of violence, and a causal relation between the fearing and the prohibition.

Suppose, plausibly enough, we have in our knowledge base axioms with the following content:

If you fear something, that will cause you not to want it.
Demonstrations cause violence.
If you don't want the effect, that will cause you not to want the cause.
If those in authority don't want something, that will cause them to prohibit it.
The police are in authority.
Causality is transitive.

From such axioms, we can prove all of the logical form of the text except the existence of the police, the demonstrating, and the fearing, which we assume. In the course of doing the proof, we unify the people doing the fearing with the police, thus resolving the problematic pronoun reference that originally motivated this example. "They" refers to the police.

One can imagine a number of variations on this example. If we had not included the axiom that demonstrations cause violence, we would have had to assume the violence and the causal relation between demonstrations and violence. Moreover, other coherence relations might be imagined here by constructing the surrounding context in the right way. It could be followed by the sentence "But since they had never demonstrated before, they did not know that violence might result." In this case, the second sentence would play a subordinate role to the third, forcing the resolution of "they" to the women. Each example, of course, has to be analyzed on its own, and changing the example changes the analysis. In Winograd's original version of this example,

(15) The police prohibited the women from demonstrating, because they feared violence.

the causality was explicit, thus eliminating the coherence relation as a source of ambiguity. The causal relation would be part of the logical form.

Winograd's contrasting text, in which "they" is resolved to the women, is

(16) The police prohibited the women from demonstrating, because they advocated violence.

Here we would need the facts that when one demonstrates one advocates and that advocating something tends to bring it about. Then showing a causal relation between the clauses will result in "they" being identified with the demonstrators.

8 Recognizing the Speaker's Plan

As presented so far, understanding discourse is seeing the world of the text as coherent, which in turn involves viewing the content of the text as observables to be explained. The focus has been on the information conveyed explicitly or implicitly by the discourse. We can call this the INFORMATIONAL account of a discourse.

But utterances are embedded in the world as well. They are produced to realize a speaker's intention, or more generally, they are actions in the execution of a speaker's plan to achieve some goal. The description of how a discourse realizes the speakers' goals may be called the INTENTIONAL account of the discourse.

Consider the intentional account from the broadest perspective. An intelligent agent is embedded in the world and must, at each instant, understand the current situation. The agent does so by finding an explanation for what is perceived. Put differently, the agent must explain why the complete set of observables encountered constitutes a coherent situation. Other agents in the environment are viewed as intentional, that is, as planning mechanisms, and this means that the best explanation of their observable actions is most likely to be that the actions are steps in a coherent plan. Thus, making sense of an environment that includes other agents entails making sense of the other agents' actions in terms of what they are intended to achieve. When those actions are utterances, the utterances must be understood as actions in a plan the agents are trying to effect. That is, the speaker's plan must be recognized—the intentional account.

Generally, when a speaker says something it is with the goal of the hearer believing the content of the utterance, or thinking about it, or considering it, or taking some other cognitive stance toward it. Let us subsume all these mental terms under the term "cognize". Then we can summarize the relation between the intentional and informational accounts succinctly in the following formula:

(17) $\textbf{intentional-account} = goal(A, cognize(B, \textbf{informational-account})$

The speaker ostensibly has the goal of changing the mental state of the hearer to include some mental stance toward the content characterized by the informational account. Thus, the informational account is embedded in the intentional account. When we reason about the speaker's intention, we are reasoning about how this goal fits into the larger picture of the speaker's ongoing plan. We are asking why the speaker seems to be trying to get the hearer to believe this particular content. The informational account explains the situation described in the discourse; the intentional account explains why the speaker chose to convey this information.

Both the intentional and informational accounts are necessary. The informational account is needed because we have no direct access to the speaker's plan. We can only infer it from history and behavior. The content of the utterance is often the best evidence of the speaker's intention, and often the intention is no more than to convey that particular content. On the other hand, the intentional account is necessary in cases like pragmatic ellipsis, where the informational account is highly underdetermined and the global interpretation is primarily shaped by our beliefs about the speaker's plan.

Perhaps most interesting are cases of genuine conflict between the two accounts. The informational account does not seem to be true, or it seems to run counter to the speaker's goals for the hearer to come to believe it, or it ought to be obvious that the hearer already does believe it. Tautologies are an example of the last of these cases—tautologies such as "boys will be boys," "fair is fair," and "a job is a job." Norvig and Wilensky (1990) cite this figure of speech as something that should cause trouble for an abduction approach that seeks minimal explanations, since the minimal explanation is that they just express a known truth. Such an explanation requires no assumptions at all.

In fact, the phenomenon is a good example of why an informational account of discourse interpretation has to be embedded in an intentional account. Let us imagine two parents, A and B, sitting in the playground and talking.

(18) A: Your Johnny is certainly acting up today, isn't he?
B: Boys will be boys.

In order to avoid dealing with the complications of plurals and tense in this example, let us simplify B's utterance to

(19) B: A boy is a boy.

Several informational accounts of this utterance are possible. The first is the Literal Extensional Interpretation. The first "a boy" introduces a specific, previously unidentified boy and the second says about him that he is a boy. The second informational account is the Literal Intensional Interpretation. The sentence expresses a trivial implicative relation between two general propositions—$boy(x)$ and $boy(x)$. The third is the Desired Interpretation. The first "a boy" identifies the typical member of a class which Johnny is a member of and the second conveys a general property, "being a boy", as a way of conveying a specific property, "misbehaving", which is true of members of that class.

Considering the informational account alone, the Literal Extensional Interpretation is minimal and hence would be favored. The Desired Interpretation is the worst of the three.

But the Literal Extensional and Intensional Interpretations leave the **fact** that the utterance **occurred** unaccounted for. In the intentional account, this is what we need to explain. The explanation would run something like this:

B wants A to believe that B is not responsible for Johnny's misbehaving.

Thus, B wants A to believe that Johnny misbehaves necessarily.

Thus, given that Johnny is necessarily a boy, B wants A to believe that Johnny's being a boy implies that he misbehaves.

Thus, B wants to convey to A that being a boy implies misbehaving.

Thus, given that boy-ness implies misbehaving is a possible interpretation of a boy being a boy, B wants to say to A that a boy is a boy.

The contents of the utterance under the Literal Extensional and Intensional Interpretations do not lend themselves to explanations for the fact that the utterance occurred, whereas the Desired Interpretation does. The requirement for the **globally** minimal explanation in an intentional account, that is, the requirement that both the content and the fact of the utterance must be explained, forces us into an interpretation of the content that would not be favored in an informational account alone. We are forced into an interpretation of the content that, while not optimal locally, contributes to a global interpretation that **is** optimal.

9 Relation to Relevance Theory

One of the other principal contenders for a theory of how we understand extended discourse is Relevance Theory (RT) (Sperber and Wilson 1986). In fact, the IA framework and RT are very close to each other in the processing that would implement them.

In RT, the agent is in the situation of having a knowlege base K and hearing a sentence with content Q. From K and Q a new set R of inferences can be drawn:

(20) $K, Q \vdash R$

RT says that the agent strives to **maximize** R in an appropriately hedged sense. An immediate consequence of this is that insofar as we are able to pragmatically strengthen Q by means of axioms of the form

(21) $P \supset Q$

then we are getting a better R, since P implies anything that Q implies, and then some. In the IA framework, we begin with pragmatic strengthening. The task of the agent is to explain the general Q with the more specific P.

This means that anything done in the IA framework ought to carry over without change into RT. Much of the work in RT depends primarily or solely on pragmatic strengthening, and where this is the case, it can immediately be incorporated into the IA framework.

From the point of view of IA, people are going through the world trying to figure out what is going on. From the point of view of RT, they are going through the world trying to learn as much as they can, and figuring out what is going on is in service of that.

The IA framework has been worked out in greater detail formally and, I believe, has a more compelling justification—explaining the observables in our environment. But a great deal of excellent work has been done in RT, so it is useful to know that the two frameworks are almost entirely compatible.

10 Research Issues

In the examples given in this paper, I have cavalierly assumed the most convenient axioms were in the knowledge base that was being used. But of course it is a serious research issue how to construct a knowledge base prior to seeing the discourses it will be used for interpreting. I believe there is a principled methodology for deciding what facts should go into a knowledge base (Hobbs 1984), and there are previous and ongoing efforts to construct a knowledge base of the required sort. For example, WordNet (Miller 1995), while shallow and lacking the required formality, is very broad, and attempts have been made to employ it as a knowledge base in text understanding (Harabagiu and Moldovan 1998). FrameNet (Baker et al. 1998) is a more recent effort aimed at deeper inference, but it is not yet as broad. The efforts of Hobbs et al. (1986), recently resumed, are deeper yet but very much smaller in scope. Cyc (Guha and Lenat 1990) is both broad and deep, but it is not clear how useful it will be for interpreting discourse (e.g., Mahesh et al. 1996). In any case, progress in being made on several fronts.

Another issue I was silent about in presenting the examples was exactly what the measure is that decides among competing interpretations. In some of the examples, factors such as redundancy in explanation and the coverage of the explanations were appealed to as criteria for choosing among them. But this was not made precise. Charniak and Shimony (1990) went a long way in setting the weighting criteria on a firm mathematical foundation, in terms of probabilities. But we still do not have very much experience in seeing how the method works out in practice. My feeling is that now the task is to build up a large knowledge base and do the necessary empirical studies of attempting to process a large number of texts with respect to the knowledge base. That of course requires the knowledge base.

I have written in this chapter only about interpretation, not about generation. It is an interesting question whether generation can be done in the same framework. At the most abstract level, it seems it should be possible. Interpreting a string of words was described as proving the existence of a situation that the string describes. It should be possible correspondingly to characterize the process of describing a situation as the process of proving the existence of a string of words that describes it. Preliminary explorations of this idea are described in Thomason and Hobbs (1997), but these are only preliminary.

The investigation of quantity implicatures should probably be located at the level of interactions between interpretation and generation. The sentence

(22) John has three children.

is usually not said when John has more than three children, even though it is still true in those circumstances. The hearer's reasoning would go something like this: The speaker said U_1, which could mean either M_1 or M_2. But she probably means M_1, because if she had meant M_2, she probably would have said U_2.

Also located in this area is the problem of how speakers are able to co-construct a single coherent segment of discourse, and sometimes a single sentence, across several conversational turns (e.g., Wilkes-Gibbs 1986).

Learning is another important research issue. Any framework that has ambitions of being a serious cognitive model must support an approach to learning. In the IA framework, what is learned is axioms. A set of axioms can be augmented incrementally via the following incremental changes: introducing a new predicate which is a specialization of an old one, increasing the arity of a predicate, adding a proposition to the antecedent of an axiom, and adding a proposition to the consequent of an axiom. But the details of this idea, e.g., when an axioms should be changed, have not yet been worked out.

Finally, there should be a plausible realization of the framework in some kind of neural architecture. The SHRUTI architecture developed by Shastri and his colleagues (e.g., Shastri and Ajjanagade 1993) looks very promising in this regard. The variable binding required by first-order logic is realized by the synchronized firing of neurons, and the weighting scheme in the abduction method is realized by means of variable strengths of activation. But again, details remain to be worked out.

Acknowledgements

This material is based in part on work supported by the National Science Foundation and Advanced Research Projects Agency under Grant Number IRI-9304961 (Integrated Techniques for Generation and Interpretation), and by the National Science Foundation under Grant Number IRI-9619126 (Multimodal Access to Spatial Data). Any opinions, findings, and conclusions or recommendations expressed in this chapter are those of the author and do not necessarily reflect the views of the National Science Foundation.