Judgmental Inference:
A Theory of
Inferential Decision-Making During Understanding

Richard H. Granger

Artificial Intelligence Project
Computer Science Department
University of California
Irvine, California 92717

ABSTRACT

In the course of understanding a text, a succession of decision points arise at which readers are faced with the task of choosing among alternative possible interpretations of what they're reading. Careful analysis of a wide range of sample texts reveals that such decisions are often based on complex evaluations of the interpretation being constructed, and sometimes cause the reader to construct and discard a number of intermediate inferences before settling on a final interpretation for a text.

This paper describes Judgmental Inference theory as a proposed scheme of evaluation metrics and mechanisms, derived from examination of inference decisions arising during text understanding. A series of programs, ARTHUR, MACARTHUR and JUDGE are briefly described, which incorporate some of the metrics and mechanisms of Judgmental Inference, enabling them to understand texts more complex than those that can be handled by other understanding systems.

1.0 Introduction

Many national newspapers carried front-page versions of the following story early this year:

[1] A Nicaraguan soldier, who last year made a public statement alleging Cuban, Ethiopian and Nicaraguan military aid to Salvadorian leftist guerrillas, today publicly retracted his story at a State Department news conference.

Why did the Nicaraguan soldier make the statements he made, a year ago and now? Why did the State Department hold these two news conferences? It is possible that the State Department had some reason for holding the news conference, intending the Nicaraguan soldier to recant; but most readers assume that the State Department had different intentions that were not fulfilled, for reasons out of their control. Indeed, most readers don't even consciously think of the former interpretation, even though it is a logical possible alternative explanation of the events.

Our analysis of examples like this has led to the identification of decision points at which human understanders are faced with the task of choosing particular inferential paths from among an array of possible alternatives. These inference decisions are based on complex evaluation metrics for judging the appropriateness of a particular inference, and on mechanisms for constructing and revising interpretations during understanding. Judgmental Inference theory (Granger [1982]) consists of a set of evaluation metrics and mechanisms derived from examination of inference decisions arising during text understanding. This paper describes how some of these judgmental metrics and mechanisms are applied during understanding.

We view this work as compatible with and complementary to research that focuses primarily on representational issues in text understanding, such as Schank and Abelson [1977], Wilensky [1980], Charniak [1980]. By examining the occurrences of inference decisions during understanding, we intend to provide a look at the mechanisms by which such representations are chosen, constructed, judged, confirmed and/or discarded during the processing of a text.

2.0 Illustration of understanders' decisions

2.1 Evaluating and supplanting inferences

Consider the following example:

[2] Kathy and Chris were playing golf. Kathy hit a shot into the rough. She wanted to let her good friend Chris win the game.

Most readers assume that the reason Kathy hit her shot into the rough was to increase her opponent's chances of winning, out of friendship. However, consider the following:

[3] Ken and Carl were playing golf. Ken hit a shot into the rough.

From reading just this two-sentence version, people infer that Ken and Carl both were playing to win, and that Ken's bad shot therefore was accidental, and will hinder his goal of winning the game. However, after readers have read the third sentence that appears in version [2], they appear to have changed this initial interpretation a great deal. It is not just that Kathy doesn't want to win the game, but also that she probably made her bad shot on purpose, not accidentally. Virtually all readers arrive at this interpretation by the end of this example, by supplanting some of their initial inferences with new ones (see Granger [1980]).

2.2 Evaluation metrics of cohesion and parsimony

Why do people arrive at this different interpretation about Kathy's action in this example? The answer is far from obvious. In particular, there is no question of logical consistency here; the interpretation that Kathy hoped to lose the game but that her bad shot was nonetheless accidental is just as logically consistent as the one that people actually infer, namely that her bad shot was intentional, not accidental.

It turns out that the scope of this phenomenon is very wide: people often arrive at interpretations that appear to involve the supplanting of initial inferences, even when that extra work is not necessary on grounds of logical consistency.

This research was supported in part by the Naval Ocean Systems Center under contract N00123-81-C-1078.

(A large number of additional text examples of this phenomenon are given in Granger [1980] and [1982].)

The decision to reject an initial inference, then, must depend on an evaluation of the representation based on some metric other than logical consistency. One such evaluation metric that was (implicitly) incorporated into previous theories of inference generation (e.g. Rumelhart [1981], Crothers [1978], Bower, Black and Turner [1979], Schank [1973]) we have termed the "cohesion metric". The cohesion metric requires that every statement in a text be connected to at least one other, resulting in all the pieces of the text representation being tied together via either referential, causal or intentional connective inferences.

Cohesion by itself is not sufficient to evaluate the goodness of a text representation, however. Another evaluation metric, identified in our previous work (Granger [1980]), measures the parsimony of a representation, with respect to the goals that motivate the events in the text. For instance, consider the following example:

[4] Doug went to a gas station. He robbed it and got away with $50.

(a) Doug went to the gas station intending to get gas, and then he changed his mind and decided to rob the station instead;

(b) Doug went to the gas station intending to rob it.

Just as in the "golf" example [3], this example can be interpreted in two different ways, both of which are not only logically consistent, but also referentially and causally cohesive, since Doug had to get to the gas station before he could rob it, regardless of his intentions in performing those actions.

Therefore, the cohesion metric does not differentiate between these two alternative interpretations, but people do: they universally seem to generate interpretation (b), which consists of a single goal (getting money from the gas station), and in fact they rarely even consciously notice the possibility of (a), which consists of two separate goals each explaining one of Doug's actions. The evaluation metric of parsimony essentially tests that an interpretation be maximally parsimonious with respect to the number of goals used to explain the events in the story; i.e., the fewer separate motives inferred to account for the story events, the better.

(Note: an evaluation of an unparsimonious interpretation will not always result in the decision to supplant inferences; sometimes readers leave "loose ends" in their interpretation, to be resolved later. See Granger [1982] for a discussion of loose ends.)

2.3 Shaping interpretations of behavior

We have identified some further evaluations that understanders perform, beyond cohesion and parsimony, which arise when a reader is led to "doubt" any part of his interpretation of a text. Such doubts can be instilled either by information presented in the text, or by "extra-textual" factors (see Granger [1981]) which may steer the reader away from an otherwise plausible interpretation. Examples of such "doubt-factors" include the reader's knowledge of the reliability of the text source (e.g., the difference between the New York Times and the National Enquirer); knowledge of an actor's deviousness (e.g., a car salesman vs. a priest); relative boredom or interest, i.e., the reader's desire to pursue possible alternative interpretations vs. just settling on a default interpretation that's "good enough". An easy way to induce a doubt factor in a reader is to simply tell him that his initial interpretation is incorrect; i.e., explicitly ask for a new and different interpretation of a text.

It turns out that readers are very capable of producing a series of such alternative interpretations of texts when they're continually told their initial interpretation is incorrect. For instance, following is a story adapted from a newspaper text, along with a series of interpretations informally elicited from a subject:

[5] The Pakistani ambassador to the United States made an unscheduled stop in Albania on his way home to what an aide of the ambassador described as "a working vacation".

Q1) Why did the ambassador go to Albania?
A1) It looks like he was on vacation — he went to Albania first and then to home, I guess in Pakistan.

Q2) No, that's not the real reason. Why did he go to Albania?
A2) Well, maybe there was some emergency reason ... it said it was unscheduled, so maybe it was that something went wrong and they had to stop there, and then they went on.

Q3) Still not it, but try again; why did he go to Albania?
A3) Ok, maybe, well he's an ambassador, so he could have been supposed to go to Albania ... so it could have been a meeting, like "shuttle diplomacy" ... but it was supposed to be a secret, so that's why they said it was unscheduled.

These different interpretations of [5] are each based on different interpretations of the actor's reasons for doing what he did. It is natural that different behavior interpretations should give rise to different text interpretations; most current theories of text representation focus primarily on representation of the events described in the text, rather than on a more "syntactic" analysis of the structure of the text itself.

Our analysis of this and similar examples has revealed a large class of inference evaluations people perform based on their attempts to decide what kind of behavior an actor has performed, for instance:

1. "simple" goal pursuit, e.g., "John was hungry, so he ate a hamburger";

2. "complex" goal pursuit, (i.e., goal interactions; see Wilensky [1979]) — e.g., "John wanted to see the football game but he also had a paper due the next day" (goal conflict);

3. deceptive or intentionally misleading behavior, e.g., "Clark wanted Lois to think he was drunk, so he smiled and fell off the barstool onto the ground";

4. accidental (non-goal-directed) behavior, e.g., "Jack smiled and fell off the barstool onto the ground"(!);

5. impromptu reactions to unplanned-for contingencies, e.g., "Bill threw himself under the jeep when he saw the man pull a gun".

Our classification scheme for dividing up the gamut of possible interpretations of behavior (e.g., intentional vs. unintentional at the top level, subdividing intentional behaviors into simple, deceptive, pre-planned, impromptu, etc., and unintentional behavior into various types of failures such as skill failure, information failure, etc.) is described in detail in Granger [1982]. We call each of these subdivisions an interpretation-"shape", since categorizing an actor's behavior into one of these classes will result in a particular shape of the representation graph constructed, and because re-interpreting an actor's behavior results in re-shaping the representation.

We have implemented two computer programs, ARTHUR and MACARTHUR, which incorporate the evaluation metrics of cohesion, parsimony, and shapes to produce interpretations of texts that cannot be handled by other text-understanding systems. Granger [1982] gives sample output of the operation of these programs on some of the text examples discussed above.

3.0 Additional categories of inference decisions

3.1 "Suspicious" understanding

It is often impossible for an understander to identify the "correct" interpretation shape for an actor's behavior. For instance, consider the following version of a story that was on the front page of a number of national newspapers earlier this year:

[6] A report by the New York State Racing and Wagering Board released today states unequivocally that leading jockeys conspired to "fix" at least 13 races in the mid-1970's, and that the jockeys have been "patently unbelievable" in denying their involvement in the scheme.

Understanding [6] requires the recognition that the observed behavior of jockeys can be very difficult to classify as either "accidental" or "deceptive". Hence, a jockey (or a jai-alai player, boxer, etc.) may lose a competition without an observer's being able to tell whether he did it intentionally or accidentally.

These are special cases of the general problem of detecting deceptive behavior by using knowledge of "cover stories". Some recent work in AI (e.g., Bruce and Newman [1978]) has pointed out that a method of maintaining separate "belief spaces" for different actors is crucial for understanding deception. However, understanding deception can also require a great deal more than this; in particular, a more subtle deceiver will typically try to cause observers to infer for themselves some false interpretation of his actions, thereby covering up the real reasons. Political propaganda, advertisements for products, and face-saving "white lies" are all examples of this kind of deception. The ability to understand (and generate) complex deceptive behavior such as this depends not only on separate belief spaces, but also on the ability to construct plausible alternative explanations for events. The more plausible the alternative explanation, the more likely the deception is to succeed in misleading understanders.

A "suspicious" understander is one who can (at least) construct alternative interpretations of events, and then can attempt to decide among them, typically by gathering additional information. Such information-gathering is based on finding a possible motive, i.e., finding a plausible explanation that the "obvious" explanation is intended to cover. The JUDGE program, currently under construction, is being designed to make use of knowledge of the shapes of alternative interpretations to detect plausible cover stories in the domain of criminal investigation. For more descriptions of cover stories and JUDGE, see Granger [1982], and Granger and Eiselt [1982].

3.2 Understanding accidents

We have also investigated the types of accidental behavior that can be described in texts, and the relations between accidental and goal-directed behavior. For example recall Ken, who accidentally hit his golf shot into the rough. Although his action of of striking the ball was intentional, the causal outcome of the ball ending up in the rough was unintended. We have classified Ken's problem as a "skill failure", i.e., an intentionally-performed physical action which results in a non-intended outcome as a result of some physical lack. There are a number of other types of intention-accident pairs like this, such as "information failure", "too-shallow planning", etc. For a further discussion of accidents and how to understand them, see Granger [1982], and Meehan [1981].

4.0 Conclusions and future research directions

4.1 What we're proposing

We have observed that people's understanding behavior is marked by an ongoing process of making inference decisions. Among the decisions understanders implicitly make are:

1. Is the interpretation referentially and causally cohesive?
2. Is the interpretation parsimonious with respect to the actors' intentions?
3. Is there reason to doubt or be suspicious of the shape of the initial interpretation?
4. Is there reason enough to revise the interpretation (supplant, re-shape, etc) or should it be left with "loose ends"?

The evaluation metrics and the construction and revision processes of Judgmental Inference theory are derived directly from our observations and analyses of some of the classes of inference decisions that readers are faced with during the task of text understanding.

We view these theories as compatible with and complementary to theories of text representations, since we intend to describe the mechanisms by which such representations are chosen, constructed, judged, confirmed and/or discarded in the process of understanding. Our theories have so far been incorporated into two working computer programs, ARTHUR and MACARTHUR, and are currently being used as the design impetus for a new computer system called JUDGE, and for a series of psychological and neurophysiological experiments, briefly described below, to test the correspondence of our theories to people's actual understanding behavior.

4.2 Minds, brains and processes

A number of researchers in the neurosciences (e.g., Arbib [1979], Geschwind [1980]) have pointed out that brain research might help guide parts of cognitive science and AI research, and vice versa. One particular issue that has been pointed out frequently is that "there is no evidence for the existence of any all-purpose computer [in the brain]. Instead, there seems to be a multiplicity of systems for highly special tasks." (Geschwind [1980], p.191). Our research on inference decisions has indeed led us away from viewing human understanding behavior as arising from a "general purpose computer"; we have ended up instead deriving a number of special-purpose mechanisms, e.g., inference pursuit, evaluation, supplanting, re-shaping, which comprise our "judgmental inference" model of understanding.

We are currently designing a number of psychological and neurological experiments on inference decisions, based on the predictions of our model (see Granger [1982]); as well as attempting to re-interpret some existing results (e.g., Rumelhart [1981], Crothers [1978], Hillyard and Kutas [1980], Black [1981]), in light of the model.

For instance, we are investigating the issue of when people evaluate their interpretations consciously vs unconsciously; our model currently fails to account for such individual differences. We hope to use the data from such experiments to find problems with our theories, and to refine the model, thereby working eventually towards some small amount of "neurological validity" in our process models of cognition.

5.0 References

Arbib, M.A. and Caplan, D. Neurolinguistics must be computational. Behavioral and Brain Sciences, 2:449-483, 1979.

Black, John. The effects of reading purpose on memory for text. Cognitive Science Technical Report 7, Yale University, 1980.

Bower, Gordon H., John B. Black, and T. J. Turner. Scripts in text comprehension and memory. Cognitive Psychology 11:177-220, 1979.

Bruce, Bertram and Denis Newman. Interacting plans. Cognitive Science 2:195-233, 1978.

Charniak, Eugene. On the use of framed knowledge in language comprehension. Yale Computer Science Research Report 137, 1978.

Crothers, Edward J. Inference and Coherence. Discourse Processes, 1:51-71, 1978.

DeJong, G.F. Skimming Stories in Real Time: An Experiment in Integrated Understanding. Yale Computer Science Research Report 158, 1979.

Doyle, J. A truth maintenance system. Artificial Intelligence 12(3), 1979.

Geschwind, N. Neurological Knowledge and Complex Behaviors. Cognitive Science, 4:185-193, 1980.

Granger, R.H. When expectation fails: Towards a self-correcting inference system. In Proceedings of the First National Conference on Artificial Intelligence, Stanford University, 1980.

Granger, R.H. Directing and re-directing inference pursuit: Extra-textual influences on text interpretation. In Proceedings of the Seventh International Joint Conference on Artificial Intelligence (IJCAI), Vancouver, British Columbia, 1981.

Granger, R.H. Judgmental Inference: Inferential Decision-Making During Understanding. Computer Science Technical Report 182, University of California, Irvine, 1982.

Granger, R.H. and K. Eiselt. 'Suspicious' Understanding: Detecting Possible Deception by Inferring Alternative Explanations. Computer Science Technical Report 185, University of California, Irvine, 1982.

Kutas, M. and S.A. Hillyard. Reading Senseless Sentences: Brain Potentials Reflect Semantic Incongruity. Science, 207:203-5, 11 Jan 1980.

Lebowitz, M. Generalization and Memory in an Integrated Understanding System. Computer Scinece Research Report 186, Yale University, 1981.

McDermott, D.V. and Jon Doyle. Non-monotonic logic I. Artificial Intelligence 13(1,2):41-72, 1980.

Meehan, J.R. Boy meets goal, boy loses goal, boy gets goal: the nature of feedback between goal-based simulation and understanding systems. Computer Science Technical Report 170, University of California, Irvine, 1981.

Rumelhart, D.E. Understanding understanding. Technical Report 100, Center for Human Information Processing, University of California, San Diego, January 1981.

Schank, R.C. Causality and Reasoning. Technical Report 1, Istituto per gli studi Semantici e Cognitivi, Castagnola, Switzerland, 1973.

Schank, R.C. and Robert P. Abelson. Scripts, Plans, Goals, and Understanding. Lawrence Erlbaum, Hillsdale, New Jersey, 1977.

Schulenburg, D. Interpreting Intentional and Unintentional Behavior. Technical Report, University of California, Irvine, 1982.

Wilensky, R. Understanding Goal-Based Stories. Research Report 140, Yale Computer Science Department, 1978. Garland Press, New York, 1980.