

Answering Why-Questions:

Test of a Psychological Model of Question Answering

Jonathan M. Golding
Memphis State University

Arthur C. Graesser
Memphis State University

Abstract

We conducted an experimental test of the Graesser and Clark (1985) model of question-answering for why-questions. This model specifies how individuals answer different types of questions by searching through various sources of information after comprehending a text. The sources of information include the passage structure and the generic knowledge structures which are associated with the content words in the query. After these knowledge structures are activated in working memory, search components narrow down a set of relevant answers. A subset of these components were tested in this experiment: (1) an arc search procedure specifying which nodes and arcs within an information source are sampled for answers to a why-question; (2) arc distance, the number of arcs in the representational network that connect the queried node to the answer node; and (3) the intersection of information between the passage structure and the generic knowledge structures associated with the query. After reading short stories, subjects were presented with questions and a number of theoretical answers to each question. Subjects were timed as they judged whether each answer was "Good" (appropriate and relevant) or "Bad" (inappropriate or irrelevant). Results supported the validity of the arc search procedure in that subjects robustly distinguished theoretically good answers from theoretically bad answers to specific questions.

This paper reports an empirical test of Graesser and Clark's (1985) model of question answering in the context of why-questions. This model is an extension of earlier models of question-answering (Graesser & Murachver, 1985; Graesser, Robertson, & Anderson, 1981; Lehnert, 1978), and is developed from a rich data base of question answering protocols and other relevant data (for details about methods, see Graesser & Clark, 1985). The model specifies (1) the major sources of information that are tapped for answers to questions and (2) the processing components that access nodes within the relevant information sources.

There are two major information sources for answers to questions: the passage structure and generic knowledge structures (GKSs). The passage structure includes explicit statements and knowledge-based inferences that are needed for establishing coherence between the explicit statements. The generic knowledge structures are either associated with explicit content words in a query or with higher level GKSs which are triggered by patterns of information (e.g., FAIRYTALE). For example, consider the question "Why did the dragon kidnap the maidens?" in the context of narrative passage N. Four information sources would be searched for answers: passage N, the GKS for DRAGON, the GKS for KIDNAP, and the GKS for MAIDEN. The four passages Graesser and Clark (1985) studied contained 25 explicit statements and 100 inferences, on the average. In addition, a typical passage activated 35 GKSs, with approximately 160 statement nodes in each GKS.

Each passage and each GKS were represented in the form of a conceptual graph structure. A structure contains proposition-like statement nodes that are interrelated by a network of categorized, directed arcs. Each statement node is assigned to one of five categories: Event, State, Goal, Action, and Style. Arc categories included: Reason, Outcome, Initiate, Manner, Consequence, Implies, Property, Set Membership, and Referential Pointer. Graesser and Clark (1985, Chapter 2) have defined the statement node and arc categories in detail. For the purpose of the present paper, it is important to note that the conceptual graph is structured according to explicit, quasi-formal constraints.

There are seven major components involved in accessing information sources and converging on a small set of answers relevant to a question. These include (1) working memory, (2) the activation of knowledge structures in working memory, (3) arc search procedures (specifying "legal" arc paths), (4) priorities among knowledge structures in working memory, (5) the intersection between/among knowledge structures, (6) distance between nodes (arc distance), and (7) constraints on knowledge structures. For example, when answering a why-question involving an action (e.g., "Why did the dragon kidnap the maidens?"), specific information would be activated in working memory (the passage structure, and the GKSs for DRAGON, KIDNAP, and MAIDEN). An entry node (E) in working memory would be found which matched the queried statement. An arc search procedure would examine paths of arcs radiating from node E to superordinate goals via forward Reason arcs and backward Manner arcs. Finally, convergence

mechanisms (e.g., intersection between structures, priorities, constraints) would converge further on a set of relevant answers.

Testing Components of the Model

The present experiment focuses on three of the model's components for why-questions: arc search procedures, arc distance, and the intersection of information between the passage structure and the GKSs associated with the query. Sixteen subjects read two short stories analyzed by Graesser and Clark (1985). After reading each story, subjects were presented with four why-event questions and four why-action questions on a CRT. Following each question were a number of different answers to that question. Subjects were asked to decide (by pressing a specific key) whether each answer was "good" (appropriate and relevant) or "bad" (inappropriate or irrelevant), based on the story. In addition to recording good/bad judgments, decision time was measured in milliseconds. Each subject made judgments for 256 answers, altogether.

An answer to a specific question was designated as "theoretically good" if the arc search procedure generated the answer. For each good answer there was a legal path of arcs between the queried node and the answer node in the passage structure. Otherwise, an answer was designated as "theoretically bad." Arc distance was defined as the number of arcs (in the passage structure) between the queried node and the answer node. Each answer was also scaled on the number of intersecting GKSs. A GKS was scored as "intersecting" if it contained a node that matched the answer node. For example, there would be a match between the answer node "the daughter cried" and "the person cried" in the GKS for KIDNAP. Based on the Graesser and Clark model, it was predicted that: (1) subjects would distinguish theoretically good from bad answers, (2) subjects would respond more quickly to answers having a greater overlap between passage knowledge and GKSs associated with the query (see also Reder, 1982), and (3) would take longer to respond to answers which were more arcs away from the entry node (see also Anderson, 1983).

Analyses robustly confirmed the validity of the arc search procedures for why-action and why-event questions. Subjects correctly judged good theoretical answers as "good" ($M = .65$ for actions and $M = .62$ for events) more than they judged bad answers as "good" ($M = .20$ for actions and $M = .23$ for events). Multiple regression analyses were performed to assess the impact of several predictor variables on the subjects' likelihood of judging an answer as good. The important predictors included whether the answer was theoretically good or theoretically bad (GOOD/BAD), the number of GKSs with nodes that intersect the

answer node (INTERSECT), and the number of arcs between the entry node and the answer node (DISTANCE). A less important predictor was the likelihood that a particular answer presented in the experiment had been generated to its associated question by an independent group of subjects in Graesser and Clark's (1985) question answering task (GENERATE). In addition, each answer was scaled on other predictor variables which were not of direct concern in this study (e.g., number of words in the answer, differences between stories, whether the answer was a state, event, or goal). Among the predictors, there was no problem of collinearity for either actions (range $r^2 = -.35$ to $.34$) or events (range $r^2 = -.25$ to $.38$), except for GENERATE and GOODBAD (action $r^2 = .59$ and event $r^2 = .50$).

The overall regression equations were significant for both answers to why-action questions, $F(7,125) = 18.50$, $p < .01$, $R^2 = .51$ and answers to why-event questions, $F(7, 135) = 12.35$, $p < .01$, $R^2 = .39$. As shown in Table 1, the beta weights indicate that GOOD/BAD was the only significant predictor variable for both of these question types. In fact, GOOD/BAD accounted for most of the predictable variance for both actions ($R^2 = .46$) and events ($R^2 = .34$). It is important to note that the GENERATE predictor was only marginally significant for the queried events. Thus, the theoretical arc search procedure (reflected in the GOOD/BAD variable) is a more robust predictor than simply the number of people who generate a specific answer to a specific question in a question-answering task.

Another set of multiple regression analyses were performed on decision latencies, using the same predictor variables. The overall regression equations significantly predicted latencies for actions, $F(7, 125) = 6.37$, $p < .01$, $R^2 = .26$, and for events, $F(7, 135) = 5.98$, $p < .01$, $R^2 = .24$. The beta-weights in Table 2 indicate that GOOD/BAD was again the only significant predictor. Decision latencies were significantly longer for good answers than bad answers.

The lack of effects for arc distance (DISTANCE) and intersecting GKSSs (INTERSECT) led to additional multiple regression analyses. It was felt that the overall speed of a judgment might have determined whether the predictor had an effect on decision latencies. Perhaps the DISTANCE and INTERSECT variables are significant only when decision latencies are comparatively fast; the effect may be masked by other processes when the decision latencies are very long. The analyses were conducted on the four fastest times and four slowest times for each answer, segregating queried actions and queried events. Once again, however, GOOD/BAD was the only consistent significant predictor. Moreover,

accuracy was essentially equal for the fast and slow judgments. Regardless of the time taken to make the judgment, the arc search procedure yielded the same degree of accuracy.

Discussion

This experiment investigated three components of Graesser and Clark's (1985) model of question answering: arc search procedures, arc distance, and the intersection between the passage structure and GKSs associated with the query. There was clear support for the arc search component. For both events and actions, subjects robustly distinguished good theoretical answers from bad theoretical answers. Decision latencies were longer for the good theoretical answers than the bad theoretical answers. These results underscore the importance of specifying the legal paths of arcs and nodes while searching for answers during question answering.

The lack of significant effects for other predictor variables awaits further investigation. The lack of an arc distance effect, especially in the fast-slow analyses is perplexing. The use of cohesive texts may have masked this effect. Perhaps the effects of arc distance and passage-GKS intersection will emerge for expository text and narratives which do not fit a clear script or prototypical story format (e.g. Keenan, Baillet, & Brown, 1984).

The absence of an effect for passage-GKS intersection might be a result of the type of question being studied. Why-questions generally probe the explicit statements and comprehension-generated inferences in the passage structure (Graesser & Clark, 1985). Consequently, the GKSs may have had a minimal role in question answering procedures. Other types of questions might highlight the role of GKSs during question answering. For example, how-questions involve GKSs to a greater extent than the passage structure. Perhaps the passage-GKS intersection will be more pronounced when how-questions are analyzed in the context of our narrative passages.

References

- Anderson, J.R. (1983). The architecture of cognition. Cambridge, MA: Harvard University Press.
- Graesser, A. C., & Clark, L. F. (1985). Structures and procedures of implicit knowledge. Norwood, NJ: Ablex.
- Graesser, A. C., & Murachver, T. (1985). Symbolic procedures of question answering. In A. C. Graesser and J. B. Black (Eds.), The psychology of questions. Hillsdale, NJ: Earlbaum.
- Graesser, A. C., Robertson, S. P., & Anderson, P. A. (1981). Incorporating inferences in narrative representations: A study of how and why. Cognitive Psychology, 13, 1-26.
- Keenan, J. M., Baillet, S. D., & Brown, P. (1984). The effects of causal cohesion on comprehension and memory. Journal of Verbal Learning and Verbal Behavior, 23, 115-126.
- Lehnert, W. G. (1978). The process of question answering. Hillsdale, NJ: Earlbaum.
- Reder, L. M. (1982). Plausibility judgments versus fact retrieval: Alternative strategies for sentence verification. Psychological Review, 89, 250-280.

Table 1

Question Answering Data for Answers to Why-Questions

Regression Coefficients of Predictor Variables¹:

Probability of Saying Good Answer

	Queried Action	Queried Event
GOOD/BAD	.61***	.48***
INTERSECT	.15**	.04
DISTANCE	-.02	.01
GENERATE	-.02	.17**

Regression Coefficients of Predictor Variables¹:

Decision Latencies

	Queried Action	Queried Event
GOOD/BAD	.16*	.25***
INTERSECT	.03	.03
DISTANCE	-.04	.10
GENERATE	-.15	-.07

* p < .10

** p < .05

*** p < .01

- ¹GOOD/BAD : whether an answer was generated by theoretical arc search procedure.
 INTERSECT: number of GKSs with a node intersecting the answer node in the passage structure.
 DISTANCE : number of arcs between queried node and answer node in passage structure.
 GENERATE : number of subjects who would generate the specific answer to the question in question-answering protocols.