Simultaneous Question Comprehension and Answer Retrieval

Scott P. Robertson, Jonathan D. Ullman, Anmol Mehta

Psychology Department
Rutgers University - Busch Campus
New Brunswick, NJ 08903

Abstract

A model is described for question comprehension in which parsing, memory activation, identification and application of retrieval heuristics, and answer formulation are highly interactive processes operating in parallel. The model contrasts significantly with serial models in the literature, although it is more in line with parallel models of sentence comprehension. Two experiments are described in support of the parallel view of question answering. In one, differential reading times for different question types were shown to be present only when subjects intended to answer the questions they were reading. In another, reading times for words in questions increased and answering times decreased when a unique answer could be identified early in the questions. The results suggest that source node activation and answer retrieval begin during parsing. Both symbolic and connectionist approaches to modeling question answering are potentially influenced by this perspective.

Question Answering

Question answering is a process that has interested researchers in several disciplines within cognitive science, especially cognitive psychology (Graesser & Franklin, 1990; Graesser, Robertson, & Anderson, 1981; Singer, 1984a, 1984b, 1986; Robertson & Weber, 1990), artificial intelligence (Dyer, 1983; Lehnert, 1977, 1978), philosophy of language (Belnap & Steel, 1976), and PDP modeling (Miikkulainen & Dyer, 1990). Question answering is also an important applied problem in query-directed information retrieval systems and in the context of education (Schank, 1986).

Question answering is interesting because it involves question-specific retrieval operations over complex mental representations. Researchers in this area have concentrated mainly on the relation between question types and retrieval heuristics, or on the heuristics themselves. Largely as a simplifying assumption, they have considered question answering to be independent of the language comprehension processes involved in question parsing or the language generation processes involved in answer production. In this paper we take issue with this view of the independence of question parsing, answer retrieval, and answer production. Following from Robertson & Weber (1990) we argue that retrieval and parsing, at least, occur simultaneously and may interact.

The main components of question answering are parsing to produce a conceptual representation of linguistic input, source node activation based on the conceptual representation, identification of retrieval heuristics appropriate for the identified question type, application of retrieval heuristics to identify or generate answer candidates, pruning of answer candidates based on pragmatic, appropriateness, and other criteria to isolate a single answer, and production of the answer in linguistic form. The most explicit models of question answering in the literature--Dyer (1983), Graesser & Franklin (1990), Lehnert (1978), and Singer (1986)--treat these as stages in a serial process. Indeed, this is the easiest thing to do since there are many dependencies among these processes, and most of the dependencies move from parsing toward production. For example, in some cases the question category can only be uniquely identified after it is determined whether the question presupposition is a motivated action or part of an unmotivated causal sequence.

The serial bias rests on many assumptions that may not be valid, however. In particular, serial models make assumptions about the need to pursue an answer in a single category or according to a unique retrieval rule. For example, in answering a

question that begins with "Why did John...," serial models would be unable to begin because it is unclear at this point whether retrieval heuristics should be applied to search goal structures (as in "Why did John go to the store?") or causal chains (as in "Why did John fall down?," if he didn't do it on purpose!). If we imagine, however, that both retrieval processes could begin, activating all causal consequents involving "John" and all goals that "John" had, then we no longer need to assume an ordered relationship between parsing and retrieval. Instead, we are faced with describing how independent, but simultaneous, processes might interact and share resources.

The TSUNAMI Model

As an alternative framework for thinking about question answering, we are developing a model called TSUNAMI, for "Theory of Simultaneous UNderstanding Answering and Memory Interaction." At this point the model is offered as a broad architecture for supporting highly interactive application of the mechanisms already identified by question answering researchers. It remains to be seen how the nature of these mechanisms will change, and what new mechanisms might be necessary, when implemented in the TSUNAMI framework.

The TSUNAMI model, depicted in Figure 1, utilizes two working memory components. One memory stores question candidates and the other stores answer candidates. These working memory components act like "blackboard" data bases in that items stored there may be inspected and altered by several processes operating at once (Erman & Lesser, 1980). The question candidate memory and answer candidate memory are the only knowledge structures that take output from processes in the model (processes are indicated by ovals). The influences of processes on these memories might be to add propositions, update proposition contents, or delete propositions. The behaviors of processes that use the data in the question and answer candidate memories, in turn, are influenced by the contents of those memories.

Parsing and Matching

Processing begins when the *parser* starts receiving input from a question. the parser utilizes grammatical, case, and pragmatic information in semantic memory to produce various propositional representations which are then stored in the question candidate buffer. Multiple arrows from the parser into the question candidate buffer suggest that the parser can produce many candidates, often only partially specified propositions, in response to the input at any

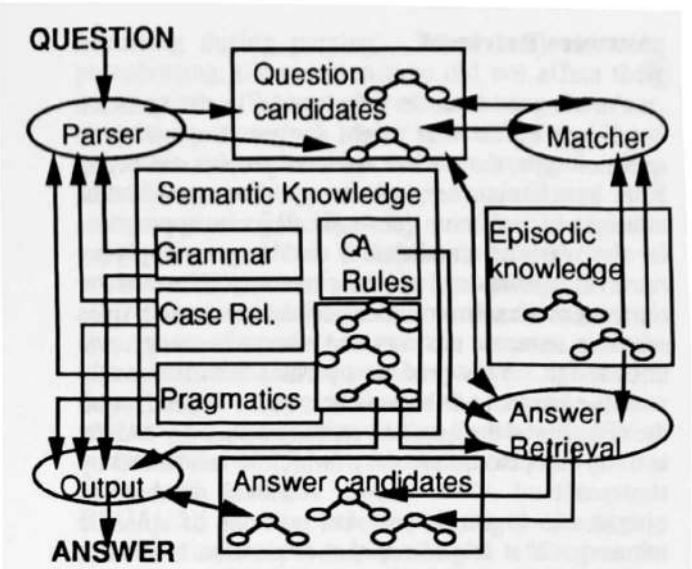

Figure 1. The TSUNAMI model of question answering.

given time. As words come into the parser, prior candidate structures may be updated or disconfirmed and deleted by the parser.

As soon as there is any information in the question candidate buffer, a *matcher* begins comparing question candidate structures with activated information in episodic memory; specifically, the subset of episodic memory that is considered relevant to the question (e.g. a story just read, a set of memories foregrounded in the conversational context, etc.). We assume that the matching process occurs in parallel for several candidates (as indicated by multiple arrows from both the question buffer and episodic memory into the matcher) but that its speed is affected by the number of candidates and the number of matches. If the matcher finds a proposition in episodic memory that corresponds to a question candidate, then this is identified as a likely source node for answer retrieval processes. Influences of the matcher on question candidates are indicated by an arrow from the matcher into the question candidate memory. Matches in episodic memory raise the activation level of this information and make it more available to future analysis by any processes that utilize episodic memory (in this way the model is like Anderson's 1983 ACT* model). When a match is found for a question candidate, other candidates become less likely interpretations of the input. We assume that partial propositions in the question candidate memory can be matched to complete propositions in memory, thereby updating the question candidate list to include *expectations*. Predicted question candidates can influence answer retrieval as the retrieval process, described below, is blind to the status or origin of propositions in the question buffer.

Answer Retrieval

As soon as there is information in the question candidate buffer that might suggest a question or question type, the *answer retrieval* process can begin. This mechanism examines question candidates, attempts to determine question categories appropriate to the various candidates, and begins applying retrieval heuristics in episodic memory. The answer retrieval mechanism utilizes question answering rules stored in semantic memory and other, relevant general knowledge. This process operates simultaneously with the parser and the matcher, but is dependent on the contents of the question candidate memory and the activity in episodic memory (which is influenced by the matcher). The answer retrieval mechanism operates to highlight relevant portions of episodic memory. If it is guessed that a question is about a goal, for example, then the answer retrieval mechanism may activate goal hierarchies in the episodic trace. On the other hand, if the mechanism is expecting a causal antecedent question, then causal sequences may become more active. This activity will affect the behavior of the matcher.

The outputs of the answer retrieval process are propositions that are candidate answers to questions that the system finds consistent with the input (and memory) and any given time. Potential answers produced by the answer retrieval process are stored in the *answer candidate buffer* as propositions. These propositions may also be partially specified, and are subject to subsequent modification and deletion by the operation of the answer retrieval process. For example, if a question candidate that spawned a retrieval process is later disconfirmed, then the answer candidate built by that process will be deleted by the answer retrieval process.

Output

The answer candidates are examined by an *output preparation* process which also utilizes grammatical, case, pragmatic, and relevant general knowledge in semantic memory. This process can influence the answer candidate set. For example, if pragmatic concerns dictate that an answer is inappropriate, then the output preparation process will delete it from the answer buffer. Finally, when one candidate remains in the answer buffer and all of the question has been input to the parser, the final answer is formulated. It is reasonable to assume that the output preparation mechanism will not commit to a final interpretation until all of the input has been processed since a final phrase on a question can change its focus, and hence the appropriate answer, tremendously.

Experiment on Comprehension Instructions

The TSUNAMI model posits that parsing, matching, and retrieval processes share resources. Effects of retrieval processes on parsing can be seen in increased reading times for the words of a question. Such increases would be due to the increased workload resulting from simultaneous processes sharing resources. In Robertson & Weber (1990), we showed that knowledge of the question type during reading of a question (when the question word was at the beginning of a question) increased word-by-word reading times but decreased answer retrieval time when compared to conditions in which the question type was not known (when the question word was at the end of a question). Increased reading times suggested parallel retrieval and parsing. Decreased answering times reinforced this interpretation by showing that the answer was "closer" and that the workload effect was related to the answer retrieval process.

In that study we also observed increased answering times for time questions ("When did...") relative to reason questions ("Why did..."). Reason questions were answered 116ms faster than time questions. We have observed this discrepancy in two subsequent extensions of that study (299ms in one case and 259ms in the other), and it was the only reliable effect in another study on presentation speed of questions (455ms). In short, the reason-time discrepancy is a highly reliable effect related in some way to differences in the retrieval processes for these two types of questions.

In this experiment we exploited the reason-time discrepancy and sought to find it during reading. Also, we asked if a reason-time effect would be present in reading times only when subjects were reading with the intention of answering a question. If subjects were not intending to answer a question, then the answer retrieval mechanism would be inactive and the reason-time discrepancy should not be apparent.

Method

Subjects. Twenty-six subjects participated in this study for credit in Introductory Psychology.

Materials. Forty-eight short (5-7 line) stories were written. In each story a character went to some location. A reason question and a time question were prepared for each story. The reason questions read "Why did <ACTOR> go to the <LOCATION>?," whereas the time questions read "When did <ACTOR> go to the <LOCATION>?"

Design and Procedure. There were three instruction conditions in the experiment: *no-story paraphrase*, *story paraphrase*, and *story answer*. In the *no story paraphrase* condition subjects read questions (self-paced one word at a time) and were told to come up with a paraphrase. When they had reached the last word of the question they were to press the response key "when the meaning of the question was understood." They then wrote down their paraphrase. After this they pressed the response key and saw a computer generated question and were asked to judge if it "meant the same thing" as the question. The latter task was intended to reinforce the paraphrase instruction. Subjects worked through eight reason questions and eight time questions randomly intermixed in a block. In the *story paraphrase* condition subjects read a paragraph-long story which was then followed by a question. The question was presented in the same manner as the *no story paraphrase* condition and subjects were instructed to come up with a paraphrase in the same way. Each question was followed by a "means the same thing" judgement and there were again eight reason and eight time questions randomly intermixed in a block. Finally, in the *story answer* condition the subjects received stories followed by questions as in the *story paraphrase* condition, but this time their instruction was to come up with answers to the questions and press the response key "when an answer comes to mind." Similarly, they were asked to judge if the second question "has the same answer" as the first. Stories were randomly assigned to conditions and rotated through the conditions across subjects. Instruction block orders were counterbalanced.

Table 1

Mean reading time (ms/syllable) for all but the last word of reason and time questions read under three comprehension instructions: no-story paraphrase (NSP), story paraphrase (SP), and story answer (SA).

	INSTR			
QUEST	NSP	SP	SA	Mean
Reason	390	367	332	363
Time	401	370	371	381
Mean	395	368	351	372

Results and Discussion

Table 1 shows the mean reading times per syllable for all of the words of the question except the last. Reading time for the last word includes time for memory retrieval and answer/paraphrase formulation, and it is not of interest for studying processes occurring during parsing. When subjects were paraphrasing, the question type did not affect their reading times. When they were answering, however, the time questions took longer to read than the reason questions. An interaction between comprehension instruction and question type confirmed this interpretation, $F(2,50)=3.24$, $p<.05$. The results support the hypothesis that question-related retrieval processes are being activated during reading.

One explanation for the reason-time discrepancy is that the set of possible answers to a reason question is a subset of the set of possible answers to a similar time question. A goal is always an acceptable answer to a time question ("John went to the store WHEN he wanted some milk") whereas a time is a bad answer to a goal question ("John went to the store BECAUSE it was Saturday") except in highly specific circumstances (e.g. if John worked at the store on weekends in our example). Hence, in the TSUNAMI framework when a time question is being processed the question buffer contains more possible question interpretations relative to a reason question, the matcher would activate more nodes, and the retriever would generate more answer candidates. This would slow the overall operating time of the parser as seen in this experiment.

Experiment on Number of Unique Answers

In this experiment we concentrated on the role that the matcher and answer retriever play in the TSUNAMI model. In the model, the matcher tries to find antecedents in episodic memory for propositions in the question buffer. This process raises the activation level of the antecedents making them more likely to serve as source nodes for the retrieval process. The fewer memory items there are that are consistent with the input at any given point, the further the answer retriever can go.

We manipulated the contents of episodic memory in such a way that sometimes the source node from which retrieval processes would begin could be identified early in parsing by virtue of a unique actor. Subjects read stories in which an action was performed at four different times for four different reasons. In each story one actor performed the action on three occasions while the other performed the action on one occasion. In the following story, for example, Mary is the unique actor and John is a multiple actor:

John went to the store to buy bread on Monday.
John went to the store to buy milk on Tuesday.
John went to the store to buy cheese on Wednesday.
Mary went to the store to buy eggs on Thursday.

Now consider questions like the following:

q1. Why did John drive to the store on Tuesday?
q2. Why did Mary drive to the store on Thursday?

In answering q1 it is impossible to identify the exact source node in memory that corresponds to the question presupposition until the end of the question. In q2, however, it is possible to identify the unique source node in memory as early at the subject, Mary. In a question answering architecture with parallelism, like TSUNAMI, answer retrieval heuristics should begin earlier when reading q2 than q1. If simultaneous parsing and retrieval compete for resources as we have argued, then increased reading times should be observed for the words in q2 relative to q1 as the parser slows down. Additionally, if the increased workload is due to retrieval processes, then the answering time at the end of the questions should be faster for q2 than q1. In strict serial models, in contrast, source node activation and application of retrieval heuristics would be delayed until after question parsing and no reading time differences should be apparent (if anything, spreading activation theory for antecedent concepts predicts longer reading times for q1 over q2, Anderson 1976).

Method

Subjects. Twenty-eight Rutgers undergraduates participated for credit in Introductory Psychology.

Materials. Sixteen four-sentence stories like the one above were constructed for the experiment. Each story consisted of four instances of the same action performed at four different times for four different purposes. In each story there were two characters. When the stories were presented, one character was associated with three actions while the second was associated with a single action. For each story, one action was chosen as the "query action," about which a question would be asked. The actor associated with the query action was varied across subjects so that for some subjects the query action was performed by the unique character and for other subjects the query action was performed by the character who did several things.

Design and Procedure. Each subject read the sixteen stories and answered two questions about each one. The entire text of each story was presented on a computer screen and subjects spent as long as they liked reading it. When they were finished they pressed a response key. At this time a prompt appeared on the screen. Each subsequent keypress revealed a word of the question, and the words appeared side-by-side in their normal positions. Subjects were instructed that on the last word of the questions they should press the response key "as soon as an answer comes to mind." Reason questions were in the form "Why did NOUN1 VERB PREP1 DET NOUN2 PREP2 TIME?" Time questions were of the form "When did NOUN1 VERB PREP1 DET NOUN2 AUX VERB2 NOUN3?" The reading times for each word were recorded.

Subjects were first asked a reason or time question about each story. Each subject was asked reason questions about eight stories and time questions about eight stories. For each subject, half of the questions were about the action performed uniquely by one actor (unique action) and half were about one of the three actions performed by the other actor (non-unique action). The question type condition (reason/time) and action uniqueness condition (unique/non-unique) were crossed. Across subjects, the stories were rotated through the conditions.

Since it was possible to answer the time and reason questions without paying attention to the actors in the story, each reason/time question was followed by a "who" question about each story. The antecedent for the who question was chosen randomly between the unique action and one of the non-unique actions. The who-question guarantied that subjects would pay close attention to the actors. Reading times were not collected for these questions.

Table 2
Mean reading time (ms/word) averaged across the five common words

Question	Actor		
	Unique	NonUnique	Mean
Reason	483	463	473
Time	507	465	486
Mean	495	464	479

Results and Discussion

Table 2 shows the mean reading times averaged across NOUN1, VERB, PREP, DET, and NOUN2 for the reason questions and time questions in the unique and non-unique actor conditions. As predicted, the time to read the questions was greater in the unique action condition relative to the non-unique action condition, $F(1,27)=7.50$, $p<.05$. There was no effect of question type and no interaction.

Our second prediction, that answering time would be faster in the unique action condition relative to the non-unique action condition, was also confirmed. The mean answer times were 2449ms vs 3554ms in the unique vs. non-unique conditions respectively, $F(1,27)=19.11$, $p<.001$. In contrast to other

experiments in our lab, reason questions were answered more slowly overall than time questions (3256ms vs 2746ms), $F(1,27)=9.39$, $p<.05$. There was no interaction.

The results support the hypothesis that if a source node can be identified early during parsing, retrieval heuristics can be identified and applied. The simultaneous operation of the parser and answer retriever slows both, but pays off in the end with a faster answer.

Final Comments

We have proposed a new architecture for question answering that has many parallel components and presented empirical evidence in support of it. A parallel view of question answering would bring this important aspect of language processing into line with current thinking on parallel processes in sentence parsing (Gorrell, 1989; McClelland & Kawamoto, 1986; Miikkulainen & Dyer, 1991; St. John & McClelland, 1990; Waltz & Pollack, 1985). Of course, the experiments support the general idea of parallelism, not the specifics of the TSUNAMI model. However, the model is general enough to incorporate many specific instantiations. As it develops it will be interesting to see how, or if, changes will be necessary in the retrieval heuristics proposed by researchers working within a serial paradigm. More than likely a new class of problems will arise having to do with conflict resolution among competing question interpretations and answer possibilities in the face of partial input.

Recently there has been considerable progress on connectionist models of sentence parsing (McClelland & Kawamoto, 1986; Miikkulainen & Dyer, 1991), and PDP models will inevitably begin to approach the problem of question answering. In this paradigm too it will be necessary to face the issue of whether the output of a parsing network should be the input to a question answering network, or whether these processes are more closely intertwined. Our results suggest the latter approach.

References

Anderson, J.R. (1976). *Language, memory, and thought*. Hillsdale, N.J.: Erlbaum.

Anderson, J.R. (1983). *The architecture of cognition*. Cambridge, MA: Harvard University Press.

Belnap, N.D., & Steel, T.S. (1976). *The logic of questions and answers*. New Haven: Yale University Press.

Dyer, M.G. (1983). *In-depth understanding: A computer model of integrated processing for narrative comprehension*. Cambridge, MA: MIT Press.

Erman, L.D., & Lesser, U.R. (1980). The HEARSAY-II speech understanding system: A tutorial. In W. Lea (Ed.), *Trends in speech recognition*. Englewood Cliffs, NJ: Prentice-Hall.

Gorrell, P. (1989). Establishing the loci of serial and parallel effects in syntactic processing. *Psycholinguistic Research, 18*, 61-74.

Graesser, A.C., & Franklin, S.P. (1990). QUEST: A cognitive model of question answering. *Discourse Processes, 13*, 279-304.

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.

Lehnert, W.G. (1977). Human and computational question answering. *Cognitive Science, 1*, 47-73.

Lehnert, W.G. (1978). *The process of question answering*. Hillsdale, NJ: Erlbaum.

McClelland, J.L., & Kawamoto, A.H. (1989). Mechanisms of sentence processing: Assigning roles to constituents. In J. McClelland & D. Rumelhart, (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. Volume 2: Psychological and biological models*. Cambridge, MA: MIT Press.

Miikkulainen, R., & Dyer, M.G. (1991). Natural language processing with modular PDP networks and distributed lexicon. *Cognitive Science, 15*, 343-400.

Robertson, S.P., & Weber, K. (1990). Parallel processes during question answering. *Proceedings of the 12th Meeting of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.

Schank, R.C. (1986). *Explanation patterns.: Understanding mechanically and creatively*. Hillsdale, NJ: Erlbaum.

Singer, M. (1984a). Mental processes of question answering. In A.C. Graesser & J.B. Black (Eds.), *The psychology of questions*. Hillsdale, NJ: Erlbaum.

Singer, M. (1984b). Toward a model of question answering: Yes-no questions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*, 285-297.

Singer, M. (1986). Answering wh- questions about sentences and text. *Journal of Memory and Language, 25*, 238-254.

St. John, M.F., & McClelland, J.L. (1990). Learning and applying contextual constraints in sentence comprehension. *Artificial Intelligence, 46*, 217-258.

Waltz, D.L., & Pollack, J.B. (1985). Massively parallel parsing: A strongly interactive model of natural language interpretation. *Cognitive Science, 9*, 51-74.