LANGUAGE RESEARCH SPONSORED BY ONR

Susan Chipman

Program Manager, Cognitive Science
Office of Naval Research
Arlington, VA 22217-5660

In contrast to DARPA, ONR has not had a defined program that is focused exclusively on language research. However, there have been some clusters of projects concerned with language that have been supported within the Cognitive Science Program and its prior incarnation, the Personnel and Training Research Program. The ONR program is a basic research program, but a mission-oriented one which seeks to generate results that will prove applicable to significant Navy applications. Traditionally, as the earlier name indicates, those applications were sought primarily in training; more recently, the targets have been broadened to include human factors applications, especially in human-system interaction. Originally, the program was a psychological research program, but it evolved into a cognitive science program, the first government research program to be so labeled. The ONR Cognitive Science Program emphasizes the use of AI techniques to model human cognitive performance, and it has also emphasized the special sub-field of AI concerned with artificially intelligent computerized instruction or tutoring systems (ICAI or ITS). Thus, the ONR Cognitive Science Program also contrasts with the DARPA Human Language Technology Program in being concerned with understanding how human language actually is produced and processed by humans. (There is also an AI program within the Computer Science Division of ONR.) The management style of the program might be best described as dynamic coherence. That is, a degree of coherence or focus is necessary in order to enhance the likelihood of identifiable impact on application areas, but the focal clusters do evolve over time, partially in response to promising proposals that are received. New proposals are judged partially by the extent to which they cohere with and enhance the current portfolio of projects, not merely on isolated merit and inclusion in the broad area of cognitive science. Two salient clusters of language-related research that emerged have been the currently active emphasis on tutorial discourse and an earlier interest in improving the readability of instructional texts and documentation. Both have had significant natural language AI aspects, but not all projects have involved computation.

TUTORIAL DISCOURSE

This cluster of projects follows earlier major investments in artificially intelligent tutoring systems. The striking effectiveness of one-on-one tutorial instruction by human tutors (estimated to be a 2 standard deviation improvement over conventional classroom instruction) has sparked great interest in efforts to emulate that effectiveness with artificially intelligent computerized instructional systems. Despite enough success in that endeavor to make the production of intelligent tutoring systems a more applied research or development activity, present intelligent tutoring systems circumvent, evade, and finesse the problem of natural language interaction in various ways because the demands of tutorial interaction are really beyond the state of the art in computerized natural language. In the ONR Cognitive Science Program, human tutorial interaction is being studied from the perspectives of linguists, psychologists, and computational linguists who aim to emulate it in artificial systems. Among the issues that arise in these studies are the size or scope of the discourse organization imparted by the tutor, the balance between the tutor's agenda and immediate responsiveness to the student, the extent to which tutors revise their plans dynamically, the nature and breadth of knowledge required to support these interactions, the relationship between tutorial interaction and normal conversational patterns, and the nature of repair and correction processes, including the use of positive, neutral and negative feedback. The ease or feasibility of emulating these features of human tutorial discourse certainly varies, but it is also true that the introduction of a computer as a conversational participant is a significant change: what is the perceived social status or role of a computer? Similarly, it is possible that ideal computerized tutorial discourse might differ from what is observed among humans. The diverse research perspectives required to address these issues typify the interdisciplinary character of cognitive science.

Apart from some informal studies conducted within larger projects concerned with building early intelligent tutoring systems, the first of these recent studies of tutorial discourse was conducted by Barbara Fox, a linguist at the University of Colorado. Her primary data were four hour-long sessions of math/science tutoring which were video-taped and transcribed in a very detailed way that records pauses and non-verbal behavior (according to the methods of Sacks, Schegloff and Jefferson). She focused on correction and repair processes in tutorial discourse. She concluded that tutors structure correction activity so as to enable students to correct their own errors, whenever possible. Tutoring, in spite of its emphasis on learning, and therefore on making and correcting mistakes, is organized by the same principle of correction that organizes everyday conversation -- the preference for self correction. She found that tutors make heavy use of pre-correction strategies and of silence, in order to accomplish this preference for self correction (on the part of the students). Pre-correction strategies signal an upcoming correction from the tutor; they serve to alert the student to the possibility of that she has made an error. The student can then engage in trying to figure out what the error might be, and then can try to correct it him/herself. Throughout this process, the tutor provides further feedback to the student to indicate whether or not the student is on the right track. Tutors also give students a considerable amount of time in answering questions before they step in to redirect or correct (depending on the context, the silence can be from 1 to 5 seconds or so). That is, tutors often wait to see if the student will self correct. And tutors do not force an immediate answer; they give the student time to address the question. If the student is displaying obvious signs of being lost or stuck, however, the tutor provides immediate guidance. Fox also studied some tutorial sessions conducted by teletype in which the tutees were led to believe that the tutor was a computer. These sessions revealed some interesting differences in interaction. Students expect tutoring computers to be capable of complex numerical computations, and they feel free to leave dead time and to make off-the-wall remarks. These students appeared to believe that they actually were interacting with a computer.

A project by Arthur Graesser at Memphis State that is still on-going includes much larger samples of interacting students and tutors. One sample was of school children being tutored in arithmetic. Another analyzed interaction patterns in 44 one-hour tutoring sessions involving college students. Undergraduate students were tutored by graduate students on troublesome topics in a research methods course in psychology (e.g., variables, statistics, factorial designs, hypothesis testing). Similar to Fox, the primary focus was on collaborative exchanges and feedback mechanisms during question asking and question answering. Graesser has identified substantial problems in knowledge tracking and feedback mechanisms between student and tutor; the pragmatic principles of politeness and cooperativity during conversation often seemed to present a barrier to effective pedagogy during tutoring. The relationship between student question asking and level of achievement has been analyzed, as well as the way tutors handled student errors. A model of tutorial interaction is being developed which specifies dialogue patterns, pragmatic assumptions, goal structures, and pedagogical strategies during question asking and answering. Although the present project does not include artificial production of tutoring dialogue, Graesser, a psychologist with a joint appointment in computer science, has previously done computer simulations of question asking and answering. He regards these studies of naturalistic tutoring as a necessary preliminary to the design of effective dialogue facilities in intelligent tutoring systems, although he believes that there might be ways in which artificial tutorial dialogue could improve upon the natural.

Two other current projects do involve artificial production of tutorial dialogue. Martha Evens, a computational linguist at IIT, has been working to develop an intelligent tutoring system with genuine natural language interaction capacity. She and her collaborators are building an intelligent tutoring system that can carry out a tutorial dialogue with first year medical students, helping them to understand the negative feedback system that controls blood pressure, guiding them in building a qualitative, causal mental model of the system. With the goal of understanding how human tutors generate tutorial dialogues in this situation, they have captured seven face-to-face and thirty-seven keyboard-to-keyboard tutoring sessions, each lasting an hour or more. The tutors are professors of physiology at Rush Medical College; the students are first year medical students from their classes. The study of human tutoring sessions reveals many examples where the expert tutors produce large-scale discourse structures: multi-turn discourse structures, multi-stage hints, series of Socratic questions, directed chains of reasoning. Investigating tutors' responses to student initiatives, Evens and her associates found that tutors always respond to student initiatives to some extent. Revelations of serious misconceptions change the tutor's agenda to elimination of the misconception. An initiative from the student that is relevant to the tutor's current agenda results in a modification of the plan to incorporate the issue raised by the students. Other student initiatives evoke a brief response, followed by a return to the tutor's agenda. Unlike Fox, they found direct negative feedback in their tutorial dialogs 25% of the time as well as direct contradictions of what the student has just said 10% of the time. Keyboard-to-keyboard communication resulted in more elaborate positive and negative feedback responses from the tutors, fewer turns, and slower initiation of student responses. The current version of the tutor, Circsim-Tutor Version 2, generates lesson plans and tactics on both large and small scales, but in the process of executing a given plan, it generates the actual dialogue a turn at a time. The next version will attempt to emulate the larger structures of the human tutors. Evens judges that the tutorial repair processes described by Fox seem extremely difficult to emulate. Therefore, she is attempting to avoid repair by studying the source of repair situations and avoiding them: the most common source of conversational misunderstanding is vague "how" questions from the tutor. Evens and her collaborators are trying to generate more specific questions. Another source of misunderstanding is the tutor's misinterpretation of very terse and ill-formed input from the student; Evens' tutor is checking those interpretations with the student. Although Evens studied tutorial interaction over a computer link (as well as face-to-face tutoring) in order to approximate the conditions of computer tutoring, although she has devoted considerable effort to dealing with the error-ridden, abbreviated, and elliptical input from the students, this tutor may interest DARPA grantees as a potential testbed for the integration of speech recognition with natural language. Speech interaction would be a highly desirable feature for computerized training systems. Given the limited resources of the ONR program, however, we are relying upon DARPA to solve the speech recognition problem.

The project in this cluster that has begun most recently is that of Johanna Moore, a computer scientist at Pittsburgh. Although the aim of her project is the artificial generation of tutorial explanations, she also has begun by studying human tutors in order to identify the properties that make them effective. She replaced the natural language component of an existing ITS with a human tutor, and gathered protocols of students interacting with the human tutor. (The existing tutor is the Sherlock tutor of skill in diagnosing problems with an avionics test station, an Air Force project. The existing tutor has been evaluated in workplace training and found highly effective; it is the first of a large number of maintenance training tutors which the Air Force plans to develop for actual, practical training use.) She then systematically compared the human's responses to those that would have been produced by the ITS, identifying two critical features that distinguish human tutorial explanations from those of their computational counterparts. First, human explainers freely exploit the previous discourse in their subsequent explanations. This facilitates understanding and learning by relating new information effectively to recently conveyed material, and avoiding repetition of old material that could distract the student from what is new. Second, human tutors make extensive use of discourse markers to express relationships among individual units of information. These markers provide cues to the structure of the explanation and the information being conveyed, and thus make the explanations easier to understand. Moore is now constructing a computational explanation planner capable of assigning appropriate discourse markers and of generating explanations that make use of prior discourse in ways done by human tutors.

TEXT READABILITY

A second cluster of language research projects has focused on improving text readability. The military services and their contractors produce enormous amounts of text, system documentation and training materials for the personnel who will operate and maintain those systems. Consequently, there has been interest in research aiming to make these materials readable and comprehensible for their users. A few years ago, the ONR program supported a cluster of projects concerned with the design of readable and comprehensible procedural instructions, a special genre that has been neglected in general educational research on reading and text design. Among these projects was an effort by David Kieras, now at the University of Michigan. In a basic research project, Kieras demonstrated an automated system that could provide rather sophisticated comments on text structure and quality, such as, "This paragraph does not seem to have a main idea." This was done without any true comprehension of the text. The text was parsed and a propositional representation of the text was constructed. Propositions were linked by repeated mentions of the same term. Comments on text coherence could then be derived.

In conjunction with a project to develop a system to aid the authors of Navy training materials (AIM), the Navy Personnel Research and Development Center provided somewhat more applied funding (6.2) for further work by Kieras on the development of a text critiquing system that might enhance the capabilities of AIM. Kieras reviewed the psycholinguistic research literature on the determinants of text readability and comprehensibility. (Current standards for readability are based on crude formulas that measure sentence length and the frequency/familiarity of words used in the text. Yet, it is known that conversions to shorter sentences can sometimes make texts less comprehensible by obscuring the connections among ideas in the text.) As a first step, Kieras put considerable effort into building a parser that could handle actual Navy training documents in production. These do have some unusual structural format features that are not found in the texts for which most existing parsers were designed. In addition, many of these texts are written by senior enlisted personnel with subject matter expertise, personnel who are not trained or talented as writers. They can present severe parsing challenges even to highly skilled human readers. Experts in computational linguistics will not be surprised to hear that the "finished" version of Kieras' parser cannot parse many of the sentences or so-called sentences in these training documents, although failure to parse might sometimes be appropriate grounds for criticizing the writing. (Kieras's parser is written in Common Lisp and is available for those who might want to use it, along with a documenting manual that explains how to add additional capabilities to it.) Kieras did go on to build a text critiquing system based on his earlier work and the broader psycholinguistic research literature. Experimentation with this system revealed some interesting problem areas: for example, the system makes too many spurious complaints about the introduction of "new referents". This happens because it does not know about semantic relations among the words in the text, such as synonymy and part-whole relations. If the F-14 has been discussed, it will respond to "the wing" as a new referent. In its present state, the critiquing system does not have a practical, reasonably friendly user interface. In addition to making errors, it does not prune or prioritize comments but outputs an overwhelming barrage. Design of an effective user interface has not yet been supported.

(NPRDC has also been interested in related work by Bruce Britton, a psychologist at the University of Georgia, who was initially supported by the Air Force. Britton has shown that revisions guided by the same principles implemented in Kieras' system do improve text comprehension. Britton has developed some simple computer programs that aid human users in doing the same kinds of analyses done by the Kieras program, relying upon the human users to do parsing and supply semantic knowledge.)

In addition to the Kieras project, several others have been supported with a view to potential applications in a system like AIM. Navy support of George Miller's WordNet project began with this rationale, although it was obvious that WordNet would be a very general lexical resource with diverse potential applications in natural language computing. WordNet might be used to aid authors in finding more frequent and familiar words to substitute for rare words in their initial drafts. Specialists in natural language computing might note an interesting irony here: very frequent words tend to be very ambiguous semantically. Thus, to make a text readable to human readers of limited ability, one is advised to substitute very ambiguous words for less ambiguous ones. Semantic ambiguity does not seem to be problematic for human readers. In addition, of course, WordNet seemed to have promise as a way of eliminating some of the erroneous comments generated by the text critiquing system.

Another project has been a system of automated text formatting developed by Thomas Bever, a psycholinguist at Rutgers. Bever's system inserts slightly larger spaces at phrase boundaries (roughly speaking) in the text. Perhaps surprisingly, this has significant effects on reading speed and comprehension performance, especially for less skilled readers, although the text retains a normal appearance. (All of the services have many personnel with rather poor reading skills. Remedial reading instruction is a major training expense.) Formatted texts have even been shown to improve performance in an entire training course. Preliminary results suggest that the effects of text formatting are much larger for texts presented on computer screens -- the expected future format of military documentation. Bever's system does not parse the text in order to insert these spaces; he has developed a set of surface rules for doing it. However, he has also shown that a neural net can be trained to do space insertion very quickly when trained by a few texts that have been marked by a human. In this way, formatting could be applied to languages other than English with a very small expenditure of effort. It might be a very useful feature for the translators' work stations that have been a target application for DARPA work in machine translation and machine-added translation. At this time, the psychological mechanism by which this formatting aids readers is unknown; presumably it relieves some of the processing burden of parsing the text. Because of its demonstrated effectiveness and unobtrusiveness, Bever's text formatting is likely to move into practical application in major Navy training manuals soon.

At one time, we had hoped that the AIM system would remain an on-going applied project with periodic upgrades that would provide a conduit for the ready application of research advances. However, a managerial decision was made that all such projects would be limited to a 3-year span. The AIM system, which is implemented on Sun computers, was declared finished and fielded as little more than a fancy word processing system that helps to meet the special formatting requirements of Navy training documents, supplemented by a MacDraw-like capability for scanning in and manipulating illustrative diagrams. It is very popular with its users. Obviously, efforts may be made to initiate a new project that would incorporate more sophisticated language processing capabilities.

Although psychological research on text design has a long history, advances in linguistic understanding change the nature of the questions that can be asked in such research and the nature of the phenomena that are noticed. In particular, as in the research on tutorial discourse, much is being discovered about larger-scale discourse structures. Related issues, such as the effective design and use of diagrams, are much less well understood. At present, these are not high priority topics in the basic research Cognitive Science Program at ONR, but they may receive attention in the future, and some projects may receive support through the 6.2 Manpower, Personnel and Training R&D Committee at ONR.