17th International Conference on Advanced Science and Technology (ICAST 2001), Chicago. pp. 47–51.

Evaluation of the Turn Planner in CIRCSIM-Tutor

Feng-Jen Yang
CS Dept., Rowan University
201 Mullica Hill Road
Glassboro, NJ 08028
yangf@rowan.edu

Michael Glass
CS Dept., U. of Ill. at Chicago
851 S. Morgan
Chicago, IL 60607
mglass@cs.uic.edu

Martha W. Evens
CS Dept., Ill. Institute of Technology
10 W. 31st St.
Chicago, IL 60616
evens@iit.edu

Abstract

CIRCSIM-Tutor is an intelligent tutoring system designed to imitate human tutorial dialogue. The turn planner is a planning module intended to be plugged into the forthcoming new version that improves dialogue coherence by post-processing the collection of sentences that constitute a single dialogue turn. In order to know how much the turn planner improves the dialogue, we are now facing the issue of how to evaluate the contribution of turn planning to the whole system. In this paper we propose a method to evaluate the adequacy from the user's perspective of the turn planner's changes to the text.

Introduction

CIRCSIM-Tutor is an intelligent tutoring system using natural language dialogue to tutor medical students in problem solving in the domain of reflex control of blood pressure. While using this system the student is presented with a predefined procedure and then is asked to predict the qualitative changes in seven core variables at three stages. These predictions are the basis of a tutorial dialogue to remediate any misconceptions revealed.

The current version of CIRCSM-Tutor generates its tutorial language a sentence at a time from pedagogical dialogue schemas. Without considering many inter-sentence linguistic issues, it generates dialogues that are comprehensible but sound unnatural. In the new version, we would like to improve dialogue coherence by modifying the sentences to take contextual information into consideration [Yang et al. 2000a, Yang et al. 2000b].

This paper outlines a method to evaluate the proposed language produced by turn planning from the perspective of user acceptability.

This work was supported by the Cognitive Science Program, Office of Naval Research under Grant No. N00014–94–1–0338 to Illinois Institute of Technology. The content does not reflect the position or policy of the government and no official endorsement should be inferred.

Related Research

There is no objective standard for the quality of natural language dialogue and texts and there are few measurable characteristics that could constitute a quality measure. This makes it hard to evaluate performance. Even the comparison of alternative systems in similar domains is virtually impossible [Fraser 1997].

Nonetheless, the evaluation of natural language systems still plays a critical role in guiding and focusing research in computational linguistics. It challenges researchers in both building advanced systems and solving hard problems. In the past decade, some conferences and workshops, such as Message Understanding Conferences (MUCs), Spoken Language Technology Workshops, and Machine Translation Workshops, have been focused on the evaluation of natural language systems.

Hirschman and Thompson [1997] describe three common areas of evaluation for natural language processing systems:

- Adequacy Evaluation: The fitness of a system for a special purpose is one of the critical factors in bringing natural language systems to market. Potential users have to know if the products on offer in a given application domain are suitable for their particular tasks or not. If so, they have to make further tradeoffs between fitness and cost and then choose the most suitable one.
- Diagnostic Evaluation: For systems where the coverage is important, the developers or end-users usually construct a large test suite to cover all of the elementary linguistic phenomena and their important combinations in the input domain. By testing systems with a large test suite, they can generate diagnostic profiles. The typical systems using this evaluation are machine translation and natural language understanding systems.
- Performance Evaluation: Most of the ideas about quantitative performance evaluations are imported from information retrieval. This evaluation comes with three levels of specificity. The first is Criterion, which addresses what to evaluate such as Precision, Speed and Error Rate. The second is Measure, which specifies the property to report in order to get the chosen criterion

such as Ratio of Hits to Hits Plus Misses, Second to Process and Incorrect Percentage. The third is Method, which is used to determine the appropriate value for a given measure such as the Analysis of System Behavior over Benchmark Tasks. In natural language systems, these approaches provide a useful way for system developers to compare different implementations of a technology or different versions of the same implementation.

In this paper we propose, strictly speaking, to evaluate the performance of the *design* of the CIRCSIM-Tutor turn planner. We can readily decide whether the software produces the desired transformations of the sentences. We want to measure the quality of each of the proposed text transformations. In Hirschman and Thompson's typology, this is an evaluation of performance on the criterion of user acceptability. The method will be by surveying users, and the measure will be counting survey results.

The Evaluation of DIAG-NLP

DIAG-NLP [Di Eugenio et al. 2001, Di Eugenio and Trolio 2000] improves the tutoring language of an intelligent tutoring system written in the VIVIDS [Munro 1994] and DIAG [Towne 1997] authoring environment. It is related to the current enterprise in that Di Eugenio et al. are producing and evaluating pairs of tutoring systems that differ only in the quality of their generated text.

The most interesting result from the evaluation of DIAG-NLP is evidence that in an intelligent tutoring system the quality of generated text matters.

The system teaches the repair of an oil-fired home heating system. It presents the student with graphical process flow and functional diagrams, allowing the student to view various indicators (e.g. an oil flow gauge) and replace various components. While solving a problem, the student is able to consult the system regarding the meaning of various indications and the likelihood that various parts are causing the problem.

In these consultations the tutoring system engages in a kind of very limited dialogue where the student is constrained to asking only a few questions by clicking on elements of the process diagram. The tutor responds in English. The responses are conditioned by the tutoring context, for example in evaluating the plausibility that a particular part is faulty it reminds the student of indicators that have already been viewed.

The contribution of DIAG-NLP beyond the original home heating tutorial is to improve the quality of the language in the consult responses. The original language is quite repetitive, for example "Oil Supply Valve always produces this abnormality when it fails. Oil Pump always produces this abnormality when it fails. System Control Module sometimes produces this abnormality when it fails." The published version of DIAG-NLP improves the language by means of linguistic aggregation on two levels, thus removing the mind-numbing repetition, and improving the layout of the text on the screen by using better indentation and spacing.

DIAG-NLP was then evaluated versus the unmodified home heating tutorial. Each of the systems was used by a number of subjects. Their performance in using the tutorial was measured, they were quizzed on their knowledge afterward, and they answered survey questions. Performance measures included items such as how many consultations were needed to solve the problems, how long they spent reading the consult responses, and how many parts were replaced until the problems were fixed. The survey questions included evaluation of qualities such as conciseness and usefulness.

The result was a win for better text. Not many of the differences in objective or subjective measures individually showed strong statistical significance. But taken in aggregate, the improved-text version performed better in eight out of nine measures, which was significant.

The Evaluation of Integrated Text and Examples

Mittal [1999] examined issues of incorporating instances of examples in expository technical text. He was interested in questions such as placement, e.g. does the example occur before, within, or after the discussion, whether examples should be presented in order of increasing complexity, whether "prompts" (attention-focusing devices) are included, and so on. His generation system produces expository and tutorial text, with incorporated examples, on the subject of LISP.

Having identified a number of factors controlling the presentation of examples in text, he produced pairs of texts minimally different in one factor. Some of the examples illustrated the concept under discussion, some did not. Experimental subjects answered a set of questions identifying which of the examples were felicitous and which were not.

It is important to note that Mittal was not evaluating the quality of a generation system, he was exploring how controlling various factors for example integration resulted in more or less comprehensible texts. The result of his evaluation guided rules for his generation system. The proposal in this paper is similar in purpose: we have identified various text characteristics to be handled by the CIRCSIM-Tutor turn planner, we want to evaluate how important and useful these various characteristics are for the users.

The Evaluation of KNIGHT

The KNIGHT system [Lester and Porter 1997] explains concepts in the domain of biology. It is a large system, producing multi-paragraph texts from a knowledge base of 180,000 anatomical, physiological, and developmental facts.

KNIGHT's performance was measured by a novel "two panel" evaluation method. In this method, the KNIGHT system constructed explanations of approximately three sentences on 60 randomly chosen topics. One panel of domain experts wrote 60 explanations on the same topics, constrained to match the computer-generated texts as to the length and target audience (freshmen biology students). A second panel of domain experts graded the explanations on an A, B, C, D, F categorial grading scale. The grading panelists were blind as to the purpose of the study, they were not even informed that half the texts were computer-written and half were human-written. The explanations received grades on overall quality and coherence (a kind of summary grade), as well as content, organization, writing style, and correctness.

For each of the five grade categories, KNIGHT performed within half a grade below the human writers. The differences were not statistically significant in any category, but the authors suggest this may be due to the sample size.

Evaluation of the Turn Planner

Primarily we will evaluate user acceptability of our turn planner's text modifications by means of user survey of dialogue on paper. We discuss that evaluation here. Later we can test two groups of students using different versions of CIRCSIM-Tutor, with and without turn planning.

The paper test is constructed as follows. For each individual type of modification made by the turn planner, we will provide two dialogues from different contexts, one incorporating the modification and one without. The dialogues will be judged by representatives of our user population, first-year medical students in physiology class who have studied the subject. This gives us access to a panel of about 50 raters.

The pairs of texts are not minimally different, meaning that besides the deliberate modification they are *not* otherwise identical. Instead, we start with a pair of texts A and B that present similarly structured arguments using different physiological variables. We construct two complementary pairs A' B and A B', where A' and B' incorporate the text planner modification. We take this approach to preserve blindness. For statistical strength we prefer to have both elements of the pair graded by the same students, but with minimal pairs the students will likely notice the variables we are testing for.

Every student will grade a selection of dialogues on a scale. Then the grading differences can be used as an indication of the validity of applying each improvement.

What follows are some of our preliminary texts, illustrating the three types of textual modifications made by the turn planner: adding discourse markers, modifying the referring expressions for physiological variables, and modifying the acknowledgments of the student's answer [Yang et al. 2000a]. We will also prepare texts with all three types of modifications together.

The Validity of Applying Discourse Markers

Human tutors tend to use discourse markers such as "so," "but," "now," "therefore," etc. to move from one discourse focus to another and make the tutorial dialogue sound more fluent and easier to understand. We have studied the selection of these discourse markers in the transcripts of human tutors that inform CIRCSIM-Tutor [Kim et al. 2000]. To evaluate the validity of applying discourse markers, we will have the medical students grade the following two dialogues. One adds the discourse markers to structural transition points in the tutor's argument, and the other disregards discourse markers. By analyzing the difference in their grading, we can have an indication of how much the discourse markers improve the acceptability of tutorial dialogue.

The following example incorporates discourse markers:

T: Can you tell me by what mechanism Inotropic State is controlled?
S: nervous system
T: Right.
So, what is the value of IS in DR?
S: decreased
T: No, it is not decreased in DR.
Remember, it is neurally controlled.
Now, what is the value of IS in DR?
S: not changed
T: Very good.

The following example disregards discourse markers:

T: Do you know what are the determinants of Cardiac Output?
S: hr sv
T: Which determinant is dominant in this case?
S: hr
T: No, the important determinant of CO in this case is SV.
HR didn't change yet.
Is the relationship from SV to CO direct or is it inverse?
S: direct
T: Good.
What is the correct value of CO?
S: down
T: Very Good.

The Validity of Improving Variable References

Human tutors tend to use abbreviated variable names and pronouns to stay in the same discourse focus. They spell out the names of physiological variables only once in a while. To evaluate the validity of improving referring expressions for physiological variables, we will have the medical students grade the following two dialogues. One

is generated with improved variable referring expressions while the other is generated without this consideration. The grading difference can then be an indication of the validity of improving variable references.

The following example is generated with improved variable references:

T: Do you know, by what mechanism is Total Peripheral Resistance controlled?
S: neural
T: Right, it is controlled by the nervous system.
Remember, the reflex will not activate until the RR stage.
So, what is the value of TPR in DR?
S: unchanged
T: Great.

The following example is generated without improvement in variable references:

T: Can you tell me what are the determinants of Cardiac Output?
S: sv and hr
T: Right, Cardiac Output is determined by Stroke Volume and Heart Rate.
Now, which determinant is dominant in this case?
S: sv
T: Right, the important determinant of Cardiac Output in this case is Stroke Volume.
And is the relationship from Stroke Volume to Cardiac Output direct or inverse?
S: direct
T: Good.
So, what is the correct value of Cardiac Output?
S: decreased
T: Absolutely.

The Validity of Improving Acknowledgment

There are a number of nuances in the responses human tutors give when acknowledging student answers. Tutors may use words such as "right," "good," "great," "absolutely," etc. to accept student's answers, but they usually don't repeat student's answers back as part of the acknowledgement. However they sometimes do repeat students' wrong answers. If the student had made some mistakes, and finally figured out the right answer, the tutor tends to acknowledge it more strongly than regular manner. CIRCSIM-Tutor gives a cumbersome, stylized response to student answers that usually repeats what the student said. To evaluate the validity of our rules for improving acknowledgments, we will have the medical students grade the following two dialogues.

The following example is generated with improvements in acknowledgments:

T: Can you tell me what are the determinants of Mean Arterial Pressure?
S: co and tpr
T: Right.
And which determinant is dominant in this case?
S: tpr
T: No, it is not.
The important determinant of MAP in this case is CO.
TPR didn't change yet.
Do you know is the relationship from CO to MAP direct or inverse?
S: direct
T: Good.
So, what is the correct value of MAP?
S: decreased
T: Absolutely.

The following example is generated without improvements in acknowledgments, using the acknowledgments currently generated by CIRCSIM-Tutor:

T: Can you tell me what is the determinant of Central Venous Pressure?
S: co
T: Correct, CVP is determined by CO.
And is the relationship from CO to CVP direct or is it inverse?
S: direct
T: Sorry, you got the wrong answer.
A change in CO has an inverse effect on CVP.
So, what is the correct value of CVP?
S: decreased
T: Correct, the value of CVP is decreased.

Assessing Performance of Integrated System

We hope that after turn planning our tutorial dialogue will be more coherent, more fluent, and easier to understand. To evaluate the contribution of turn planning as a whole, we plan to provide two versions of CIRCSIM-Tutor: one applies turn planning while the other disregards it, and we will separate our users into two groups.

Just as with DIAG-NLP [Di Eugenio and Trolio 2000, Di Eugenio et al. 2001] there are a number of different measures available for comparison. One is pre-test and post-tests of student knowledge and problem-solving ability, others are measures of the student's interaction with the tutoring system, such as number of erroneous variable predictions or time on task. We can also survey the students afterward as to user acceptability.

Discussion and Conclusions

We have noticed three basic approaches to evaluation of text generation and dialogue. First is to have people judge the quality of the text directly. This was the approach taken by Lester and Porter [1997] in evaluating KNIGHT. Second is to expose the people to the text in order to accomplish some task, then survey the people afterward. Di Eugenio et al. based some of their measures of DIAG-

NLP on this approach. Third is to expose the people to the text in order to accomplish some task, then measure the performance of the people in accomplishing the task. Di Eugenio et al. did this for some measures, and Mittal [1999] took this approach for determining rules for integrating examples and text.

The last approach, measuring how varying text influences task performance, would seem to be both the most objective and the most utilitarian form of evaluation. For us, it is also the most expensive. It is expensive partly because it requires two versions of an entire working tutoring system, and it is expensive because it would conflict with other tests during one of our infrequent opportunities to test new versions of CIRCSIM-Tutor. We prefer to save this kind of evaluation until we have some confidence in our turn planning rules.

Furthermore comparison of student performance with tutoring systems that differ only in text quality presents a difficulty: the system and the student are both oriented toward achieving the goals of completing the tasks and learning the material. In the DIAG-NLP experiment, the students who saw the better text performed only very slightly better on the knowledge post-test. (The differences were larger, but still not highly significant, on other performance measures.)

To put the problem crudely, it is not clear how bad text has to be before students stop learning.

Mittal reports the same phenomenon with goal-oriented text reading. Recall that his subjects were required to judge the felicity of examples incorporated into expository texts, the examples having been presented according to various rules. Initially, his experimental subjects were likely to judge the felicity all examples correctly. They simply spent enough time with each text, no matter how infelicitous or badly presented were the examples, until they figured it out. Only after he limited the available time to answer the question did useful differences appear.

The approach of Lester and Porter, using a panel of judges to rate text directly, hearkens back to Pirsig [1974]. Pirsig discusses how he had students in his university rhetoric classes rank the quality of different texts, then compared the students' relative rankings to his own. He makes the claim that people can obtain fair agreement on writing quality, even without being able to define or measure it.

The belief that direct rating of text acceptability is both cheap and effective led us to our proposed evaluation method for the turn planner text modifications in CIRCSIM-Tutor.

Acknowledgments

Reva Freedman introduced us to the idea of turn planning, showed us why it is needed, and pointed us in the right direction.

References

Di Eugenio, B., Glass, M., Trolio, M., and Haller, S. 2001. Simple Natural Language Generation and Intelligent Tutoring Systems. Presented at the Workshop on Educational Dialogue Systems held in conjunction with the Artificial Intelligence in Education conference.

Di Eugenio, B., and Trolio, M.J. 2000. Can Simple Sentence Planning Improve the Interaction Between Learners and an Intelligent Tutoring System? In *Building Dialogue Systems for Tutorial Applications* (AAAI Fall Symposium, Cape Cod). Menlo Park, CA: AAAI Press, technical report FS-00-01.

Fraser, N.M. 1997. Spoken Dialogue System Evaluation: A First Framework for Reporting Results. *Proceedings of the Fifth European Conference on Speech Communication and Technology*, Rhodes, Greece, pp. 1907–1910.

Hirschman, L., and Thompson, H. 1997. Overview of Evaluation in Speech and Natural Language Processing, In Varile, G., Zampolli, A., Cole, R., Mariani, J., Uszkoreit, H., Zaenen, A. and Zue, V. (editors), *Survey of the State of the Art in Human Language Technology*, Chapter 13.1, Cambridge, UK: Cambridge University Press.

Kim, J. H, Glass, M., Freedman, R., and Evens, M.W. 2000. Learning the Use of Discourse Markers in Tutorial Dialogue for an Intelligent Tutoring System. *Twenty-Second Annual Conference of the Cognitive Science Society, Philadelphia*, pp. 262–267.

Lester, J. and Porter, B. 1997. Developing and Empirically Evaluating Robust Explanation Generators: The KNIGHT Experiments, *Computational Linguistics*, vol. 23, no. 1, pp. 65–101.

Mittal, V.O. 1999. *Generating Natural Language Descriptions with Integrated Text and Examples*, Mahwah, NJ: Lawrence Erlbaum Associates, chapter 9.

Munro, A. 1994. Authoring Interactive Graphical Models. In de Jong, T., Towne, D.M., and Spada, H., eds., *The Use of Computer Models for Explication, Analysis, and Experiential Learning*. Berlin: Springer-Verlag.

Pirsig, R.M. 1974. *Zen and the Art of Motorcycle Maintenance*, New York: William Morrow and Co.

Towne, D.M. 1997. Approximate Reasoning Techniques for Intelligent Diagnostic Instruction. *International Journal of Artificial Intelligence in Education*, vol. 8 pp. 262–283.

Yang, F., Kim, J., Glass, M. and Evens, M.W. 2000a. Turn Planning in CIRCSIM-Tutor. *Proceedings of the 13th International Florida Artificial Intelligence Research Society Conference (FLAIRS 2000)*, Orlando, FL, pp. 60–64.

Yang, F., Kim, J., Glass, M. and Evens, M.W. 2000b. Lexical Usage in the Tutoring Schemata of CIRCSIM-Tutor: Analysis of Variable References and Discourse Markers. *Proceedings of the 5th* International *Conference on Human Interaction with Complex Systems (HICS 2000)*, Urbana, IL, pp. 27–31.