The KNIGHT Experiments: Empirically Evaluating an Explanation Generation System*

James C. Lester
Department of Computer Science
North Carolina State University
Raleigh, NC 27695-8206
(lester@adm.csc.ncsu.edu)

Bruce W. Porter
Department of Computer Sciences
University of Texas at Austin
Austin, Texas 78712
(porter@cs.utexas.edu)

Abstract

Empirically evaluating explanation generators poses a notoriously difficult problem. To address this problem, we constructed KNIGHT, a robust explanation generator that dynamically constructs natural language explanations about scientific phenomena. We then undertook the most extensive and rigorous empirical evaluation ever conducted on an explanation generator. First, KNIGHT constructed explanations on randomly chosen topics from the Biology Knowledge Base. This is an immense structure that contains more than 180,000 facts. We then enlisted the services of a panel of domain experts to produce explanations on these same topics. Finally, we submitted all of these explanations to a second panel of domain experts, who graded the explanations on an A–F scale. KNIGHT scored within "half a grade" of the domain experts. Its performance exceeded that of one of the domain experts.

Introduction

The issue of evaluation has posed a notoriously difficult problem for the field of artificial intelligence. It is still unresolved. Moreover, because of the complexity and subjectivity of natural language, evaluation is particularly challenging for projects in computational linguistics. Only in the past few years have researchers in natural language understanding begun to make headway on evaluation (Sundheim 1991). For several reasons, which we discuss in the following section, the field of explanation generation has not witnessed the development of a similar "empiricist" school. However, as a noted researcher in explanation generation has observed, "Empirical evaluation is vital if we are to develop practical explanation systems" (Cawsey 1992).

Support for this research is provided by grants from the National Science Foundation (IRI-8620052 and IRI-9120310), a contract from the Air Force Office of Scientific Research (F49620-93-1-0239), and donations from the Digital Equipment Corporation.

An empirical approach to evaluation offers several benefits. It would permit researchers to substantiate claims about the effectiveness of their architectures and design principles. It would also enable them to calibrate the successes of their systems, both relative to other systems and also to human "explanation generators." This in turn would provide a means for the field to measure its progress. Finally, the conclusions of well designed empirical studies should be considerably less equivocal than those derived from "proof-of-concept" systems.

We have developed and empirically evaluated KNIGHT (Knowledge-Integration-based Generator of History-sensitive Text) (Lester & Porter 1991a; 1991b; Lester 1994), a robust explanation generation system that dynamically constructs natural language explanations about scientific phenomena. To generate explanations, KNIGHT extracts knowledge structures from a large-scale knowledge base, organizes them into hierarchical discourse plans, and employs the FUF realization system (Elhadad 1992) to translate them into smooth English prose. The details of KNIGHT's operation may be found in (Lester 1994).

We conducted a rigorous empirical evaluation of KNIGHT. In the course of this study, we subjected its explanations to intense scrutiny by domain experts. An immense undertaking, this evaluation is by far the most extensive that has ever been conducted on an explanation generator with sophisticated discourse planning and realization facilities. Our study employed two panels of domain experts. Experts on the first panel served as "writers," i.e., they produced explanations in response to questions. Experts on the second panel served as "judges," i.e., they analyzed different dimensions of explanations and assigned grades. The results of this study are both surprising and very encouraging.

The Challenges of Empirical Evaluation

Traditionally, research projects in explanation generation have not culminated in empirical evaluations.

Conducting a formal study with a generator has posed difficulties for at least three reasons:

- the absence of large-scale knowledge bases,
- the problem of robustness, and
- the subjective nature of the task.

First, the field of explanation generation has experienced a dearth of "raw materials." The task of an explanation generator is three-fold: to extract information from a knowledge base, to organize this information, and to translate it to natural language. Unless an explanation generator has access to a sufficiently large knowledge base, the first step—and hence the second and third—cannot be carried out enough times to evaluate the system empirically. Unfortunately, because of the tremendous cost of construction, large-scale knowledge bases are scarce.

Second, even if large-scale knowledge bases were more plentiful, an explanation generator cannot be evaluated on them unless it was specifically designed to perform robustly. In very practical terms, a generator is likely to halt abruptly when it encounters unusual and unexpected knowledge structures. Frequent premature termination makes it very difficult to evaluate the quality of explanations for more than a small number cases. Hence, it is our (unverified) conjecture that most implemented explanation generators would meet with serious difficulties when applied to a large-scale knowledge base.

Third, explanation generation is an ill-defined task. Ideally, we would like to "measure" the coherence of explanations. Although it is clear that coherence is of paramount importance for explanation generation, there is no litmus test for it.

Given these difficulties, how can one evaluate the architectures, algorithms, and knowledge structures that form the basis for an explanation generator? The traditional approach has been to

1. conduct an analytical evaluation of a system's architecture and principal,
2. demonstrate that it can produce well-formed explanations on a few examples.

We believe that while these evaluation techniques are necessary, they are not sufficient. By far the vast majority of research on explanation generation has adopted this approach. However, there are three notable exceptions. By varying pragmatic information such as tone, Hovy enabled PAULINE to generate many different paragraphs on the same topic. PAULINE's texts were not formally analyzed by a panel of judges, and it did not produce texts on a wide range of topics (it generated texts on only three different events.) However, this project is a significant achievement in terms of evaluation *scale* because of the sheer number of texts it produced: PAULINE generated more than 100 different paragraphs on the same subject. In a second landmark evaluation, Cawsey undertook a study in which subjects were allowed to interact with her explanation generation system, EDGE (Cawsey 1992). Subjects posed questions to EDGE about the operation of four circuits. Cawsey analyzed the system's behavior as the dialogs progressed, interviewed subjects, and used the results to revise the system. Although EDGE does not include a realization system (other than simple templates), it was sufficiently robust to be used interactively by eight subjects. Finally, Mittal developed and evaluated a generator that produced descriptions integrating text and examples (Mittal 1993). The level of formality of his empirical evaluation far surpassed his predecessors. For example, the degree to which he controlled for specific factors, e.g., the effect of example positioning, example types, example complexity, and example order, is remarkable.

Clearly, all of these projects have significantly raised the standards for the evaluation of explanation generation systems; they point the way toward a much more extensive, rigorous evaluation. However, given the three problems outlined above (the absence of large-scale knowledge bases, the problem of robustness, and the subjective nature of the task), how can we advance the state of the art in evaluation methodology?

In our own work, we have taken three steps to address this issue. First, we and our colleagues have undertaken a mammoth knowledge base construction project. This has been a seven year effort involving many domain experts, graduate students, and programmers. Second, we have designed and implemented a very robust system. As discussed in previous chapters, the explanation planner, its knowledge base accessing system, and the realization system were all designed to perform well when operating in a large-scale knowledge base environment. Third, we have developed a method for minimizing the problem of subjectivity by employing many human subjects in our study. The quality of an explanation is (and always will be) a subjective measure. However, by seeking the analysis of a panel of judges, we can minimize the effects of subjectivity: such a panel will rarely reach a consensus, but its collective opinion provides persuasive evidence about the quality of explanations. By taking these three steps, we were well positioned to conduct a rigorous empirical evaluation.

The Explanation Generator's Input

KNIGHT draws its representational structures from the Biology Knowledge Base, a massive structure representing the domain of plant biology (Porter *et al.* 1988). It is one of the largest knowledge bases in existence. The backbone of the knowledge base is its taxonomy, a very large hierarchical structure containing all of the knowledge base's biological objects (representing botanical anatomy) and biological processes (representing botanical physiology and development). It also includes the hierarchy of all of the relations that may hold between concepts. The reader may find it instructive to view a sample knowledge structure from the Biology Knowledge Base. Figure 1 depicts a small portion of the representation of the concept *photosynthesis*. This is a typical fragment of the semantic network that constitutes the knowledge base. Each of the nodes in this network is a concept, e.g. *photosynthesis*, and each of the arcs is a relation in the knowledge base. For example, the *transducer* for *photosynthesis* is the concept *chlorophyll*. The Biology Knowledge Base currently contains more than 180,000 explicitly represented "triples," i.e., facts of the form (*Unit Slot Value*).

To ensure that the knowledge base was not tailored for the purposes of explanation generation, the designers of the explanation generator and the representation team entered into a "contractual agreement":

> Under no circumstances could the designers of the explanation generator request the knowledge engineers to alter the structures of the knowledge base other than for purposes of consistency or completeness.

The designer of the explanation generator was able to request representational changes only if knowledge was inconsistent or missing. In short, this agreement eliminated all requests for representational modifications that would simplify the task of explanation generation. This agreement has resulted in a unique experiment in which the representational structures were *not* tailored for the task of explanation generation.

Experimental Design

Explanation Generation: Knight

Because KNIGHT's operation is initiated when a client poses a question, our first task was to select the questions it would be asked. To this end, we combed the Biology Knowledge Base for concepts that could furnish topics for questions. Although the knowledge base focuses on botanical anatomy, physiology, and development, it also contains a substantial amount of information about biological taxons. Because this latter area is significantly less developed, we ruled out concepts about taxons. In addition, we ruled out concepts that were too abstract, e.g., *Spatially-Extended-Entity*. We then requested KNIGHT to generate explanations about the 388 concepts that passed through these filters.

To thoroughly exercise KNIGHT's organizational abilities, we were most interested in observing its performance on longer explanations. Hence, we eliminated explanations of concepts that were sparsely represented in the knowledge base. To this end, we passed the 388 explanations through a "length filter": explanations that consisted of at least 3 sentences were retained; shorter explanations were disposed of. This produced 87 explanations, of which 48 described objects and 39 described processes. Finally, to test an equal number of objects and processes, we randomly chose 30 objects and 30 process. So that we would not influence the selection process, a random number generator was used to choose the final explanations.

Two Panels of Domain Experts

To address the difficult problem of subjectivity, we assembled 12 domain experts, all of whom were graduate students in biology. To ensure the integrity of our results, we carefully maintained the following condition:

> None of the participants were informed about the purpose of the study.

Because we wanted to gauge KNIGHT's performance relative to humans, we assigned each of the experts to one of two panels: the *Writing* Panel and *Judging* Panel. By securing the services of such a large number of domain experts, we were able to form relatively large panels of four writers and eight judges (Figure 2). To ensure that the human-generated explanations would be of the highest possible quality, we assigned the four most experienced experts to the Writing Panel. The remaining eight experts were assigned to the Judging Panel to evaluate explanations.

To minimize the effect of factors that might make it difficult for judges to compare KNIGHT's explanations with those of domain experts, we took three precautions. First, we attempted to control for the length of explanations. Although we could not impose hard constraints, we made suggestions about how long a typical explanation might be. Second, to make the "level" of the explanations comparable, we asked writers to compose explanations for a particular audience, freshman biology students. Third, so that the general topics of discussion would be comparable, we asked judges to focus on anatomy, physiology, and development.

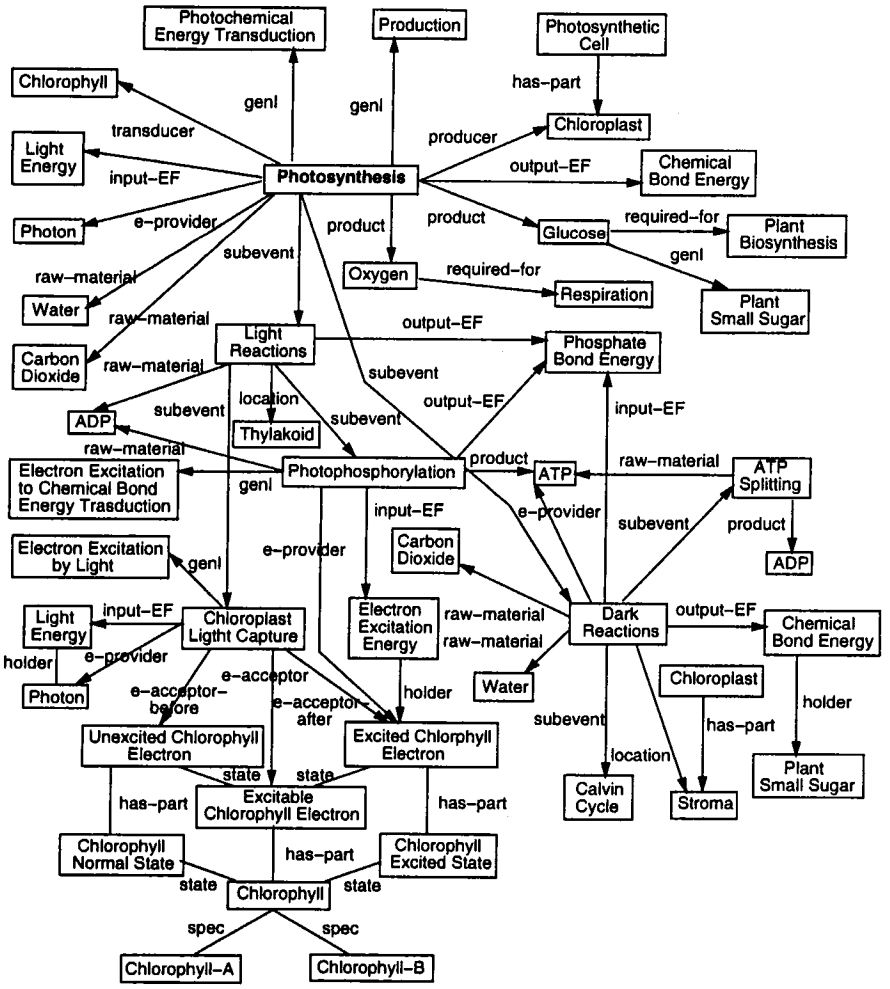

Figure 1: A Sample Representational Structure

Explanation Generation: Humans

To ensure that the difficulty of the concepts assigned to the writers were the same as those assigned to KNIGHT, the writers were given the task of explaining *exactly* the same set of concepts that KNIGHT had explained. Because we wanted to give writers an opportunity to explain both objects and processes, each writer was given an approximately equal number of objects and processes. Each of the 4 writers was given 15 concepts to explain, and each concept was assigned to exactly one writer. We then transcribed their handwritten explanations and put them and KNIGHT's explanations into an identical format. At this point, we had a pool of 120 explanations: sixty of these pertained to objects (30 written by biologists and 30 by KNIGHT), and the other sixty pertained to processes (also 30 written by biologists and 30 by KNIGHT).

Explanation Evaluation

We then submitted the explanations to the panel of eight judges. The judges were not informed of the source of the explanations, and all of the explanations appeared in the same format. Each judge was given fifteen explanations to evaluate. Judges were asked to rate the explanations on several dimensions: overall quality and coherence, content, organization, writing style, and correctness. To provide judges with a familiar rating scale, they were asked to assign letters grades (A, B, C, D, or F) to each dimension of the explanation.

Because carefully evaluating multiple dimensions of explanations is a labor-intensive task, time considerations required us to limit the number of explanations submitted to each judge. Hence, we assigned each judge 15 explanations, which on average required an

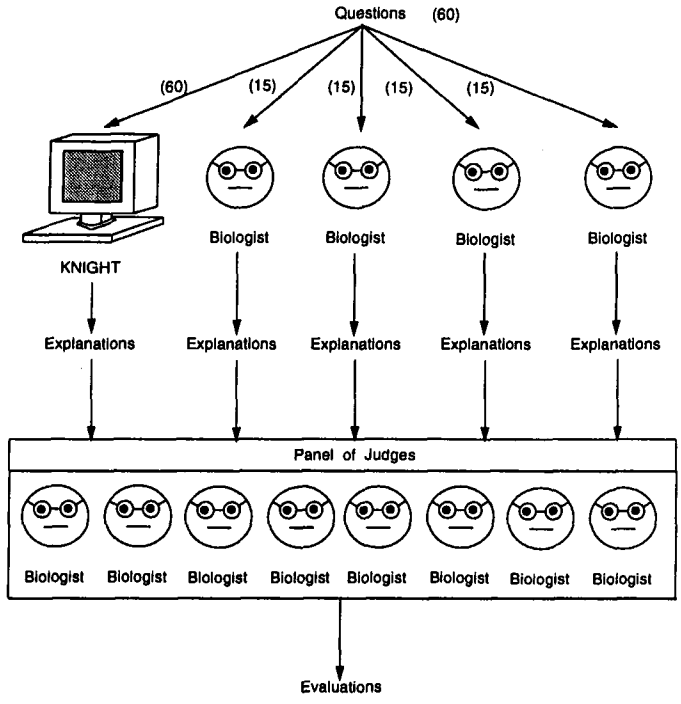

Figure 2: A Formal Empirical Evaluation

hour to evaluate. We assigned explanations to judges using an allocation policy that obeyed the following four constraints:

- Each judge received explanations that were approximately evenly divided between objects and processes.
- Each judge received explanations that were approximately evenly divided between those that were produced by KNIGHT and those that were produced by biologists.
- No judge received two explanations of the same concept.
- The explanations written by each writer were not evaluated by only one judge; rather, they were distributed to at least two judges.

It is important to emphasize again that the judges were not made aware of the purpose of the experiment, nor were told that any of the explanations were computer-generated.

Results

By the end of the study, we had amassed a large volume of data. To analyze it, we converted each of the "grades" to their traditional numerical counterparts, i.e., A=4, B=3, C=2, D=1, and F=0. Next, we computed means and standard errors for both KNIGHT's and the biologists' grades. We calculated these values for the overall quality and coherence rating, as well as for each of the subscores of content, organization, writing style, and correctness. On the overall rating and on each of the subscores, KNIGHT scored within approximately "half a grade" of the biologists (Table 1).[1]

Given these results, we decided to investigate the differences between KNIGHT's grades and the biologists' grades. When we normalized the grades by defining an "A" to be the mean of the biologists' grades, KNIGHT earned approximately 3.5 (a B^+). Comparing differences in subscores, KNIGHT performed best on correctness and content, not quite as well on writing style, and least well on organization.

Because the differences between KNIGHT and the bi-

[1] In the tables, $\pm$ denotes the standard error, i.e., the standard deviation of the mean.

ologists were narrow in some cases, we measured the statistical significance of these differences by running standard t-tests.[2] KNIGHT's grades on the content and correctness subscores did not differ significantly from the biologists' (Table 2). Of course, an insignificant difference does not indicate that KNIGHT's performance and the biologists' performance was equivalent—an even larger sample size might have shown a significant difference—however, it serves as an indicator that KNIGHT's performance approaches that of the biologists on these two dimensions.

To gauge how well KNIGHT generates explanations about objects—as opposed to processes—we computed means and standard errors for both KNIGHT's explanations of objects and the biologists' explanations of objects. We did the same for the explanations of processes. For both objects and processes, KNIGHT scored within "half a grade" of the biologists. Again, we measured the statistical significance of these differences. Although there was a significant difference between KNIGHT and biologists on explanations of processes, KNIGHT and the biologists did not differ significantly on explanations of objects (Tables 3 and 4).

As a final test, we compared KNIGHT to each of the individual writers. For a given writer, we assessed KNIGHT's performance relative to that writer in the following way: we compared the grades awarded to KNIGHT and the grades awarded to the writer on explanations generated in response to the same set of questions. This analysis produced some surprising results. Although there were substantial differences between KNIGHT and "Writer 1," KNIGHT was somewhat closer to "Writer 2," it was very close to "Writer 3," and its performance actually *exceeded* that of "Writer 4." KNIGHT and Writers 2, 3, and 4 did not differ significantly (Table 5).

Conclusions

That there was no significant difference between KNIGHT's explanations and the biologists' explanations on content and correctness is particularly striking. This indicates that the local and global content determination methods performed well. Moreover, it suggests that KNIGHT's "content" discourse knowledge and the domain knowledge in the Biology Knowledge Base are well represented. The lower grades on the organization subscore indicates that either the discourse knowledge is in need of revision, or that we need a more context-dependent approach to organization. The somewhat lower grades on the writing style subscore indicates that the realization system needs further work, which is undoubtedly the case.

These results call for further analysis and experimentation. A particularly intriguing kind of experiment is an *ablation* study, in which different aspects of the system are ablated (removed or degraded), and the effects of the ablation on the explanations are noted. By performing a series of these experiments, we can determine which aspects of KNIGHT and its representations contribute most significantly to its success.

In summary, it is very encouraging that KNIGHT scored within "half a grade" of the domain experts. Moreover, it is interesting that its performance actually *exceeded* that of one of the writers. These results demonstrate that an explanation generation system, which has been given a well represented knowledge base, can construct natural language responses whose quality approximates that of domain experts.

[2] All t-tests were unpaired, two-tailed. The results are reported for a 0.05 level of confidence.

References

Cawsey, A. 1992. *Explanation and Interaction: The Computer Generation of Explanatory Dialogues.* MIT Press.

Elhadad, M. 1992. *Using Argumentation to Control Lexical Choice: A Functional Unification Implementation.* Ph.D. Dissertation, Columbia University.

Lester, J. C., and Porter, B. W. 1991a. A revision-based model of instructional multi-paragraph discourse production. In *Proceedings of the Thirteenth Cognitive Science Society Conference*, 796–800.

Lester, J. C., and Porter, B. W. 1991b. A student-sensitive discourse generator for intelligent tutoring systems. In *Proceedings of the International Conference on the Learning Sciences*, 298–304.

Lester, J. C. 1994. *Generating Natural Language Explanations from Large-Scale Knowledge Bases.* Ph.D. Dissertation, The University of Texas at Austin, Austin, Texas.

Mittal, V. O. 1993. *Generating Natural Language Descriptions with Integrated Text and Examples.* Ph.D. Dissertation, University of Southern California.

Porter, B.; Lester, J.; Murray, K.; Pittman, K.; Souther, A.; Acker, L.; and Jones, T. 1988. AI research in the context of a multifunctional knowledge base: The botany knowledge base project. Technical Report AI Laboratory AI88-88, University of Texas at Austin, Austin, Texas.

Sundheim, B., ed. 1991. *Proceedings of the Third Message Understanding Evaluation Conference.* Los Altos, Calif.: Morgan Kaufmann.

Generator	*Overall*	*Content*	*Organization*	*Writing*	*Correctness*
KNIGHT	2.37±0.13	2.65±0.13	2.45±0.16	2.40±0.13	3.07±0.15
Human	2.85±0.15	2.95±0.16	3.07±0.16	2.93±0.16	3.16±0.15

Table 1: Comprehensive Analysis

	Overall	*Content*	*Organization*	*Writing*	*Correctness*
Difference	0.48	0.30	0.62	0.53	0.09
t statistic	-2.36	-1.47	-2.73	-2.54	-0.42
Significance	0.02	0.14	0.07	0.01	0.67
Significant?	Yes	No	Yes	Yes	No

Table 2: Differences and Significance

Generator	*Grade*
KNIGHT	2.65±0.19
Human	2.93±0.19
Difference	0.28
t statistic	-1.05
Significance	0.30
Significant?	No

Table 3: Explanation of Objects

Generator	*Grade*
KNIGHT	2.10±0.24
Human	2.77±0.17
Difference	0.67
t statistic	-2.23
Significance	0.03
Significant?	Yes

Table 4: Explanation of Processes

KNIGHT	*vs. Writer 1*	*vs. Writer 2*	*vs. Writer 3*	*vs. Writer 4*
KNIGHT	1.93±0.29	2.73±0.23	2.73±0.27	2.07±0.23
Human	3.60±0.16	3.40±0.23	2.80±0.28	1.60±0.23
Difference	1.67	0.67	0.07	0.47
t statistic	-5.16	-2.03	-0.17	1.42
Significance	0.00	0.05	0.86	0.16
Significant?	Yes	No	No	No

Table 5: KNIGHT vs. Individual Writers