Scaling Up Explanation Generation: Large-Scale Knowledge Bases and Empirical Studies*

James C. Lester
Department of Computer Science
North Carolina State University
Raleigh, NC 27695–8206
lester@eos.ncsu.edu

Bruce W. Porter
Department of Computer Sciences
The University of Texas at Austin
Austin, Texas 78712–1188
porter@cs.utexas.edu

Abstract

To explain complex phenomena, an explanation system must be able to select information from a formal representation of domain knowledge, organize the selected information into multi-sentential discourse plans, and realize the discourse plans in text. Although recent years have witnessed significant progress in the development of sophisticated computational mechanisms for explanation, empirical results have been limited. This paper reports on a seven year effort to empirically study explanation generation from semantically rich, large-scale knowledge bases.

We first describe Knight, a robust explanation system that constructs multi-sentential and multi-paragraph explanations from the Biology Knowledge Base, a large-scale knowledge base in the domain of botanical anatomy, physiology, and development. We then introduce the Two Panel evaluation methodology and describe how Knight's performance was assessed with this methodology in the most extensive empirical evaluation conducted on an explanation system. In this evaluation, Knight scored within "half a grade" of domain experts, and its performance exceeded that of one of the domain experts.

Introduction

In the course of their daily affairs, scientists explain complex phenomena—both to one another and to lay people—in a manner that facilitates clear communication. Similarly, physicians, lawyers, and teachers are equally facile at generating explanations in their respective areas of expertise. In an effort to computationalize this critical ability, research in natural language generation has addressed a broad range of issues in automatically constructing text from formal representations of domain knowledge. Research on text planning (McKeown 1985; Paris 1988; McCoy 1989 1990; Hovy 1993; Maybury 1993) has developed techniques for determining the content and organization of many genres, and explanation generation (Cawsey 1992; Moore 1995) in particular has been the subject of intense investigation. In addition to exploring a panorama of application domains, the explanation community has begun to assemble these myriad designs into a coherent framework. As a result, we have begun to see a crystallization of the major components (Suthers 1993).

Despite this success, empirical results in explanation generation are limited. Although techniques for developing and evaluating robust explanation generation should yield results that are more conclusive than those produced by prototype, "proof-of-concept" systems, with only a few notable exceptions (Kukich 1983; Hovy 1990; Cawsey 1992; Mittal 1993; Robin 1994), most work has adopted a research methodology in which a proof-of-concept system is constructed and its operation is analyzed on a few examples. While isolating one or a small number of problems enables researchers to consider particular issues in detail, it is difficult to gauge the scalability and robustness of a proposed approach. A critical factor contributing to the dearth of empirical results is the absence of semantically rich, large-scale knowledge bases. Knowledge bases housing tens of thousands of different concepts could furnish ample raw materials for empirical study, but no work in explanation generation has been conducted or empirically evaluated in the context of these knowledge bases.

To empirically study explanation generation from semantically rich, large-scale knowledge bases, we undertook a seven year experiment. First, our domain experts (one employed full-time) constructed the Biology Knowledge Base (Porter *et al.* 1988), a very large structure representing more than 180,000 facts about botanical anatomy, physiology, and development. Second, we designed, implemented, and empirically evaluated KNIGHT, a robust explanation system that extracts information from the Biology Knowledge Base, organizes it, and realizes it in multi-sentential and multi-paragraph expository explanations of complex

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

biological phenomena. Third, we developed a novel evaluation methodology for gauging the effectiveness of explanation systems and employed this methodology to evaluate KNIGHT. This paper describes the lessons learned during the course of the "KNIGHT experiments."[1]

The Task of Explanation Generation

Explanation generation is the task of extracting information from a formal representation of knowledge, imposing an organization on it, and realizing the information in text. The overall task is typically decomposed into two subtasks, *explanation planning* and *realization*. Explanation planning itself has two subtasks: *content determination*, in which knowledge structures are extracted from a knowledge base, and *organization*, in which the selected knowledge structures are arranged in a manner appropriate for communication in natural language. To communicate complex ideas, an explanation system should be able to produce multisentential explanations such as the one in Figure 1, which shows several explanations from the domain of botanical anatomy, physiology, and development. To perform these tasks successfully, an explanation planner must have access to *discourse knowledge*, which informs its decisions about the content and organization of textual explanations. The organizational aspect of discourse knowledge plays a particularly important role in the construction of extended explanations.

> Embryo sac formation is a kind of female gametophyte formation. During embryo sac formation, the embryo sac is formed from the megaspore mother cell. Embryo sac formation occurs in the ovule.
>
> Embryo sac formation is a step of angiosperm sexual reproduction. It consists of megasporogenesis and embryo sac generation. During megasporogenesis, the megaspore mother cell divides in the nucellus to form 4 megaspores. During embryo sac generation, the embryo sac is generated from the megaspore.

Figure 1: Explanation produced by KNIGHT from the Biology Knowledge Base

Evaluating the performance of explanation systems is a critical and non-trivial problem. Five evaluation criteria should be applied.

- *Coherence*: A global assessment of the overall quality of explanations generated by a system.
- *Content*: The extent to which the explanation's information is adequate and focused.
- *Organization*: The extent to which the information is well organized.
- *Writing style*: The quality of the prose.
- *Correctness*: For scientific explanations, the extent to which the explanations are in accord with the established scientific record.

In addition to performing well on the evaluation criteria, if explanation systems are to make the difficult transition from research laboratories to fielded applications, they should exhibit two important properties, both of which significantly affect scalability. First, these systems' representations of discourse knowledge should be easily inspected and modified. To develop explanation systems for a broad range of domains, tasks, and question types, discourse-knowledge engineers must be able to create and efficiently debug the discourse knowledge that drives the systems' behavior. The second property that explanation systems should exhibit is robustness. Despite the complex and possibly mal-formed representational structures that an explanation system may encounter in its knowledge base, it should be able to cope with these structures and construct reasonably well-formed explanations.

Given the state of the art in explanation generation, the field is now well positioned to explore what may pose its greatest challenge and at the same time may result in its highest payoff: generating explanations from semantically rich, large-scale knowledge bases. Large-scale knowledge bases encode information about domains that cannot be reduced to a small number of principles or axioms. These knowledge bases consist of highly interconnected networks of (at least) tens of thousands of facts.

One such knowledge base is the **Biology Knowledge Base**, a large structure that encodes information about botanical anatomy, physiology, and development. One of the largest knowledge bases in existence, it is encoded in the KM frame-based knowledge representation language. The backbone of the Biology Knowledge Base is its taxonomy, which is a large hierarchical structure of biological objects and biological processes. The Biology Knowledge Base currently contains more than 180,000 explicitly represented triples, and its deductive closure is significantly larger.

It is important to note that the authors and the domain experts entered into a "contractual agreement" with regard to representational structures in the Biology Knowledge Base. To eliminate all requests for representational modifications that would skew the knowledge base to the task of explanation generation, the authors entered into the following agreement: they could request representational changes only if knowledge was inconsistent or missing. This facilitated a unique experiment in which the representational structures were not tailored for the task of explanation generation.

Accessing Large-Scale Knowledge Bases

By interposing a KB accessing system between an explanation planner, which performs global content determination, and a knowledge base, it is possible to keep an explanation planner at "arm's length" from

[1] Details of KNIGHT's architecture, implementation, and evaluation may be found in (Lester 1994).

the representation of domain knowledge. In addition, it can help build explanations that are coherent. An important technique for generating coherent explanations is by extracting *views* (McCoy 1989 1990; Suthers 1993). For example, the concept *photosynthesis* can be viewed as either a *production* process or an *energy transduction process*. Viewed as *production*, it would be described in terms of its *raw materials* and *products*: "During photosynthesis, a chloroplast uses water and carbon dioxide to make oxygen and glucose." Viewed as *energy transduction*, it would be described in terms of *input energy forms* and *output energy forms*: "During photosynthesis, a chloroplast converts light energy to chemical bond energy." In short, the view that is taken of a concept has a significant effect on the content that is selected for its description.

KNIGHT has a robust KB accessing system that extracts views of concepts represented in a knowledge base. Each view is a coherent subgraph of the knowledge base describing the structure and function of objects, the change made to objects by processes, and the temporal attributes and temporal decompositions of processes. Each of the nine accessors in the library can be applied to a given concept—the "concept of interest"—to retrieve a view of that concept. There are three classes of Accessors: those that are applicable to all concepts (*As-Kind-Of* and *Functional*), those that are applicable to objects (*Partonomic-Connection* and *Sub-Structural*), and those that are applicable to processes (*Auxiliary-Process*—which includes *Causal*, *Modulatory*, *Temporal*, and *Locational* sub-types—*Participants*, *Core-Connection*, and *Sub-event*, and *Temporal-Step*).[2]

In addition to coherence, robustness is also an important design criterion. The KB accessors achieve robust performance in four ways: (1) They do not assume that essential information will actually appear on a given concept in the knowledge base. (2) They employ a type checking system that exploits the knowledge base's taxonomy. (3) When they detect an irregularity, they return appropriate error codes to the explanation planner. (4) They tolerate specialized (and possibly unanticipated) representational vocabulary by exploiting the relation taxonomy. By using these techniques in tandem, we have developed a KB accessing system that has constructed several thousand views without failing.

A Discourse Programming Language

The "discourse-knowledge programming language" of *Explanation Design Packages* (EDPs) emerged from an effort to accelerate the representation of discourse knowledge without sacrificing expressiveness. For a given query type, domain, and task, a discourse-knowledge engineer must be able to represent the discourse knowledge needed by an explanation system for responding to questions of that type in that domain about that task. Pragmatically, to represent discourse knowledge for a broad range of queries, domains, and tasks, a formalism must facilitate *efficient* representation of discourse knowledge. Therefore, important goals for the design of a discourse formalism are ease of reuse and ease of modification.

EDPs give discourse-knowledge engineers the proper abstractions for specifying the content and organization of explanations. They combine a frame-based representation language with embedded procedural constructs. To mirror the structure of expository texts, an EDP contains a hierarchy of nodes, which provides the "global organization" of explanations. EDPs are schema-like (McKeown 1985; Paris 1988) structures that include constructs found in traditional programming languages. Just as prototypical programming languages offer conditionals, iterative control structures, and procedural abstraction, EDPs offer discourse-knowledge engineers counterparts of these constructs that are precisely customized for explanation planning. Moreover, each EDP names multiple KB accessors, which are invoked at explanation planning time. Because EDPs are frame-based and are implemented in KM, the representational language used by the Biology Knowledge Base, they can be easily viewed and edited by knowledge engineers using the graphical tools commonly associated with frame-based languages. This has proven to be very useful for addressing a critical problem in scaling up explanation generation: maintaining a knowledge base of discourse knowledge that can be easily constructed, viewed, and navigated by discourse-knowledge engineers.

A discourse-knowledge engineer can use EDPs to encode discourse knowledge for his or her application. In our work, we have focused on two types of texts that occur in many domains: *process descriptions* and *object descriptions*. For example, in biology, one encounters many process-oriented descriptions of physiological and reproductive mechanisms, as well as many object-oriented descriptions of anatomy. In the course of our research, we studied many passages in biology textbooks. These passages focused on explanations of the anatomy, physiology, and reproduction of plants. We manually "parsed" each passage into a discourse tree. The discourse trees were expressed in an informal language centering around viewpoints (Acker *et al.* 1991; Suthers 1993). The viewpoints were in turn expressed in an informal language of structure, function, and process which is commonly found in the discourse literature, e.g., (McKeown 1985; Paris 1988). Our final step was to generalize the most commonly occurring patterns into abstractions that covered as many aspects of the passages as possible.

[2] In addition to the "top level" accessors, the library also provides a collection of some twenty "utility" accessors. These include procedures for extracting particular aspects of views previously constructed by the system.

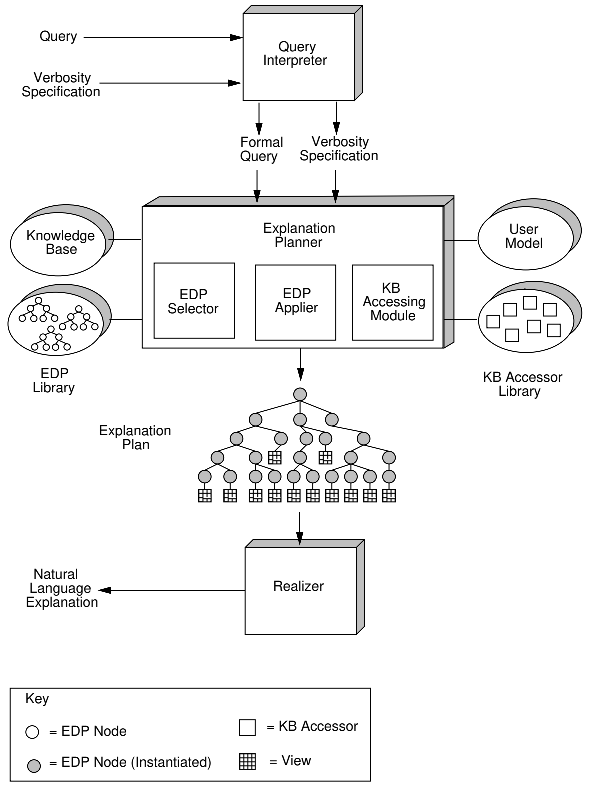

Figure 2: An Architecture for Explanation Generation

After generalizing the commonly occurring patterns into abstractions, we encoded the abstractions in two Explanation Design Packages. These EDPs can be used by an explanation planner to generate explanations about the processes and objects of physical systems.

Planning Explanations

We have designed an architecture for explanation generation and implemented a full-scale explanation generator, KNIGHT,[3] that is based upon this architecture.[4] Explanation generation begins when the user poses a query, which includes a verbosity specification that comes in the form of a qualitative rating expressing the desired length of the explanation (Figure 2). The query interpreter—whose capabilities have been addressed only minimally in our work—translates the query to a canonical form, which is passed, along with the verbosity specification, to the explanation planner.

The heart of an explanation generator is its explanation planner. The explanation planner invokes the EDP Selector, which chooses an Explanation Design Package from the EDP Library. The explanation planner then applies the EDP by conducting an in-order traversal of its hierarchical structure. As the plan is constructed, the explanation planner updates the user model to reflect the contextual changes that will result from explaining the views in the explanation plan, attends to the verbosity specification, and invokes KB Accessors to extract information from the knowledge base. Recall that the Accessors return "views," which are subgraphs of the knowledge base. The planner attaches the views to the explanation plan; they become the plan's leaves. Planning is complete when the explanation planner has traversed the entire EDP.

The planner passes the resulting explanation plan to a realization component for translation to natural language. The realizer, FARE (Callaway & Lester 1995), is built on top of a unification-based surface generator with a large systemic grammar (Elhadad 1991). Explanation generation terminates when FARE has translated all of the views in the explanation plan to natural language.

Example Behavior

To illustrate the behavior of the system, consider the concept of *embryo sac formation*. Figure 3 depicts the semantic network in the Biology Knowledge Base that represents information about *embryo sac formation*. When KNIGHT is given the task of explaining this concept,[5] it applies the *Explain-Process* EDP. KNIGHT first finds the topics of the *Explain Process* exposition node, which are *Process Overview*, *Output Actor Fates*, *Temporal Information*, and *Process Details*. During its traversal of this tree, it begins with *Process Overview*, which has a `High` centrality rating and an inclusion condition of `True`. KNIGHT executes the COMPUTE INCLUSION algorithm with the given verbosity of `High`, which returns `True`, i.e., the information associated with the topic should be included.

Hence, it now begins to traverse the children of this topic node, which are the *As Kind Of Process Description*, *Process Participants*, and *Location Description* content specification nodes. For the *As Kind Of Process Description*, it computes a value for the local variable *?Reference Concept*, which returns the value *female gametophyte formation*. It then instantiates the content specification template on *As Kind Of Process Description*, which it then evaluates. This results in a call to the *As-Kind-Of* KB Accessor, which produces a view. Similarly, KNIGHT instantiates the content specification expressions of *Process Participants Description* and *Location Description*, which also cause KB Accessors to be invoked; these also return views. Next KNIGHT visits the *Location Partonomic Connection* node, which is an elaboration of *Location Description*. However, because its inclusion condition is

[3] All of the explanation planning algorithms, as well as the KB Accessors, were implemented in Lucid Common Lisp on a DEC Station 5000.

[4] See (Lester & Porter 1991) for a discussion of KNIGHT's approach to user modeling.

[5] In this example, KNIGHT is given a HIGH verbosity specification.

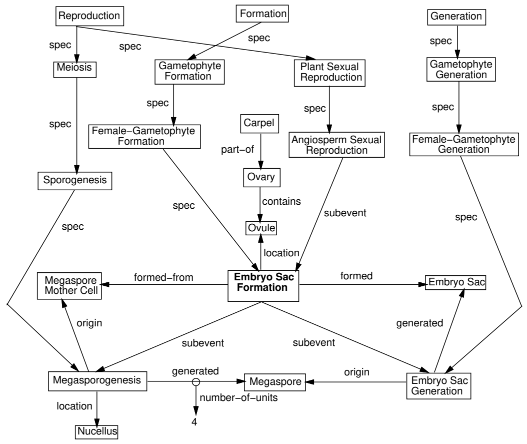

Figure 3: A representation of embryo sac formation

not satisfied, this branch of the traversal halts. Next, KNIGHT visits each of the other topics of the *Explain Process* exposition node: *Temporal Information* and *Process Details*. Because it was given a `High` verbosity specification, both of these are used to determine additional content. When the views in the final explanation plan are translated to text by the realization system, KNIGHT produces the text shown in Figure 1.

Evaluation

How can one evaluate the architectures, algorithms, and knowledge structures that form the basis for an explanation generator? To address these problems, we have developed the **Two-Panel Evaluation Methodology**. To ensure the integrity of the evaluation results, a central stipulation of the methodology is that the following condition be maintained throughout a study:

> **Computer Blindness:** None of the participants can be aware that some texts are machine-generated or, for that matter, that a computer is in any way involved in the study.

Experimental Design

The Two-Panel Evaluation Methodology involves four steps: (1) generation of explanations by computer; (2) formation of two panels of domain experts; (3) generation of explanations by one panel of domain experts; and (4) evaluation of all explanations by second panel of domain experts.

Explanation Generation: Knight. Because KNIGHT's operation is initiated when a user poses a question, the 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., *Object*. 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 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.

Two Panels of Domain Experts. To address the difficult problem of subjectivity, we assembled 12 domain experts, all of whom were PhD students and post-doctoral scientists in biology. 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 the *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 4). 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 writers to focus on anatomy, physiology, and development.

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

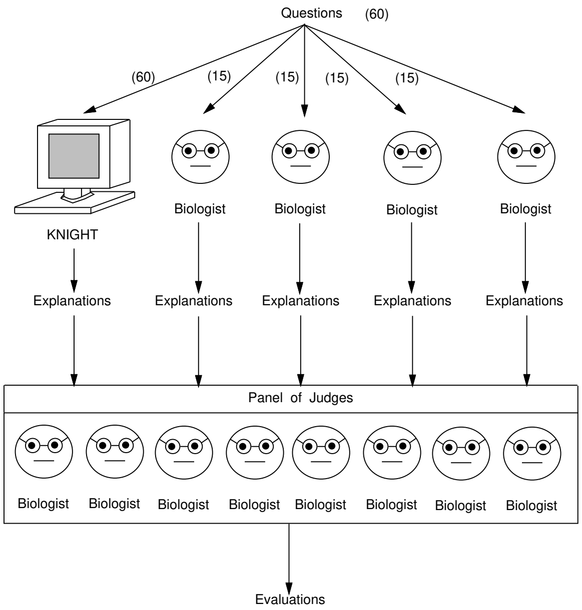

Figure 4: The Two-Panel Methodology in the KNIGHT Experiments

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 explanation on each of the dimensions. 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 hour to evaluate. We assigned explanations to judges using an allocation policy that obeyed the following four constraints: (1) Each judge received explanations that were approximately evenly divided between those that were produced by KNIGHT and those that were produced by biologists. (2) Each judge received explanations that were approximately evenly divided between objects and processes. (3) No judge received two explanations of the same concept. (4) 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, etc. 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 dimensions of content, organization, writing style, and correctness. On the overall rating and on each of the dimensions, KNIGHT scored within approximately "half a grade" of the biologists (Table 1).[6]

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 dimensions, 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 biologists were narrow in some cases, we measured the statistical significance of these differences by running standard t-tests.[7] KNIGHT's grades on the content, organization, and correctness dimensions 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 three dimensions.

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. 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 3).

Related Work

By synthesizing a broad range of research on knowledge base access (McCoy 1989 1990; Suthers 1993),

[6] In the tables, ± denotes the standard error, i.e., the standard deviation of the mean.

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

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	No	Yes	No

Table 2: Differences and Significance

schemata (McKeown 1985; Paris 1988; McCoy 1989 1990), and top-down discourse planners (Cawsey 1992; Suthers 1993; Hovy 1993; Moore 1995). KNIGHT provides a "start-to-finish" solution to the problem of automatically constructing expository explanations from semantically rich, large-scale knowledge bases. Perhaps its most important contribution lies in its evaluation methodology. With regard to evaluation, it is perhaps most closely related to five NLG projects that have been empirically evaluated: PAULINE (Hovy 1990), EDGE (Cawsey 1992), the EXAMPLE GENERATOR (Mittal 1993), ANA (Kukich 1983), and STREAK (Robin 1994). 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.); nevertheless, it is a significant achievement in terms of evaluation *scale* because of the sheer number of texts it produced. In a second landmark evaluation, Cawsey undertook a study in which subjects were allowed to interact with her explanation generation system, EDGE. Cawsey analyzed the system's behavior as the dialogs progressed, interviewed subjects, and used the results to revise the system. Although EDGE was not subjected to a tightly controlled, formal evaluation, it was sufficiently robust to be used interactively by eight subjects.

The EXAMPLE GENERATOR, ANA, and STREAK were each subjected to formal (quantitative) evaluations. Mittal and Paris developed and formally evaluated a generator that produced descriptions integrating text and examples. Rather than evaluating the explanations directly, subjects were given a quiz about the concept under consideration.[8] The degree to which the experiments controlled for specific factors, e.g., the effect of example positioning, example types, example complexity, and example order, is remarkable. ANA and STREAK were both subjected to quantitative, corpus-based evaluations. Kukich employed a corpus-based methodology to judge the coverage of ANA's knowledge structures. STREAK was evaluated with a corpus-based study that produced estimates of its sub-language coverage, extensibility, and the overall effectiveness of its revision-based generation techniques. Although neither of these studies employed human judges to critique text quality, the rigor with which they were conducted has significantly raised the standards for evaluating generation systems.

To summarize, KNIGHT is the only system to have been evaluated in the context of a semantically rich, large-scale knowledge base. It is also the only system to have been evaluated in a kind of restricted "Turing test" in which the quality of its text was evaluated by humans in a head-to-head comparison against the text produced by humans (domain experts) in response to the same set of questions.

Conclusion

Explanation generation is an exceedingly complex task that involves a diversity of interacting computational mechanisms. To investigate the issues and problems of generating natural language explanations from semantically rich, large-scale knowledge bases, we have designed and implemented KNIGHT, a fully functioning explanation system that automatically constructs multi-sentential and multi-paragraph natural language explanations. This work has demonstrated that (1) separating out knowledge-base access from explanation planning can enable the construction of a robust system that extracts coherent views from a a semantically rich, large-scale knowledge base; and (2) Explanation Design Packages, a hybrid representation of discourse knowledge that combines a frame-

[8] In a second analysis *without human judges*, the system developers compared selected features of the EXAMPLE GENERATOR's output with text from textbook and obtained encouraging results.

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 3: KNIGHT vs. Individual Writers

based representation with procedural constructs, facilitate the iterative refinement of discourse knowledge.

To gauge the effectiveness of these techniques, we developed the Two-Panel Evaluation Methodology and employed it in the evaluation of KNIGHT. KNIGHT scored within "half a grade" of the biologists. There was no significant difference between KNIGHT's explanations and the biologists' explanations on measures of content, organization, and correctness, nor was there a statistically significant difference in overall quality between KNIGHT's explanations and those composed by three of the biologists. KNIGHT's performance exceeded that of one of the biologists. These findings demonstrate that an explanation system that has been given a well represented knowledge base can construct natural language responses whose quality approximates that of humans.

Acknowledgements

We would like to thank our principle domain expert, Art Souther, for leading the knowledge base construction effort; Charles Callaway and the NLG students for their work on the realization system; Erik Eilerts, for building the knowledge base editing tools; Michael Elhadad, for generously assisting us with FUF; Peter Clark and Charles Callaway for helpful comments on previous drafts of this paper; Dan Suthers for insights on the problems of evaluating explanation systems; and the other members of the Biology Knowledge Base Project: Liane Acker, Brad Blumenthal, Rich Mallory, Ken Murray, and Jeff Rickel.

References

Acker, L. H.; Lester, J. C.; Souther, A. F.; and Porter, B. W. 1991. Generating coherent explanations to answer students' questions. In Burns, H.; Parlett, J.; and Redfield, C., eds., *Intelligent Tutoring Systems: Evolutions in Design*. Hillsdale, New Jersey: Lawrence Earlbaum. 151–176.

Callaway, C. B., and Lester, J. C. 1995. Robust natural language generation from large-scale knowledge bases. In *Proceedings of the Fourth Bar-Ilan Symposium on Foundations of Artificial Intelligence*, 96–105.

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

Elhadad, M. 1991. FUF: The universal unifier user manual version 5.0. Technical Report CUCS-038-91, Department of Computer Science, Columbia University.

Hovy, E. H. 1990. Pragmatics and natural language generation. *Artificial Intelligence* 43:153–197.

Hovy, E. H. 1993. Automated discourse generation using discourse structure relations. *Artificial Intelligence* 63:341–385.

Kukich, K. 1983. *Knowledge-Based Report Generation: A Knowledge Engineering Approach to Natural Language Report Generation*. Ph.D. Dissertation, University of Pittsburgh.

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

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

Maybury, M. T. 1993. Communicative acts for generating natural language arguments. In *Proceedings of the Eleventh National Conference on Artificial Intelligence*, 357–364.

McCoy, K. F. 1989 1990. Generating context-sensitive responses to object-related misconceptions. *Artificial Intelligence* 41:157–195.

McKeown, K. R. 1985. *Text Generation: Using Discourse Strategies and Focus Constraints to Generate Natural Language Text*. Cambridge University Press.

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

Moore, J. D. 1995. *Participating in Explanatory Dialogues*. MIT Press.

Paris, C. L. 1988. Tailoring object descriptions to a user's level of expertise. *Computational Linguistics* 14(3):64–78.

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.

Robin, J. 1994. *Revision-Based Generation of Natural Language Summaries Providing Historical Background*. Ph.D. Dissertation, Columbia University.

Suthers, D. D. 1993. *An Analysis of Explanation and Its Implications for the Design of Explanation Planners*. Ph.D. Dissertation, University of Massachusetts.