Automated Index Generation for Constructing Large-scale Conversational Hypermedia Systems*

Richard Osgood and **Ray Bareiss**
The Institute for the Learning Sciences
Northwestern University
Evanston, Illinois 60201
osgood@ils.nwu.edu and bareiss@ils.nwu.edu

Abstract

At the Institute for the Learning Sciences we have been developing large scale hypermedia systems, called ASK systems, that are designed to simulate aspects of conversations with experts. They provide access to manually indexed, multimedia databases of story units. We are particularly concerned with finding a practical solution to the problem of finding indices for thes units when the database grows too large for manual techniques. Our solution is to provide automated assistance that proposes relative links between units, eliminating the need for manual unit-to-unit comparison. In this paper we describe eight classes of links, and show a representation and inference procedure to assist in locating instances of each.

Introduction

Interaction with a knowledge-based system typically provides a user with only limited information. For example, a diagnostic system typically returns a classification in response to a sequence of situational features. If an explanation is provided, it is usually a trace of the system's inference process. In contrast, consultation with a human expert typically provides a wealth of information. An expert knows which questions to ask in a problem solving situation, why those questions are important, which questions not to ask, how to interpret and justify the actual results, alternative methods of data collection, *et cetera*. Unfortunately, these aspects of expertise have proven difficult to represent with current AI formalisms. As a practical alternative, builders of knowledge-based systems have turned to hypermedia to capture such knowledge in a partially represented form [Spiro and Jehng, 1990].

For the last three years, we have been developing a class of large-scale hypermedia systems called ASK systems [Ferguson *et al.*, 1992], that are designed to capture important aspects of a conversation with an expert. An ASK system provides access to a multimedia database containing short video clips of interviews with experts, archival video material, and text passages. Currently, these systems are indexed in two ways. ASK systems can be built by human "indexers" (our term for knowledge engineers) who use a question-based methodology and some supporting tools to create relative links between pieces of the material [Osgood and Bareiss, 1992]. Our experience shows that as the size of the system's database grows beyond about 100 stories, (depending on the degree of interrelatedness) the process of identifying relevant connections between stories becomes prohibitively difficult for indexers. (The term *story* refers to an individual content unit in the database and is not limited to the traditional narrative sense.) We call this phenomenon the *indexer saturation problem:* an indexer cannot remember enough about the contents of the database to make all appropriate connections, and the prospect of exhaustive search for all connections is onerous [Conklin, 1987]).

The second way in which the problem arises is when authors must index their own stories. School Stories is a collaborative hypermedia authoring environment for telling and interconnecting stories about grade K-12 experiences in US public schools. There is no separate indexer role in the system. Authors notice a connection between a story in the system and one they know. This new story is entered into the system and linked directly by its author to the eliciting story at the point it is told. Unfortunately, no easy way exists for an author to find links between a new story and the rest of the database.

We are beginning to provide automated assistance to achieve more complete interconnectivity in all our ASK systems than is possible with our current manual in-

*This research was supported, in part, by the Defense Advanced Research Projects Agency, monitored by the Air Force Office of Scientific Research under contract F49620-88-C-0058 and the Office of Naval Research under contract N00014-90-J-4117, by the Office of Naval Research under contract N00014-89-J-1987, and by the Air Force Office of Scientific Research under contract AFOSR-89-0493. The Institute for the Learning Sciences was established in 1989 with the support of Andersen Consulting, part of The Arthur Andersen Worldwide Organization. The Institute receives additional support from Ameritech and North West Water which are Institute partners, and from IBM.

dexing methods. The contents of each story are represented as input to a computerized search process which compares simple representations of the input story to that of other stories in the story base and proposes connections between them to an indexer or author. Although fully automated indexing of stories would be ideal, we do not believe it to be practical, given the current state of the art of knowledge representation. It will require a more complete representation of story content as well as large amounts of commonsense knowledge to infer automatically the same set of the connections typically made by human indexers.

Given our desire to build a practical tool today, we have decided to employ a partial representation of story contents and very limited auxiliary knowledge. The cost of this decision is the requirement to keep a skilled human "in the loop", to determine the relevance of proposed links, and to maintain a story representation that can be easily processed by both machines and humans (see, *e.g.*, semiformal knowledge structures [Lemke and Fischer, 1990]). This decision balances the strengths of humans (*e.g.*, feature extraction and categorization) and computers (*e.g.*, rapid search and record keeping), enabling us to build a useful tool and solve a problem intractable to either machine or human alone.

The remainder of this paper discusses the ASK model of hypermedia, our representation of stories, the specific procedures for inferring links between stories, and our ongoing research.

The ASK Model of Hypermedia

ASK systems are based on a simple theory of the memory organization that might underlie conversation about problem solving.[Schank, 1977; Ferguson *et al.*, 1992]. This general theory argues that coherence in a conversation comes from the connectivity of human memory, *i.e.*, there is alignment between expression and thought (see *e.g.*, [Chafe, 1979]). We hypothesize that after hearing a piece of information in such a conversation, there are only a few general categories of follow-up information that represent a natural continuation of the thread of the conversation rather than a major topic shift. The categories can be thought of the poles of four axes or dimensions. These eight poles represent the most general kinds of questions that a user is likely to have in a conversation about problem solving. The browsing interface of an ASK system reifies this model of conversation by placing each relative link between stories in one of these eight general categories [Ferguson *et al.*, 1992]. Users can find their specific questions in the category that best describes the question.

The four dimensions are *Refocusing, Causality, Comparison,* and *Advice*. The *Refocusing* dimension concerns both adjustments to the specificity of topic under consideration as well as relevant digressions like clarifying of the meanings of terms or describing situations in which the topic arises. One pole, **Context,** points to the big picture within which a piece of information fits. The other, **Specifics,** points to examples of a general principle, further details of a situation, definitions of terms, or descriptions of parts of the whole, *et cetera*.

The *Causality* dimension arises directly out of the human desire to understand a situation in terms of its antecedents and consequences. We group temporal order and the causal chain because people typically collapse the distinction. The **Causes (or earlier events)** pole points to how a situation developed. The **Results (or later events)** pole points to the outcome of a situation.

The *Comparison* dimension concerns questions of similarity and difference, analogy and alternative, at the same level of abstraction as the reference story. The pole, **Analogies,** points to similar situations from other contexts or from the experiences of other experts. The **Alternatives** pole points to different approaches that might have been taken in a situation or differences of opinion between experts.

Finally, the *Advice* dimension captures the idea of carrying away a lesson, either negative or positive, for use in the problem solver's situation. The **Opportunities** pole points to advice about things a problem solver should capitalize upon in a situation. The **Warnings** pole points to advice about things that can go wrong in a problem solving situation.

The Partial Representation of Stories

Our approach to devising a representation for stories has been to provide a domain-independent representational frame that is instantiated with domain-specific fillers (Figure 1). A primary purpose of the frame is to enforce consistency of feature selection by an indexer. The representation is simple, indexical, and natural for human indexers to employ. It is just detailed enough to support the types of inference needed to recognize relationships between stories. In this and subsequent sections, we will describe a model of naive intentionality expressed in this frame structure and inference procedures specific to the conversational categories. We will offer examples of each from the School Stories application.

Because all of the stories of interest in the School Stories domain (K-12 school experiences) concern human intentional behavior, our representation is based upon the intentional chain [Schank and Abelson, 1975]. This is the simple model implicit in the design of the upper section of the frame shown in Figure 1. First, agents play roles and have beliefs that influence their selection of a course of action. Second, to play out those roles, agents establish goals and plans to achieve them. Finally, actions based on those plans and goals yield both intended and unintended results.

When representing a story, an indexer must instantiate the slots of this domain-independent frame with

AgentRole: *athlete*—the role the agent plays in the story	
BeliefType: *strong doesn't mean dumb*—the agent's belief inducing the goal	
IntentionLevel: *actually did*—the level of intentionality(goal, plan, or act)	
IntentionType: *get good grades*—the goal, plan or action of an agent	
OutcomeTypes: *positive emotional*—the results of the IntentionType	
SituationType: *conflict with others*—a name linking multiple interacting frames	
TimeOfOccurrence: *after reference*—sequencing information for frames	
StoryType: *literal example*—story application information	

Figure 1: A representational frame for describing one scene of a story

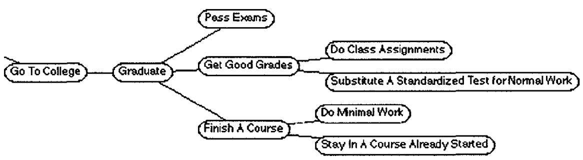

Figure 2: IntentionType Slot Fillers Near *Get Good Grades*

fillers representing the key domain concepts of the story. To achieve representational consistency, fillers are chosen from pre-enumerated taxonomies—one for each slot. Each filler exists in a domain specific hierarchy. The semantics of the hierarchies are intentional for the **IntentionType** slot, for example, *getting good grades* is a way to *graduate* (Figure 2) and categorical for the rest, *e.g.,* for the **AgentRole** slot, *a teacher without leverage* is a kind of *teacher.* Figure 1 also shows examples of fillers drawn from the School Stories domain.

A priori enumeration of all slot fillers is not intended. Rather our idea is to provide an indexer-extensible set for which the initial enumeration serves as an example. Indexers can enter a new term in the hierarchy by determining its similarity to pre-existing fillers. Assessment of the similarity of fillers during representation works because it is conducted by indexers in the target system's task context—the same one in which they would have judged the appropriateness of hand-crafted relative links between stories. In effect, the similarity of concepts is represented in equivalence classes, not computed from features [Porter, 1989], *i.e.,* similar concepts have a common parent in the hierarchy. To infer links, these hierarchies of equivalence classes are processed by inference procedures described in the next section.

The representational frame or *scene* captures the intentionality of a single agent. The upper portion the Figure 1 frame says: *an athlete actually did get good grades* by believing that *being strong doesn't mean being dumb* and this had a *positive emotional* impact on him/her.

In the frame's lower part in Figure 1 we include three additional slots. The **SituationType** slot functions both to group frames together and to describe the kind of agent interaction in those frames enabling the reprepresentation of interactions among multiple agents, sometimes with conflicting goals [Schank and Osgood, 1991]. Indexers employ multiple frames—one or more for each agent, filling just the slots in each that they feel apply as in Figure 3. For example, a situation about how to handle student boredom is captured by selecting *Being Bored* to fill the **SituationType** slots of two frames of the same story, one about a *Student* who *Shows Lack of Interest* and the other about a *Teacher* who *Assigns An Independent Activity.*

The frame representation deliberately overspecifies situations. This makes feasible inferences of the same type at two different levels of abstraction. For example, similarity between stories can be assessed at the level of an entire situation through the fillers of the **SituationType** slot. Similarity can also be assessed between stories at the level of agent activity through fillers of the top section of the frame in Figure 1.

The **TimeOfOccurrence** slot supports sequencing of scenes in stories to establish intrastory causal/temporal relationships. For example, the term *at reference* indicates the relative point in time of the main action of the story, while drawing a lesson from the story happens *after reference,* another time designation.

The **StoryType** slot allows the indexer to advise the inferencing mechanism to identify what the story might be useful for and what the level of abstraction of the story content is. For example, if a story contains useful cautionary advice this slot will contain the value *Warnings.* If a story is a good explicit of example of something, *Literal Example* would fill this slot.

Inference Procedures

We have implemented inference procedures for all of the link types specified by the ASK model of hypermedia. In concept, inference procedures compare one of the representation frames of a reference story with all other frames in the story base. Operationally, inference is implemented as path finding, not exhaustive search and test. Links from slot fillers in the reference story frames are traversed in the concept hierarchy to identify sibling fillers which instantiate slots of other stories. Inference procedures are implemented as deductive retrieval rules which exploit the relationships between slot fillers. Each rule can create one of

AgentRole	*Student*
BeliefType	
IntentionLevel	*Actually Did*
IntentionType	*Show Lack of Interest*
OutcomeTypes	*Successful*
SituationType	*Being Bored*
TimeOfOccurrence	*At Reference*
StoryType	*Opportunity*

AgentRole	*Teacher*
BeliefType	
IntentionLevel	*Actually Did*
IntentionType	*Assign Independent Activity*
OutcomeTypes	*Successful Positive*
SituationType	*Being Bored*
TimeOfOccurrence	*At Reference*
StoryType	*Opportunity*

Figure 3: Two scenes for the story Entertaining the Troublemaker

more links depending on whether or not the link type is symmetric, *e.g., analogies/analogies,* or complementary, *e.g., context/specifics.*

There are many senses of each link type. A particular rule finds only one. Summaries of each rule we have implemented are listed. Each is described as a process which indexes a new story, the reference story, with respect to existing stories in the the database which are potential follow-up stories.

Context, Specifics, and *Examples* are the implemented **Refocusing** rules. In a reference story scene if the parent concept of the situation or the agent's activity (*e.g.,* in the concept hierachy for situations, *interpersonal struggles* is the parent of *being bored*) occurs in a potential follow-up story scene, the context link is proposed. If on the other hand it is a child concept that is present in the follow-up story scene, then the specifics link is proposed. When a specifics link has been proposed and the follow-up story scene also has the story type of literal example, then an examples link is also proposed.

Earlier Events, Later Events, Causes and *Results* are the **Causality** rules. When absolute temporal information is available in a reference story scene, and a potential follow-up story scene describes the same situation or similar agent activity and has an earlier absolute time designation, the earlier events link is proposed. The later events link is proposed analogously. When absolute temporal information is not available in a reference story scene, and a potential follow-up story scene has the same agent activity but an earlier position in the intentional chain (*e.g.,* in Figure 2, *graduate* is earlier in the intentional chain than *get good grades*), a causes link is proposed. A results link is proposed if the follow-up is later than the reference story scene in the intentional chain. Also, when a reference story scene is missing a belief to explain an agent's activity or situation, causes links are proposed to all follow-up story scenes that can supply one. A results link is proposed if the reference and follow-up story scenes are about similar situations or have similar agent activity and the follow-up story scene can provide the reference scene with missing outcome information.

Analogies and *Alternatives* are the **Comparison** rules. If a reference and follow-up story scene have agents with similar beliefs, situations, or activities (as determined taxomomically, *e.g.,* in Figure 2, *pass exams* is a peer of *get good grades*), then an analogies link is proposed between them. However, if in otherwise similar story scenes, a dissimilar value is found in exactly one of the slots used above to compute similarity, then an alternatives link is proposed instead.

Warnings and *Opportunities* are the **Advice** rules. In similar reference and follow-up story scenes, if one has a story type of one of the advice link types and the other does not, then a link of that type is proposed from the former to the latter. The indexer provides these story type values when representing the story.

When we first defined the system, the information needs of these inference procedures determined the definition of the frame as well as the parts of domain concept hierarchy vocabulary that are explicitly mentioned in the rules, *e.g.,* a story type of, *literal example,* used in the *examples* link inference. Likewise these rules operate in conjunction with the representations of similarity built into the equivalence classes of the hierarchy. The effectiveness of machine-assisted relative indexing is dependent upon the tuning of this relationship between rules and representation. Experience with tuning the School Stories system indicates that this task is within the capabilities of our indexers.

An Example: Indexing School Stories

Automated inference helps the authors working on School Stories find appropriate links between stories. While our work in this area is ongoing, the examples below illustrate the kinds of links between stories that can be inferred from the simple representation of stories described above.

One story entitled *Entertaining the Troublemaker* begins:

> One problem for smart kids is to keep from boring them in school. Each year that I was in school, my teachers had to find some way to keep me out of trouble since I was both bored and rambunctious. In the second grade I ran messages for the teacher. In the third I built baseball parks out of oak tag. In the fourth I wrote songs. These events turn out to be most of what I remember from those years. School for me was one long attempt to avoid boredom and trouble.[1]

The author has represented it in two scenes in Figure 3. As part of its search for links, the system runs

[1] This story was written by Roger Schank for Group-Write: School Stories.

AgentRole	*Student*
BeliefType	
IntentionLevel	*Actually Did*
IntentionType	*Disrupt Class*
OutcomeTypes	*Successful*
SituationType	*Being Bored*
TimeOfOccurrence	*At Reference*
StoryType	*Literal Example*

Figure 4: A Scene from A Different Bag of Tools

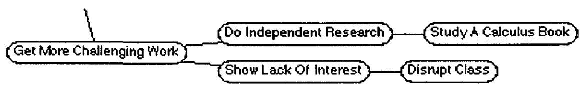

Figure 5: IntentionType Fillers Near *Disrupt Class*

its inference procedure for finding examples links listed above against the frames in Figure 3 and Figure 4. The rule for examples first specializes the fillers for each of the slots of the frame for the reference story (Figure 3). For instance, one causal specialization of the **IntentionType** slot with filler, *Show Lack of Interest* is *Disrupt Class, i.e.,* one way to show lack of interest is to the disrupt class (Figure 5). Because the candidate follow-up story frame in Figure 4 has the **StoryType** slot filler *Literal Example,* the system proposes an examples link to story *A Different Bag of Tools,* the story represented partially by the frame in Figure 4, which reads:

> I had learned to do integrals by various methods shown in a book that my high school physics teacher, Mr. Bader, had given me.
>
> One day he told me to stay after class. "Feynman," he said, "you talk too much and you make too much noise. I know why. You're bored. So I'm going to give you a book. You go up there in the back, in the corner, and study this book, and when you know everything that's in this book, you can talk again"[2]

AgentRole	*Student*
BeliefType	
IntentionLevel	*Actually Did*
IntentionType	*Leave Class*
OutcomeTypes	*Successful*
SituationType	*Getting What You Want*
TimeOfOccurrence	*At Reference*
StoryType	*Literal Example*

Figure 6: A Scene from the story A Deal's a Deal

In this simple case, our representation was sufficient to infer a possible *examples* link.[3] The system continues its search and finds additional ways in which to connect these same two stories. It finds a *similarity* link (one sense of *analogies*) as well through **SituationType**: *Being Bored.* The human indexer can accept one or both of these links for inclusion in School Stories. The system goes on to propose as many other links as the story representations and rules will permit. The author accepts or rejects them as appropriate.

The representation-rule combination excludes some close yet still inappropriate links, as well. The frame for the story *A Deal's a Deal* in Figure 6 does not qualify as an *examples* link for our original story because, while it is has the **StoryType** slot filler *Literal Example,* the **IntentionType** filler *Leave Class* is not a specialization of *Show Lack of Interest* (Figure 5). In the story a *Deal's a Deal* the students were upset because a teacher had broken a promise. It was not that they were bored.

How well the approach excludes near misses depends on the assignment of filler terms to equivalence classes in the concept hierarchies. This assumes that agents do similar things for the same reasons. This kind of similarity limits inadvertent feature matching, because similarities are derived within the context of a specific unambiguous hierarchy locale. In the above example, one construal of *Leave Class* could conceivably be to *Show Lack of Interest,* but that is not the reason in *A Deal's a Deal.* In that story the agents *Leave Class* as a way to *Refuse to Cooperate with a Teacher. Showing Lack of Interest* is a weaker reason and is not represented as similar, *i.e.,* not placed in the same local context of the intentional hierarchy (Figure 5).

These simple examples illustrate how richly connected the stories in our test domain are and how, with a simple representation and processes, these links can be inferred. Given the human commitment to fill out frames for stories and to verify each system-proposed link, such a method significantly reduces the cognitive load human indexers face.

Ongoing Research

This work raises a number of research issues: balancing a fine grained representation against the ability to do simple syntactic feature matching, extending domain concept hierarchies consistently, and testing the effectiveness of the inference rules for machine assisted indexing.

It is difficult to determine just how much detailed domain knowledge should be represented in the content hierarchies to support the kinds of inferencing we have envisioned. There is a trade-off between the coarseness of the representation and its usefulness for infer-

[2] This story was extracted by Ian Underwood for GroupWrite: School Stories from Feynman, R (1985) *Surely you're joking, Mr. Feynman: adventures of a curious character.* New York: W. W. Norton.

[3] In a group story-telling environment authors do not maintain strong causal/temporal threads by telling a sequence of related stories. Therefore the conversational categories have analogical semantics. In the case of an examples link (one sense of **specifics**), one story is an example of the kind of thing discussed in general terms by a story which is probably about another situation written by another author.

ring links by feature matching. At one extreme we could have used fine grained representations that enrich expressiveness but make overall determination of similarity between stories very difficult, because the representations must be processed deeply to compensate for acceptable variation in representation. At the other extreme we could have reified complex relationships into flat propositional features which reduces inferencing to simple feature matching. For example, we rejected the use of complex role relations as a way to represent multiple interacting agents in the **AgentRole** slot, *e.g., student who is infatuated with the teacher but the teacher does not respond favorably.* Use of such unduly extended filler names flattens the representation lessening the ability to infer links, because the internal structure of the filler is not accessible to inference [Domeshek, 1992]. We have tried to find an acceptable balance in our representation between flat and deep representation. Our principle is to provide just the amount of representation needed by the inference rules we have defined.

It is the indexer's job to define the domain concept hierarchies and to use these as fillers in frames for stories. These fillers establish equivalence classes for inferencing. Also where they are placed in the hierarchy represents a prediction about where future indexers will find fillers to describe their stories. Therefore, consistency and economy in the selection of the hierarchy vocabulary is required by both machine and human. We do not yet know how consistent the human extension of domain hierarchies will be. Our experience to date suggests that indexers sometimes overlook or misinterpret the semantics of existing fillers. In many domains, different vocabularies tend to be used in different situations. The result is the creation of synonymous categories. Indexers may also misuse the hierarchy by placing elements of widely divergent levels of abstraction at the same level in the hierarchy. Our current solution is to use the simplest partial concept hierarchy that will support the desired inferences—a corollary of the principle governing representation for rules stated above.

Finally, we have not yet subjected the conversational category-based inference rules for machine assisted linking to a systematic comparison with the link sets derived by human indexers independently. We have however conducted some informal checks on the system's performance in one domain (School Stories). The automated approach found a superset of the links human indexers found in a sample of 16 stories selected at random from the database. We are beginning to apply our technique in a very different domain, *i.e.,* military transportation planning.

These open issues have not prevented us from seeing some significant benefits to indexers already from machine-assisted knowledge acquisition as described herein. Ideally, as our inference procedures are improved and as our confidence grows that the indexes generated converge with those humans would produce, we may be able to grant autonomy to some of them, enabling our ASK hypermedia systems to generate some classes of relative links dynamically. Whether or not that proves possible, we are creating an optimal partnership between human and tool, enabling large-scale relative indexing which neither human nor machine can do alone.

Acknowledgments: The dynamic indexing tool was written by Paul Brown and Paul Rowland.

References

Chafe, W. 1979. The flow of thought and the flow of language. In Givon, T., editor 1979, *Discourse and syntax.* Academic Press, New York. 159–181.

Conklin, E. 1987. Hypertext: An introduction and survey. *IEEE Computer* 2:17–41.

Domeshek, E. 1992. *Do the Right Thing: Component Theory for Indexing Stories as Social Advice.* Ph.D. Dissertation, Yale University, New Haven, CT.

Ferguson, W.; Bareiss, R.; Birnbaum, L.; and Osgood, R. 1992. ASK systems: An approach to the realization of story-based teachers. *The Journal of the Learning Sciences* 2:95–134.

Lemke, A. and Fischer, G. 1990. A cooperative problem solving system for user interface design. In *Proceedings of the Eighth National Conference on Artificial Intelligence*, Menlo Park, CA. AAAI Press/The MIT Press.

Osgood, R. and Bareiss, R. 1992. Index generation in the construction of large-scale conversational hypermedia systems. AAAI-93 Spring Symposium on Case-Based Reasoning and Information Retrieval.

Porter, B. 1989. Similarity assessment: Computation vs. representation. In *Proceedings: Case-Based Reasoning Workshop*, San Mateo, CA. Morgan Kaufman Publishers.

Schank, R. and Abelson, R. 1975. *Scripts, Plans, Goals and Understanding.* Lawrence Erlbaum Associates, Hillsdale, NJ.

Schank, R. and Osgood, R. 1991. A content theory of memory indexing. Technical Report 2, The Institute for the Learning Sciences, Northwestern University, Evanston, IL.

Schank, R. 1977. Rules and topics in conversation. *Cognitive Science* 1:421–441.

Spiro, R. and Jehng, J. 1990. Cognitive flexibility and hypertext: Theory and technology for the nonlinear traversal of complex subject matter. In Nix, D. and Spiro, R., editors 1990, *Cognition, Education, and Multimedia: Exploring Ideas in High Technology.* Lawrence Erlbaum Associates, Hillsdale. 163–205.