Beyond Question Answering (QA) toward Investigative Discourse: Creating Networks of Functionally-Linked Information Objects with Agentized, Contextualized Filters

David A. Evans, Gregory Grefenstette, Jimi Shanahan, Victor Sheftel, Yan Qu, David A. Hull

Clairvoyance Corporation
5001 Baum Boulevard, Suite 700, Pittsburgh, PA 15213-1854, USA
{dae, grefen, jimi, v.sheftel, y.qu, d.hull}@clairvoyancecorp.com

June 2002

Abstract. Current question answering (QA) systems respond to direct questions with relevant "facts" or short passages text. We believe such now-traditional QA systems only primitively address users' actual information needs. Direct QA is merely one form of a larger set of functions that are properly analyzed as *investigative discourse*, which can be regarded as a constantly active process of fetching and filtering information that may bear on a task. We have been developing an *investigative discourse* system—*ViviDocs*—that automatically creates linked networks among information objects (such as words, phrases, sentences, whole texts, etc.) in a work environment. We describe here what we mean by *investigative discourse* systems, and the agentized, contextualized filters that power them. We believe our work defines and clarifies a research program that is relevant to the QA research community.

1. Information Management and Investigative Discourse

Developing technology for information management (IM) is a challenge because our systems cannot be based on the perfection of any single function—such as superior information retrieval, for example—but rather must derive their usefulness from an interaction of many functions. Effective IM will depend on the integration (and exploitation) of models (1) of the user, (2) of the context, and (3) of the application (or information purpose) with (4) the processing of source data. Integration will be the dominant factor in making systems useful.

We have asked ourselves how we might mobilize information in the user's environment to achieve such integration. It is obvious that IM tasks are highly contextualized, highly linked to other tasks and related information—never tasks in isolation. An effective IM system will automatically link varieties of information objects (IOs), dynamically preparing answers to implicit information needs. To this end, rather than focus on a system that performs a single "end-to-end" function, such as processing a request for information, we have been focusing on the critical components of a system (which we call "ViviDocs") that operates *behind* more ordinary tasks, such as reading messages or writing reports. These tasks are not, explicitly, directed at finding information, but when performed in the workplace, these tasks continually generate new information needs, whose responses require grounding in a web of authoritative information.

Every time a user engaged in work reads or writes, the user spontaneously generates new information needs: to understand the text he or she is reading or to supply more substance to the arguments he or she is creating. Simultaneously, each IO (word, entity, term, concept, phrase, proposition, sentence, paragraph, section, document, collection, etc.) encountered or produced creates context for the other IOs in the same discourse. The interaction between implicit information needs and developing context gives rise to the concept of *investigative discourse*. By *investigative discourse*, we mean the hierarchy of information needs that the presentation and ordering of these IOs spawns.

In contrast to the relatively syntactically based typology of questions and answers that we are more accustomed to—*who, what, when, where* ("factoid") questions vs. *how* questions vs. *why* questions vs. *what if* questions—*investigative discourse* types fall into a hierarchy of increasing complexity. On the first level, we have implicit information needs that are local to the information objects mentioned: factoids (such as those supplied by current QA systems), definitions, localizations, elaborations on information objects mentioned. On a higher level, we have argumentative and discovery needs: authoritative evidence for facts, recognition of arguments being made, finding support for and against arguments, discovery of unmentioned

information (e.g. third parties associated with mentioned parties).

2. Investigative Discourse Support via Agentized, Contextualized Filters

For the past six years, Clairvoyance technology and development efforts have focused on the problems of text mining and decision support. Recently, we have begun integrating this experience into a new system, *ViviDocs*, for supporting *Investigative Discourse*.

The author of a document cannot have known what the reader's special circumstances are. The information in the document that bears on those circumstances may not be apparent or easily accessible. Under *ViviDocs*, the system interprets what the reader is reading and provides "missing" or "implicit" or "related" information that the reader might want to see in the document in light of his or her needs.

In *ViviDocs*, while a person reads or writes a text (an e-mail message; a report), the components of the text are continually analyzed into candidate IOs. Agents are generated for each new IO. These agents identify an appropriate (typically local) context for the IO—represented by other text or information in the environment—and then actively fetch and filter relevant information concerning the IO in (latent) information sources the user has access to. We call such agents "Agentized, Contextualized Filters" (ACFs). While ACFs can perform many functions in theory, one class of ACFs quite concretely attempts to provide answers to implicit questions. Such ACFs associate one IO to another in relations that reflect an end state of investigative discourse.

On its simplest level, the effective relation between a particular IO in use (such as the document being written or the message being read) and the environment of available information (including the resources and working directories of an organization or work group) is that of *an implicit question in search of an answer*. As an illustration of what we have in mind, consider the information that is implied by the statements in a text, for example, the following:

"The Battle of Stalingrad represented a major turning point for the Germany Army in World War II. The German generals were out-foxed by the Russian Generals by being drawn into the city. The Russians eventually wore them down, cut off their supply lines, and made retreat impossible."

First, there are many specific and general references in the text—to the "Battle of Stalingrad," the "German Army," the "German generals," the "Russian generals," for example. Some readers (and maybe some writers) might want to know *what* the Battle of Stalingrad was and *where* and *when* it took place. They might well ask *which* German Army was engaged (with *which* Russian Army), and *who* the German and Russian generals were that led their respective forces.

Second, the text suggests a time sequence—and implies a linear causal chain of events—that relate to the outcome: the Germans were drawn into the city; they wore down; their supply lines were cut; they could not retreat, and ultimately were defeated. A reader (writer) might want to know more about the events *before* and *after* the named events; *simultaneous* events in the battle; *contemporaneous* events in other theaters of the war.

Third, the text asserts several propositions: the German Army was defeated (i.e., lost the battle); the outcome was a major turning point of the war; the Germans were drawn into the city; the Russian generals out-foxed the German generals; etc. A reader (writer) might well want to know whether any of these assertions or implications is *true*—more specifically, *what kind of evidence* exists to support the propositions and *what arguments* (for or against) the conclusions might be made.

Finally, the text may be playing a role with respect to a task in a context. It might be the beginning of an essay on major battles of World War II; the response to a question on an examination; one message in a sequence of messages being exchanged among colleagues exploring the reasons the Germans lost World War II; a summary for a supervisor who is charged with making a recommendation on a course of action. In short, in the context of the task, and in the context of the user's work environment, there may be many possible roles that the text might play, the purpose of which is to *establish new information or knowledge*, or *support a decision*.

Again, IOs and contexts can exist at various levels of granularity and abstraction. An IO might be a word, a phrase, a sentence (proposition), a paragraph, a whole document, a set of documents. The proper interpretation of an IO may require using (extracting) content from immediately surrounding text, from an embedding larger section of a document, from a set of (similar, recent) documents, from a message file, from a database in a workgroup. And the IOs to which links are made may be words, phrases, sentences, lists, tables, documents, sets of documents, etc.

4. Investigative Discourse as Basis for IO Models

The relations between IOs suggested by the example are reminiscent of some of the relations that have been employed for modeling discourse. Rhetorical Structure Theory (RST) [1], for example, is based in part on the notion that segments of discourse (text)

bear one of a relatively small number of relations to one another, including *elaboration*, *evidence*, *justification*, *antithesis*, *concession*, etc. Such an approach to modeling discourse—especially to account for cohesion [2] and coherence [3]—is not new. However, the relation types used in discourse theory relate information at the level of the utterance (or the proposition) and are designed to reflect speech acts [4] or possibly the links in a plan [5]. They do not apply to or attempt to account for relations among arbitrary IOs.

If we seek to apply insights from discourse theory to the modeling of relations in IO space, we must imagine that each IO is part of a discourse or plan. Two IOs create a minimum discourse (much as two points define a line). A meaningful relation between any two IOs is possible to the extent that they can be interpreted as playing one of a handful of roles vis-à-vis one another.

We suggest that *investigative discourse* offers a particularly attractive model for such an analysis. In particular, we regard one IO (e.g., the anchor) in its context as representing an *information need* that the second IO satisfies.

5. MEMEX Redux: Networks of Information

The notion of linked information was already present in the original MEMEX vision [6]. Many people regard the World Wide Web as the practical realization of MEMEX since the Web offers a concrete example of linked IOs. Parts of a document may be linked to whole other documents or parts of other documents; the link lattice can be used to move from point to point along pathways of relevance (or, at least, association). But the network itself is relatively static and the types of links are quite general—and must be created "by hand," explicitly. Thus the possible interpretations of information must be decided at link time—by individuals creating links, reflecting their unique perspectives. The possibility that the "same" information might be linked to multiple, distinct other objects, depending on the information needs of the user, cannot be accommodated. Such a static approach is limited. True "knowledge networks" will be subject to constant change and "re-linking" of information, dynamically. Thus, the original vision of MEMEX—as a knowledge network—has not been realized in the Web.

6. Challenges for Research

The program of research and implementation we have undertaken requires us to address a number of challenging problems. Our ultimate success depends on a proper grounding for such problems in theory and a concrete and effective computational interpretation in practice. We outline below a few of these problems, some of which have special relevance to issues in QA systems more generally.

(1) Establishing a formal grounding for ACFs as link types.

We understand what it means to link elements in declarative knowledge networks. One object (say, the concept "dog") is linked to another (say, the concept "poodle") by a relation (say, the relation "is-a") and the attributes and values of the one object (say, "dog" has the attribute "animal-type" with value "mammal") may be available to the other under the operation of the relation. The link (relation) type may provide for a symmetrical or asymmetrical sharing of attributes and values ("mammal" "is-a" "animal" but "animal" NOT "is-a" "mammal"); it may explicitly record a hierarchy or precedence of relations ("john" is "father-of" "bill"); it may restrict the scope of a subordinated object ("cervix" as "part-of" "neck" vs. "cervix" as "part-of" "uterus").

All this is fine for relatively static and declarative semantics. But what about links in the environment of virtually linked information objects? We assume that the link function is to "pass" information from one object to another (or at least to provide the potential to pass such information). But the objects themselves may be in flux; and the relation of one object to another—which might serve as a basis for establishing context, for example—can change as a result of information being passed. There are problems of feedback loops, of state changes, etc.

(2) Managing feedback

This includes the problem of what to do when linked information is activated (= imported into the task and used). Does the link get updated? Does the content of the contextually-circumscribed information change? Do changes propagate?

If we discover that the German general at Stalingrad was Paulus and now have this information available to the text—whether it is explicitly added to the text or merely made part of the background, implicit context—how should we update our filters? In a traditional, adaptive-filtering case, such new information (positively associated with the IO anchoring the filter) would cause the filter to tune toward more such information. But that might not always be an appropriate response, as in the case when the user, having learned that Paulus was the German general, desires to learn about *other* generals, and not more about Paulus.

(3) Defining and managing context

This includes the problem of how we find an implicit question in a body of text as well as the problem of

what influence we allow other environmentally present (relevant) information to have on the process.

More generally, there are at least three kinds of context: (a) the global context, (b) the task context, and (c) the focus. If we have introduced Paulus into the Battle-of-Stalingrad text, the general context tells us about the person, his interests, etc. The task context tells us about his participation in the battle of Stalingrad. The focus may tell us about his particular role in the event—a restricted text context. General context is stable, and does not need to be updated as the text changes. Task context is fairly stable, but may require occasional updates. Focus is the most dynamic; it changes even as the user writes text and needs to be updated frequently.

(4) Typing possible responses

We have a theory of investigative discourse that supports a hierarchy of answer types. These range from the relatively simple to the very complex and include (a) definition ("factoids" such as *who*, *what*, *when*, *where*, etc.), (b) descriptions (contextualized facts), (c) elaborations (information that expands the background of a contextualized fact), (d) explanations (a set or sequence of facts that are causatively related to one another or the anchor IO), (e) arguments (a set of facts that reflects alternative points of view on the anchor IO), (f) synthesis (a set of facts ordered to reflect steps in a logical process, oriented to a goal or outcome), and (g) discovery (a set of facts that represents new knowledge).

How do we factor the information we find (collections of passages; other IOs) into such categories of "answers?" How do we know which answer types are the correct ones (given a large amount of data from many sources)? This problem also subsumes the issue of topic (response) modeling and dealing with duplications, contradictions, etc.

This also subsumes the problem of moving from implicit questions to implicit argumentation, as when the best answer is an array of information "for" and "against" a particular interpretation of a question. Moving toward "dialectical" processing of facts-as-answers will be especially useful for the QA research community.

(5) Adding texture to linked information

Rather than just providing a "fact" or other response as an answer, we might want to give the user a sense of the structure of the space of answers (via topic modeling, perhaps); or the location of centers of importance (via semantic hubs and authorities, perhaps); or of related topical "regions" (via semantic-space abstractions).

(6) Creating effective models of utility and cost

The natural calibration of an ACF is to "maximize utility." We understand what this means in the case of traditional (adaptive) filters; the quantitative interpretation amounts to a tradeoff between true- and false-positive responses. But in the investigative discourse model, other measures may be more appropriate, such as importance, novelty, diversity, confidence/support.

Similarly, we will probably need an explicit measure of "cost"—the computing resources, time, interference with other tasks—that a process requires.

(7) Defining and managing user profiles

How we model the user affects utility. For example, if we model the user as an analyst who already knows a great deal about a topic, then we probably want to maximize the novelty aspect of any information we link to the user's work and discount the information already in the user's background (files, past work, workgroup, etc.). As in the case of context models, user profiles may need to be updated based on system (and user) feedback.

7. Concluding Remarks

Our work on the problem of modeling QA as generalized investigative discourse and using such a model as the basis of a system that can generate an agentized network of relations among IOs is just beginning. We plan on having a first prototype system by the early autumn of this year.

References

1. Mann, W.C., & Thompson, S.A. 1988. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text, 8 (3).* 243-281.

2. Haliday, M. A. K. & Hasan, R. 1976. *Cohesion in English.* London: Longman.

3. Hobbs, Jerry R. 1985. On the coherence and structure of discourse. *Technical Report CSLI-85-37*, Center for the Study of Language and Information, Stanford University.

4. Evans, David A. 1985. *Situations and Speech Acts: Toward a Formal Semantics of Discourse.* New York, NY: Garland Publishing, Inc., vii+211pp.

5. Hobbs, Jerry R., and Evans, David A. 1980. Conversation as planned behavior. *Cognitive Science, 4,* 349-377.

6. Bush, V. 1945. As we may think. *Atlantic Monthly, 176, 1, (Jul),* 101-108.