Abstract: We describe an interactive approach to question answering where the user and the system first negotiate the scope and shape of information being sought and then cooperate in locating and assembling the answer. The system, which we call HITIQA\(^{11}\), has access to a large repository of unprocessed and unformatted data, and is additionally equipped with limited capabilities to search and navigate this set. The user asks questions in English, for example “What are the effects of pollution on commercial fishing on the Black Sea?” Mediating between the data and the user is a Dialogue Management System that attempts to make some sense of what the user is asking and what can be retrieved from the database, and then reconcile these through a dialogue. The purpose of the Dialogue Manager is to make the communication between the human and the machine possible and as efficient as it can be, but it does not necessarily imply a full understanding thus leaving the initiative fully in user’s hands. The HITIQA system has been designed primarily for information analysts who require answers to complex, analytical questions, not just finding simple facts.

1. INTRODUCTION

Question answering is of course a form of dialogue. In the simplest instance of this process, a question is posed by one party to the conversation and it is immediately answered by the other:

Q: Are you ready for lunch?
A: Sure, let’s go!

In another setting, a simple question-and-answer information exchange may look like the following:

Q: What’s the first flight out to DC?
A: 6:10 AM.

What makes such exchanges succeed has been explored by various researchers in linguistics and computational linguistics (e.g., Lehnert, 1987; Ferguson, 1998;\

\(^{11}\) HITIQA stands for High-Quality Interactive Question Answering.

T. Strzalkowski and S. Harabagiu, (eds.), Advances in Open Domain Question Answering, 149-188
© 2008 Springer.
Traum, 1994) and we are not going to discuss all these right now. Suffice it to say that something like the following must be at play:
1. The first party has some information need or another goal to achieve (change someone’s belief state, make them do something, e.g., go for lunch)
2. He/She communicates this through a straightforward utterance, possibly in the form of a question.
3. The other party understands (hopefully) the request and its purpose.
4. He/She performs the requested action, which may involve searching through memory or through some help material (e.g., flight schedule) and communicating the answer.

Of course other things may happen in-between, such as misunderstandings, deceptions, ambiguities, etc., but let’s just sweep them under the rug for now. In order to implement the four step process above in an automated system, we need the machine to understand the question well enough to be able to locate and return that bit of information requested by the user, or alternatively to act as expected. In other words, the system responding “6:10 AM” will be taken to mean actually: “the earliest flight on a regular weekday from where you are to DC is at 6:10 AM local time”. This is how the answer is understood by the human user, whether or not the machine actually “means” it. Therefore, in order provide effective communication, the system must understand what the expected answer is, preferably before any search begins. This, however, requires a substantial amount of knowledge separate from the data repository itself, and that knowledge must be rich enough to make the subsequent search little more than a verification of a fact, in other words, the system needs to know what it is looking for.

The problem with this scenario is that most of the system’s world knowledge is actually in that searchable database, and thus is not directly interpretable. To turn this data into usable knowledge requires a great deal of effort, the task that can never be quite completed. Requiring that QA system’s performance depends primarily on having access to a rich apriori knowledge base would clearly limit the practical utility of this technology. We want a QA system to give us access to data we had no time or resources to process or process fast enough. This means that a very different approach is required. One intriguing possibility is to think of the QA system less as a self-reliant oracle but rather as an extension of its human user, a tool smart enough to exploit what the user knows in order to find what he does not. This is the basis of our approach to dialogue with data. However, before we explore this possibility further, let’s digress a little into the area of human-computer dialogue and then work our way back to QA.

2. DIALOGUE WITH INFORMATION AND SERVICES

2.1 Current Research in Human-Computer Dialogue

The field of natural language human-computer dialogue is closely linked to the speech (or dialogue) acts theory, which postulates that speakers’ utterances carry