Task-based Interaction with an Integrated Multilingual, Multimedia Information System: A Formative Evaluation

Pengyi Zhang, Lynne Plettenberg, Judith L. Klavans, Douglas W. Oard and Dagobert Soergel

College of Information Studies, 4105 Hornbake (South Wing), University of Maryland, College Park, MD 20742

{pengyi, lpletten, jklavans, oard, dsoergel} @umd.edu

ABSTRACT

This paper describes a formative evaluation of an integrated multilingual, multimedia information system, a series of user studies designed to guide system development. The system includes automatic speech recognition for English, Chinese, and Arabic, automatic translation from Chinese and Arabic into English, and query-based and profile-based search options. The study design emphasizes repeated evaluation with the same (increasingly experienced) participants, exploration of alternative task designs, rich qualitative and quantitative data collection, and rapid analysis to provide the timely feedback needed to support iterative and responsive development. Results indicate that users presented with materials in a language that they do not know can generate remarkably useful work products, but that integration of transcription, translation, search and profile management poses challenges that would be less evident were each technology to be evaluated in isolation.

Categories and Subject Descriptors

H.3.0 [**Information Systems**]: Information Storage and Retrieval.

General Terms

Design, Experimentation, Human Factors.

Keywords

User studies, cross-language information retrieval, multimedia.

1. INTRODUCTION

In an increasingly interconnected global environment, users such as intelligence and business analysts need to quickly process a vast array of formal and informal information sources – print, Web, radio and television, sometimes in more than one language. They need to see things as they happen and to find things that happened in the past. They need to identify specific facts, opinions and trends to produce reports that decision makers can act on. Advances in technologies such as Automatic Speech Recognition (ASR) and Machine Translation (MT) can be exploited to build systems that help to make this possible. We report on a formative evaluation of one such system, Rosetta, that integrates

ASR and MT, query-based search, explicit and implicit relevance feedback, and tools to support organization of information that is found and authoring of new documents that report and analyze what has been found.

Imagine a scenario in which a breakdown in civil order or the emergence of armed conflict necessitates rapid evacuation of US citizens from the affected region. Diplomatic personnel and security forces will need immediate access to reporting in local media and through informal sources (e.g., personal blogs) if they are to expeditiously locate US citizens, plan evacuation routes, position transportation resources, and focus limited security resources where they can have the greatest effect. This could require a far greater number of analysts able to understand local languages than would be available in such a fast moving situation. Coupling ASR with MT to allow additional analysts that have expertise in evacuating noncombatants (but no knowledge of the local languages), might be an attractive solution, provide that information systems could be designed that these analysts could use effectively. While ASR and MT are presently far from perfect, those capabilities are evolving rapidly. This leads to two questions:

(1) Given the present state of the art in ASR and MT, how can we design systems to best support task performance?

(2) What characteristics of present ASR and MT technology would, if improved, yield the greatest benefit for this task?

As Figure 1 illustrates, problems of this kind are defined by two constraints: (1) by the users' task, and (b) by the information sources that are available. Given these constraints, a co-design problem emerges in which designers must iteratively refine (1) the capabilities of the system and (2) the process by which that system is used. That yields an enormous design space; in this paper we focus on what we have learned about the influence of ASR and MT capabilities on process design, and the influence of the use process on integrated system design.

After a (necessarily brief) description of the Rosetta system and review of related research, we describe the study design. We then report findings and conclude with future research plans.
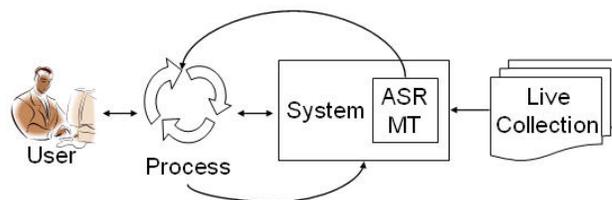

Figure 1. The framework for the study.

2. ROSETTA

Rosetta is a multilingual multimedia information system developed at the IBM T. J. Watson Research Center; it integrates components from IBM and six universities. As Figure 2 indicates, ASR and MT are run in real time to index live non-English multimedia and text content using English terms and to support on-demand display of passages and full text in English. For this study we focused on television news and Web news sites. Present source languages include Arabic, Chinese, and English, although English sources were not used for this study. Searchers control Rosetta in three ways:

- by specifying structured or unstructured queries,
- by explicitly designating passages as useful, redundant, or off-topic,
- by implicitly indicating passage utility through their behavior (e.g., copying or organizing passages).

Template-based structured queries that were integrated later were not yet a focus of the studies reported in this paper. Results can be displayed in text (e.g., side-by-side display of translated and untranslated Web pages), as a storyboard that aligns images extracted from video with English text, or as video with automatically generated English captions.

3. RELATED RESEARCH

The study draws on research in many fields: cross-language information retrieval (CLIR), user- and task-profiling, evaluation of ASR and MT systems, and evaluation of integrated systems.

3.1 User-centered Evaluation of ASR and MT

Reliable and informative techniques for automated evaluation of transcription accuracy have been an essential factor in the remarkable improvements in ASR over the past two decades. ASR error rates on broadcast news have been below 20% for many languages for some time [21], and transcription accuracy for broadcast news is now approaching the limits of human agreement on that task (typically somewhere around 5%). MT has begun to benefit from the same type of cycle, with the rapid adoption of automated accuracy measures over the past decade [11]. MT evaluation is by far the harder task because of the large set of acceptable translations caused by lexical choice and word ordering. User-centered evaluation is therefore still a key component of core MT research and typically takes one of two forms: manual assessment of accuracy and fluency [22], or task-based evaluation in which MT results form the basis for task performance (e.g., reading comprehension tests) [27]. Our work falls into this latter paradigm, with tasks that are significantly more complex than those that have previously been studied.

3.2 User-centered Evaluation of CLIR

There are three broad types of CLIR systems: those based on query translation, those based on document translation, and those that use some aspects of both [15]. Query translation systems translate the user's queries, retrieve documents, and then (if necessary) translate those documents for presentation to the user [10, 18]. Document translation systems instead translate the collection and then index it, simplifying subsequent search and display. Rosetta uses real-time document translation and incremental indexing to accommodate live content.
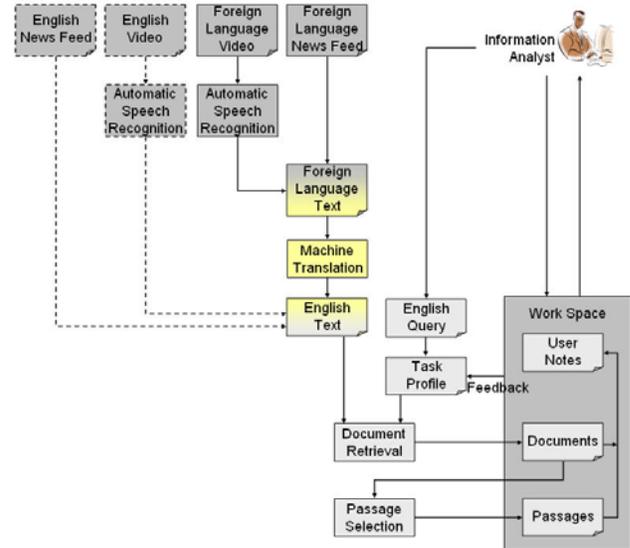

Figure 2. Rosetta system architecture.

Ogden and Davis [19] were among the first to study the utility of CLIR systems in interactive settings. They found that users were able to reliably assess the topical relevance of translated documents. Subsequent studies [17] began to explore process issues, finding that searchers sometimes recognized untranslated terms as useful and added them to their query. As the quality of machine translation improved, the focus of CLIR user studies expanded from merely enabling users to find documents (e.g., for subsequent human translation) to also support information use (e.g., by translating the full text). The question answering task in the interactive track of the Cross-Language Evaluation Forum (iCLEF) is an example of that more comprehensive perspective [8]. The studies reported in this paper continue to broaden the perspective by adding a focus on complex tasks with live multimedia content.

3.3 Evaluation of Integrated Systems

Although evaluation of automated support for profile construction has been studied using model-based approaches for some time (e.g., [4]), surprisingly few user studies have been reported. Moreover, the research literature is somewhat bifurcated, with content-based approaches being found mostly in the information retrieval literature [6] and more strictly behavioral approaches (often called "collaborative filtering" or "recommender systems") being found most often in the machine learning literature [26, 24]. The same evolution is evident for evaluation of relevance feedback in the context of CLIR. Early work focused on model-based evaluation using "canned" relevance judgments (e.g., in the topic tracking task of the Topic Detection and Tracking evaluations) [1]. Orengo and Huyck [20] appear to have been the first to investigate relevance feedback in a CLIR application. Our studies extend that line of research.

Integration across multiple media has also proven to be beneficial in at least one automated evaluation setting: Yang and Hauptmann [28] found that by combining information from several error-prone components (speech recognition, closed captions, OCR of on-screen captions and scene text, and face recognition), the Informedia system was able to identify people appearing in news programs more accurately than with single sources. Our work draws on some of these same insights in an interactive setting.

4. METHODOLOGY

4.1 Participants

Rosetta is intended for trained analysts performing complex tasks, but analysts are not available for a long-term study. We therefore selected another type of trained analyst: six information/library science students, one practicing librarian, and one history Ph.D. student. Information analysts are comfortable with a range of information gathering techniques, and are tasked with synthesizing data from a variety of sources, assessing the credibility of information, and evaluating claims based on supporting evidence [9]; so are librarians. The major difference between analysts and librarians is that professional analysts have more domain knowledge. This domain context can be very important to the task [14], so we provided our participants with descriptions of situations or background readings to partially compensate for this difference. We also designed some task scenarios around the expertise of our one history Ph.D. student (an expert on China).

The study involved 16 three-hour sessions between June 2006 and December 2006, with 3 to 8 participants each. Participants responded to advertisement and were screened through interviews. We required them to be native English speakers and to have formal training or extensive experience in search. They were paid $15 per hour. Additional users participated remotely from time to time, including some system developers, a retired military intelligence analyst, and people new to the project (e.g., observers and faculty members). Involving developers in the evaluation provided context for interpreting results, and also served as an informal expert review, as individual component developers reviewed each other's work.

4.2 Task Scenario Development

In order to simulate the activities of professional analysts as realistically as possible, we experimented with scenario designs to characterize a broad range of capabilities. A scenario consists of three sections: information need, output format, and context.

Information needs ranged from specific factual questions to broad analysis. All of the tasks sought information that was available from non-English materials in the collection. We ran experiments on two news collections of major news agencies in Arabic, Chinese, and English: one frozen trilingual news collection that permitted replication (dated approximately from 2002 to 2005), and a second collection of live television and Web news (with four-minute latency for transcription and translation) of the most recent three months. Retrospective bias (the inability of users to ignore their knowledge of later events when drawing conclusions) is a well known problem with user studies on frozen collections, so we principally chose to use the live collection. Task scenarios for the live collection were designed two days before the sessions and were tested the night before the study to ensure that completion was feasible in the allotted time. Task designs varied, and participants were asked to perform some combination of:

- fact gathering,
- compilation of biographical dossiers,
- hypothesis testing,
- comparison of the way an event was reported in different regions or at different times,
- comparison of responses to an event by particular groups.

The topics addressed included:

- Saddam Hussein trial,
- North Korean missile testing,
- imprisonment of reporter Zhoa Yan,
- death of Abu Musab al-Zarqawi,
- armed conflict between Israel, Palestinian fighters, and Hezbollah.

Some participants were given long-term scenarios, returning to the same topic in several sessions to compile an update on developments since their last report. We asked participants to prepare report and analysis typical of those produced by professional analysts [13], including:

- short text reports (with supporting citations),
- event timelines,
- short PowerPoint presentations,
- marking locations on a map,
- completing templates that we provided,
- organizing evidence in a chart for formal analysis,
- engaging in rapid verbal reporting as information was found.

Scenarios included two types of context: (1) topical background information and (2) situational context that explained why the information was needed, what role the user was playing (e.g. "Civilian employee assisting the extraction of U.S. citizens from Lebanon"), and the intended audience (e.g. "US Secretary of State"). Prior work indicates that this kind of situational context helps limit the natural tendency of participants to interpret the same instructions differently [3].

4.3 Example Scenarios

To illustrate, we present two scenarios:

1) how a searcher completed a fact-finding task,
2) how a searcher used visual clues from video clips.

We then report on some search results to give a sense of how much users were able to achieve.

4.3.1 Fact-finding Task Scenario - Hezbollah

Participants were given the scenario shown in Figure 3 and maps of Israel, the Palestinian territories, and the surrounding region.

Task Scenario: Hezbollah (abridged version)   **Time: 60 min.**

Foreign (U.S., Canadian, Australian, and European) citizens are evacuating Lebanon as a result of the recent armed conflict between Israel, Palestinian fighters, and Hezbollah [Hizbullah].

You are assisting with the extraction of US citizens. Compile sites of recent armed conflict (in the last month) in this area. Your supervisor will use these data to develop evacuation plans.

For each attack you find, place a number on the map and complete as much as you can of the following template:

Location: Date:

Type of attack: Number killed/wounded:

Include attacks in areas not shown on the map. For multiple attacks, list each occurrence.

Figure 3. Hezbollah task scenario

An actual user began with the query `+(Hezbollah palestin*) +(Israel*) +(attack* bomb* shoot*)`. He scanned the

passages the system returned, looking for words that look like names of places. He quickly identified a relevant passage taken from an Arabic news website. Viewing the translated text, he highlighted "Al-Harmel syrian border rocket bombs attacks, late evening," and clicked on a button to add this text to his User Notes. He located Al-Harmel on the map and marked its location.

Meanwhile the system has updated its task profile, based on the inference that the marked text about the bombing is relevant, and retrieved new results. The user updated his result list and saw several new passages. Two of these appeared relevant: the user added them his User Notes and marked places on the map; a third seemed to be an analysis of recent statements from leaders on both sides; the user marked it irrelevant. Updating the results list again removed the irrelevant passage, hid the passages already added to the User Notes, and retrieved new passages that matched the evolving task profile.

Later the user changed his strategy. He broadened his query to `+(Hezbollah) +(attack* bomb*)` and used checkboxes to limit his search to video. The system incorporated the new query with the existing task profile to retrieve a new list of passages. The user continued until he had gathered enough information.

The user compiled the notes he had gathered. He identified 76 distinct attacks, represented them using the task template provided, and attached the compiled notes as an appendix to his report.

4.3.2 The Use of Video - Saddam Trial

Participants were asked to gather opinions of ordinary Iraqi people (what news reporters refer to as "man-on-the-street" opinions) about the trials and sentence of Saddam Hussein. The report was to assist a special task force whose tasked with providing security for parties involved in the trial.

A user began by brainstorming for query words to represent opinions of a "man on the street." She tried `+Saddam +Trial +(Iraq* public) +(reaction opinion)` and a few others and soon realized that it was not possible to express the "man-on-the-street" notion with keywords. She decided to see what she could learn by viewing some video clips. She simply used the query `Saddam trial` and restricted her search to video sources.

As she moved her mouse over a thumbnail in the document list (Figure 4), the thumbnail changed to a slide show of key frames from that video clip. She noticed that a key frame from the video clip showed someone on the street speaking to the camera. She clicked on the Storyboard tab to view the full story.

As she viewed the storyboard (not shown here), she saw three images of people talking to the camera. They appeared to be ordinary Iraqi people. There was a snippet of text associated with each of the three pictures. The first man said "it is the best news in my life, but I cannot assume the sentence is fair or not." A woman said "it is fair because Saddam committed many crimes..." A third person said it is a "political game" and "unfair." The user gathered these pieces of information. She browsed more video clips, and was able to gather several examples of the public reaction to Saddam's trials and sentence.

In this example, the user used visual cues to locate concepts that could not be easily expressed with query terms. The visual context information was essential for the success of the search.

Figure 4. Results list with mouse over.

4.4 Procedure

User sessions typically lasted three hours: one hour each for training, executing the task, and debriefing.

4.4.1 Training

Although our users were trained searchers, they had not searched over the imperfect output of the ASR and MT cascade. Therefore, we provided specific instructions in adapting standard search strategies. At the first session, we conducted a user background survey to gather information about participants' demographics, computer skills, and language abilities. Following each training session, users completed a task background knowledge survey to assess familiarity with the topic.

Since the system was constantly being improved, users needed frequent updates. These updates also helped users who had previously tried a feature, found it lacking, and were disinclined to try again. We gave reintroductions to features that seemed underused, based on survey data. Because we used the same pool of users throughout the study, we had the opportunity to provide additional training on analysis techniques to better profile our target user group. We developed several training modules for specific techniques (e.g., competing hypothesis analysis).

4.4.2 Search

Each user received a hard copy and an electronic copy of the scenario. The system logged the activities of the user, such as queries, page views and resizing windows. Observers (also information science students) recorded users' information-seeking and other aspects of their interaction with the system. Each observer was assigned to one searcher and was positioned to easily read the searcher's screen. Observations were recorded using a separate interface to the same system that allowed editing directly into the search log in real-time. Thus the actions of the user as recorded by the system and by the observer were logged in the same place and linked by time. In some sessions, participants were asked to submit their reports every fifteen minutes via email, enabling us to track product development over time.

4.4.3 Debrief

After the search, participants were debriefed using the following techniques:

Reaction paper: Following each scenario, participants spent ten minutes providing written feedback on the system, the scenario, the study methodology, and anything else they found notable.

Satisfaction Questionnaire: Participants completed a modified version of the Questionnaire for User Interaction Satisfaction (QUIS) [5] to assess users' subjective satisfaction with specific aspects of the system. This survey was modified throughout the study to gather data on new and altered system features.

Post-Search Interview: Each observer interviewed the participant he or she had observed to gather further information on user experience, search strategies, opinions about the system, and satisfaction/frustration. Observers generated tailored questions on the spot based on their observation. Several standard questions were added to an interview protocol used for later sessions. This interview was also an opportunity to verify and further explore observations.

Group Discussion: At the end of each session, we conducted group discussions to pose specific questions from the development team, discuss search tactics, and generate creative ideas for system improvements. Participants found it particularly helpful to hear from other users of the system and to compare strategies and reactions. Group discussion greatly improved the learning curve of the group.

4.5 Data Analysis

Our evaluation design required that we report results to the development team within one week of each study session, and that the development team identify, prioritize, and implement changes just as rapidly. This enabled a continuous cycle of feedback and system improvements. Our feedback cycle proceeded as follows.

Immediately following each session, we read through the observation logs, interview notes, group discussion notes, and reaction papers to gather comments about what did not work properly and specific feature recommendations. We entered both into a bug- and recommendation-tracking database.

Broader issues from the same data sources, together with survey results, were compiled into a session summary report, including:

- a session overview (task scenario, users, and dataset),
- findings about user behavior and usability,
- features examined in the session,
- findings associated with these features,
- system design implications,
- users' comments on training and the user interface.

Task background knowledge survey results and users' reports informed the development of upcoming scenarios. Search log analysis was informal and focused on answering questions of interest to the development team, e.g., "How often are the terms added to a query drawn from the results on the screen?" Periodic analysis of the search logs generated points for group discussion as well as some of the findings below.

More in-depth analysis of the search and observation logs, and of relationships found in the scenario background survey and scenario outputs, is planned for future phases of the research.

5. FINDINGS

5.1 User Accomplishments

The complexity of the integrated system appeared to be a challenge for novice users. After training and a few trials, users developed search routines for completing a task. System output is not very readable although it was produced by state-of-the-art ASR and MT methods. However, users were able to compensate for less-than-fluent ASR-MT output to complete a variety of tasks.

Four users worked on the fact-finding Hezbollah task shown in Figure 3. They had an average self-reported knowledge level of 5.5 out of 10 about the conflicts. Within 60 minutes, they were able to report an average 52 instances of such conflicts (along with 16 inaccurate results). The total number of 202 distinct instances of attacks reported from all users was used as the relevance pool for recall. Table 1 shows the number of attacks each user reported and their precision/recall scores:

Table 1. Hezbollah scenario: number of attacks reported

User	1	2	6	7	Average
Correct / All reported	59/91	49/51	37/53	64/76	52/68
Precision	65%	96%	70%	84%	76%
Comparative recall	29%	24%	18%	32%	26%

In another scenario, users were asked to collect biographical information on Moqtada al-Sadr, the leader of the Mahdi Army in Iraq, and report on Sadr's activities in the past week, which included violent clashes in Sadr city and organizing a rally to support Hezbollah. Five users participated in this session. Within 60 minutes, they were able to collect information about most of the questions with relatively high accuracy:

Table 2. Sadr scenario: answers (correct/all reported)

User	1	2	4	5	7
Country of origin?	1/1	1/1	1/1	1/1	1/1
Related terror groups?	0/0	1/1	3/3	4/4	1/1
Terror activities?	0/0	2/2	2/2	2/2	2/2
Related people?	1/3	1/1	3/4	2/2	4/6
Current location?	1/1	0/0	0/1	1/1	0/0
Raid?	Yes	Yes	Yes	Yes	Yes
Rally?	Yes	Yes	Yes	Yes	Yes

In the following sections, we present major findings on search, result selection and use, and issues arising from ASR-MT, most followed by design implications, and draw together more common themes.

5.2 Search

Our study of the retrieval component focused on users' search tactics and experiences in the integrated environment. By and large, users said that they searched ASR-MT text as if they were searching English text: "It is exactly the same way when I'm

searching English. I should have tried different strategies though." However, there were exceptions.

In performing our assigned tasks, users appeared more concerned with recall than with precision. They expressed a preference for receiving a larger list of potentially relevant documents from which they could select, rather than a smaller, finely honed, system-selected list. This is one reason for users' starting broadly in order not to lose anything (see Section 5.2.1). The emphasis on recall also underlies users' reactions to the relevance feedback mechanism (see Section 5.2.2).

5.2.1 Query Formulation, Choice of Search Terms

Users often started with a general query and narrowed it down to specific aspects. General queries allow users to explore the collection and gain a sense of how a subject or term appears (e.g. how a name is spelled or whether a term appears at all). This knowledge can then be leveraged in more elaborate strategies.

Queries attempt to compensate for translation problems

Specific query terms were often less effective for ASR-MT text. Specific named entities are usually the most effective search terms in specific free-text searches of large English collections (e.g., Google); however, they were significantly less effective here because they were less likely to be translated correctly. For example, city names were less likely to retrieve relevant results than country names. This is not surprising given that the collection was transcribed and translated using corpus-based techniques. Also, because two of the source languages use non-Latin character sets, a single proper name may have several correct transliterations. The transliteration chosen by the system may not be the one chosen by the user. Users learned to work around this issue with some success by broadening their search, for example from Dogubeyzait (a city in Turkey) to Turkey.

This collection poses more problems in selecting query terms. One user said, "I tried to use simpler words and phrasing because of not knowing what comes through in translation." Users had difficulty with one foreign word having several English translations, although sometimes they were able to recognize translation errors and identify for themselves the correct word choice. In the task scenario about Iran's nuclear program, one user recognized a mistranslation and used it in searching. He retrieved several documents containing the expression "nuclear file" instead of "nuclear program"; the user recognized "file" as a mistranslation that might also appear in other documents. He used "nuclear file" in his query and retrieved several additional results.

Implications for design: Some systems make thesaurus-based query term suggestions (spelling variants, synonyms, broader, narrower, and related terms), and/or suggest query terms used by other users. In our context this takes on added importance: since there is no one-on-one mapping from one language to another, it is quite common that one foreign word could have several translations in English. These alternate translations could be suggested as query terms, possibly making an effort to array translations according to different senses of the foreign language term. Going a step further, the system could use relevance feedback to build task profiles with foreign language terms and use these to search foreign language text.

5.2.2 Relevance Feedback and Task-profiling

The system used explicit and implicit relevance feedback to create task profiles, which in turn were used to retrieve better results.

Users are reluctant to provide negative feedback

We had originally thought that by providing a combination of both positive and negative feedback, a user could "train" the system to build an accurate task profile and massage his results list into a very precise set of relevant passages. However, we found that users are more comfortable giving only positive feedback to build a task profile that generates a lot of results, and then paring the list down themselves: "I have an impression of [the negative feedback mechanism] as being roughly analogous to using 'NOT' in a search, which I generally do with great caution."

Before they used the task-profiling system, we asked users if they would be willing to provide negative feedback. Results were mixed: "I would like to mark off-topic because I don't want to see them again for this task." "I'm lazy…Why would I bother marking something old or off-topic?" We initially attributed this reaction to the users' lack of experience with task-profiling. However, as users became more comfortable with the feedback mechanism, they continued to use it to expand their search results rather than reduce them. This behavior points to a deeper lack of trust in task-profiling systems. Users were afraid that based on negative feedback, the system, while retrieving fewer irrelevant results, would also miss significant relevant results.

Users want the task-profiling process exposed

Users expressed a desire to know how task-profiling and feedback work: "How does the [task-profiling component] decide what results to display first?" This desire for transparency has also been found in studies of recommender interfaces, such as [26].

While we expected some questions, we found that the need for explicating the task-profiling process to the user was especially important in a translation environment. Because the text had been automatically transcribed and translated, users had less confidence in the results and were therefore initially skeptical that a feedback mechanism could profile their needs. To quote: "If I found a paragraph that I like talking about the missing soldier and I mark that as 'relevant' meaning I want more about that particular soldier, the feedback returned will give me every missing soldier."

Implications for design: We offered extensive explanation of the task-profiling function. We focused on the shortcomings of keyword search, such as term ambiguity and the difficulty posed by abstract concepts such as "changes" or "relationship," for which task-profiling could compensate. Further explanation could be incorporated into the interface itself: users should have the option to view and possibly change terms added to the query or results excluded by the system. When passages are determined to be redundant, displaying similarity measures between them and other results marked relevant helps users understand why. Studies have shown that allowing users to interact with the feedback attributes via tradeoff categories has potential for building trust and improving user performance [24].

Users require flexibility from task-profiling

Due to the nature of television and online news sources, many documents in the collection contain information about multiple

stories, and all of a story or only a single aspect may be relevant to a user's task. Furthermore, in an ASR-MT environment, short passages of coherent and relevant text may be interspersed with less coherent and thus less useful text. To quote: "When I mark something relevant because I know what the document really meant and I really want more of what the document really meant, the system returns more bad translations."

Therefore our system implements a three-tiered feedback functionality, gathering information to support task profiling at several levels: the document level, the system-specified passage level, and the user-specified passage level.

At the document level, users were able to rate entire documents using a single drop-down menu. Users also provided feedback for system-specified passages, which were extracted from the document set resulting from their keyword query based on their task profile. For even greater flexibility, users could specify their own passages of 5 words or more by highlighting them in the full text of a document and adding them to their notes, which the system interpreted as positive feedback.

Even with passage-level granularity and user-specified passages, our users, all highly proficient keyword searchers, were concerned that the system was not responsive enough to their needs throughout the search process: "You should be allowed to change your mind easily – change terms, keywords, dates."

Implications for design: For greater flexibility, systems should gather feedback at several levels and allow users to control the boundaries of the information to which they are asked to react. Interface design should emphasize connections between various feedback mechanisms and include a combination of hard controls (keyword queries, date, language, and media type filters), and soft controls (feedback mechanisms) for greater responsiveness.

5.3 Result Selection and Use

In the presentation of results, users preferred extracted continuous passages consisting of several sentences over sets of Google-style Key-Word-In-Context (KWIC) snippets. Users had a relatively easy time deciding whether a document was relevant to the task. They managed to get reliable results for factoid questions by using a number of strategies to overcome the problems of machine translation, but experienced great difficulty answering higher-level questions about opinions, reactions, and so on.

5.3.1 Relevance Judgments

Users formed relevance judgments of ASR-MT text in the same way they would for native English text. Keywords were enough for users to make the decision.

Keyword highlighting was very important. In a user's words: "Keywords here and there are enough to decide whether to click on it. Then I clicked on the results to see if they make total sense." Another user reported looking at the context around highlighted keywords: "I scanned the highlighted words, also the context words around them, to see if it is about the topic I'm looking for. For example, this one, it says 'missing soldier' but has a 'Sweden' before it, so I did not click on it." One user considered the number of highlighted keywords.

Extracted passages were useful. Users relied on extracted passages heavily: If the extracted passage seemed to be about the topic, users would then look at the full-text.

One user reported that ranking did not work as well with ASR-MT text as with English: "It [the system] says it is sorted by rank, but I don't think the top ones are any better than the rest of them." That user did not rely on order of presentation when deciding which documents to view.

5.3.2 Facts, Opinions, and Sense-making

Users reported confusion in ASR-MT text about particular events or issues, especially on details, but they were able to get fairly accurate answers to factoid and simple opinion questions in a limited amount of time. Some strategies used include:

- *Using multimedia features:* Additional information provided by non-text media can solve some of the problems posed by ASR-MT output. Users reported using video clips and photos on Web sites to identify and resolve different translations of a single name. In another example, a user said, "I saw a graphic on the screen for the city. I couldn't tell from the words about the city name, but could tell from the screenshot."

- *Pulling multiple pieces of information together*: "Everything that seemed to relate to that question I will note them, and after I read maybe 3-4 snippets, I tried to pull things together and answer the questions."

- *Working with multiple questions at the same time*: "I go in with the idea of one question I am looking for, if I happened to have answer to other question, I will note that and go back to my initial questions."

- *Relying on context informatio*n: "I scanned it, looked at the context of the highlighted word and sentence. Context is important."

- *Searching for more documents to resolve contradictio*n: "If I found a contradiction, I would search for additional sources on those two dates, and see what other sources say about it."

It was difficult for users to answer high-level questions, e.g. to compare the reaction to particular events in different countries. It seemed that users tried to work around ASR and MT problems, but did not yet develop search strategies that would work well to find enough information for this level of usage. Even "having pre-established knowledge about the topic didn't help that much."

When attempting high-level questions, users often drew inferences from the context. However, the inferences were not always correct, especially when there were different stories or news headlines on one translated page. In one example, two adjacent stories were displayed on a result page, one about a Muslim cleric, the other about a U.S. military base in Cuba. A user incorrectly inferred that when the Muslim cleric mentioned "White House," he referred to U.S. actions in Cuba. Another user said, "Another problem is that included with many of the articles were reader comments that might be completely wrong, but unless you looked at the original article you wouldn't know that the bottom half of the page was reader comments, and not the main article."

When sense-making was too hard, users ceased trying to understand a particular document. "I had to find articles I could understand, a lot of times I would just skip something if it was too hard to understand."

Implications for design: Context clues from the original source must be preserved. There need to be clear boundaries for different news stories; otherwise, with ASR-MT text it is very difficult for

users to identify where a new story starts. If there are images or videos in the document, the system should visually present any associations between images and text.

Systems should also provide easy ways for the users to multitask, e.g., allowing them to work on multiple questions at the same time, pull multiple pieces of information together, and switch back and forth between various parts of the system.

Users commented that higher-level sense-making would be easier if the system would provide post-search editing of notes, concept-mapping, and knowledge organization tools with visualization of the topical space surrounding the task and the user profile.

5.4 Issues Arising from ASR-MT

During the post-search interviews and group discussions, we asked users to compare experience with ASR-MT to their experience with English news. Many of the issues reported by the users are specific to ASR-MT text. Search-specific issues are discussed in Section 5.2. This section focuses on issues of use.

Users infer corrections for some translation problems

When reading the ASR-MT text, users were able, by and large, to deal with the word choice problem. Many foreign words have more than one possible English translation, and the system does not always select the right one. For example, in the passage, "Where she something women and Brainstorming women," the word *brainstorming* was actually a bad choice for *attacking* or *accusing*. The user was able to infer from the context that "brainstorming was actually something like berating."

Often the noun and the adjective modifying the noun were reversed in the ASR-MT text because they were in that order in the foreign language. This did not seem to present a problem.

There was inconsistency in translation of proper names. For example, even in the same portion of a video, a cleric's name was translated in two different ways (probably due to a variation in pronunciation that led to a different transcription). Even so, users could still identify them as the same person from the context.

Implications for design: One user wanted the ability to right-click on a translated word and receive a list of alternate translations, reminiscent of the capabilities found in stand-alone CLIR systems [18]. This would allow users to better understand the text, and to suggest a better translation (see below).

Missing, untranslated, and non-English words pose serious problems

The users were constantly unsettled by sentences that would just stop rather than end. A user said, "Missing words makes answering questions the most difficult – but [I] can make some sense or make a guess of it."

Users struggled with strange, non-English words in the translation. For example, "Galatasaray cat or a mistake of meat drainage?" "Galatasaray" is not an English word, and does not seem to be a proper name. Some words appeared to come from nowhere and were not related to the rest of the sentence. Users often ignored such expressions.

Implications for design: The system should provide all available information on non-translated words/phrases, however little; It should not simply drop them or replace them with strange non-English words or "???". Instead the system should provide some information about the word: is it a noun, is it a name, is it a sub-

ject or an object, is it part of a larger phrase? Sometimes users could tell whether it was the subject, the object or the action that was missing, but the system should help as much as possible. The system should also provide any available information about why the word was not translated, for example, if the audio was not clear, or if the word had no English equivalent.

Jumbled sentences obscure higher-level information

Users found sentence-level problems such as misplacement of negation difficult to deal with. A user said, "Sometimes the word 'not' really didn't belong to the first verb of the sentence, but the second verb, which then changed the whole meaning of the phrase." Another user: "Structure of sentences is off – subjects rearranged – reversing order – leaving out words… Noun phrases are in inappropriate places – the direct object is somewhere entirely else." This makes it very difficult to resolve referential words/phrases from the context, since many of the referenced noun phrases were not in the right place.

As a result, users found it was difficult to identify relationships, find higher-level opinions, and make sense of what they read. It was easier for users to answer factoid questions than to identify the relationships between the subjects under discussion. For example, it was possible for a user to infer what was said (the content of speech), but he had a difficult time identifying the speaker: "[The most difficult part] is not knowing who said what."

It is even more difficult to understand an opinion from the ASR-MT text, since subtle meanings such as the tone of the author and the attitude of the media are often lost. As one user put it: "You don't have much variation with the translated words, and subtle meanings cannot be captured in the translated text."

Implications for design: The system should provide aids to decrease the cognitive burden on users. For example, the system could present a list of related named entities in the context, such that users do not have to read or memorize them to make connections.

The system could compensate for the loss of subtle meanings by multimedia features: one user said that he listened to the tone and looked at the facial expression of the broadcaster to try to recover the lost subtle information.

Users need more source material than when searching English

Users have a harder time judging what is reliable in translated text. They read more documents with ASR-MT text than with original English text: "I would not have to read more if the materials were in English… if things are repetitive, I tend to trust them."

This was caused at least in part by the lack of a basis for assessing the accuracy of the ASR-MT produced text and exacerbated by contradictions in the text caused by translation. For example, in one task about a missing soldier, an article stated that the soldier had been released, then later, that he was not. So "I need to gather more information to support my understanding."

Implications for design: Participants suggested that information about the ASR and MT processes be made available to the user in the interface. For example, if reasonable confidence estimates from the ASR-MT cascade were available [2, 25], the system could display words with a high level of confidence in black, and terms with a lower degree of confidence in shades of gray, enabling users to make more informed decisions.

The system should also allow users to easily pull out pieces of information from multiple sources to put together a whole story. Clustering repetitive information together would make it easier for the user to find more documents for verification.

Users are eager to interact with the linguistic processes

Users copied the ASR-MT output into User Notes, or a separate Word file, and edited the text, correcting many of the errors mentioned above. They also repeatedly requested the ability to provide feedback to the system about mistranslations.

Implication for design: Users could provide feedback to the machine translation component to resolve word choice, word order, and name variation problems discovered during result examination and use. Interactive machine translation systems solicit input from users (often professional translators [19]) to achieve better readability. For some problems we mentioned above, users with no particular language skills can make good corrections. It would be useful to store users' corrections for feedback to the ASR and MT systems and to improve the text that is used for search and presentation to the next user. This would be in keeping with the evolving paradigm of end-users as information producers.

5.5 Common Themes

5.5.1 Transparency and Trust

Users requested greater transparency of the ASR and MT processes in the form of more information about non-translated words and visual representations of translation confidence. They also requested detailed information on the inner workings of the task-profiling process. Both of these requests stem from the same motivation: desire for transparency. Further, users did not trust the ASR-MT and task-profiling processes, as evidenced by their reading more results for verification than they would normally, their reluctance to provide negative feedback, and their general preference for recall over precision when searching. These tendencies, caused by a lack of trust, slowed users down.

Users perform better when they feel that they have a good understanding of the system's behavior [12]. This desire on the part of the users has been the motivation for interface designs that expose aspects of processing that would otherwise be hidden. As the functionality and "intelligence" of applications increases, so does the need for a transparent interface [16]. Increasing the transparency of a process by explaining it within the interface has been shown to build feelings of trust (e.g. [12, 24]) and this is the goal of several of our design implications.

5.5.2 Context

Overall, we learned that users experience a heightened need for context in a ASR-MT environment where information is incomplete. In this situation, there is a greater than usual need to help users orient themselves through interface design tactics. Orientation problems are compounded in an environment with low quality/less coherent text. Our users require a higher-than-normal level of support to maintain their orientation in the system.

Therefore the system used a variety of standard visual techniques to help users find their place, including: changing the appearance of visited links, highlighting search terms throughout the document review process, highlighting the words/sentences that were used to create snippets and passages, and highlighting the last viewed result when the user returned to the results list.

Users also reported relying on highlighting of passages or snippets selected by the system and portions of the text they selected themselves when viewing the full text. This helped them to stay oriented by structuring the text around areas known to be important for the task at hand. A user said, "it is very helpful when I click to view the full-text of a document, the system will automatically locate to the relevant passage, with the keywords highlighted, so that I don't have to browse through the news from the beginning."

Furthermore, we found that the data of which the users could be 100% certain became more important in this environment. Therefore we emphasized the metadata—publication date and time, source, document length, and whether the document had been viewed − in the interface.

The need for semantic context was also evident, for example in users' preference for extracted passages over shorter KWIC snippets, and their requests for the ability to view the video clips immediately preceding or following a clip they had retrieved.

5.5.3 Integration

We found that integrating visual and text sources can compensate for the compounding errors resulting from the ASR-MT cascade. We have repeatedly observed users drawing visual information from videos or photos on websites to establish information that was not available in the low-quality text. For example, several spellings of the same name can be recognized because they appear next to pictures of the same person. Some of these tactics for taking advantage of multiple channels of information, which we have observed in our users, might be performed automatically.

This integration also makes possible new types of searching, such as the use of visual cues to locate concepts not easily expressed by query terms as reported in Section 4.3.2. It may also be possible to leverage tacit information such as voice inflection and facial expression, for example, to gauge how "big" or tragic a story is based on reporters' voices and expressions. The opportunities for new types of use warrant further investigation.

6. CONCLUSION AND FUTURE WORK

Our study design allowed us to see different things than quantitative or component studies. For example, rather than calculating how well the system or a component performed, we learned which components were most effective in helping users complete their tasks and which additional components needed to be added or invented, based on coping behaviors we observed. By user-testing a system still under development, we were able to establish that many tasks were feasible even at this early stage, and to focus system improvements on those aspects that most affected user performance.

Our goal was to improve the process by which the user interacted with the system as well as the capabilities of the system itself. We found that this process was characterized by a greater than normal need for system transparency and contextual information. Further we found that there is a potential for components to support other components for greater overall effectiveness of the integrated system. Although present technology is not perfect, it is accurate enough that we can learn much from observing its use in end-to-end, task-based studies.

We have developed several hypotheses about users' information seeking behavior in an integrated multilingual, multimedia infor-

mation system. We would like to test these hypotheses with more controlled experiments on the system's components, in particular, to tease apart what contributions ASR and MT each makes. This could lead to more firmly grounded design implications at every level of integration.

On the theoretical level, we plan to express our results in the framework of a general theory of how people deal with incomplete information.

On the practical level, we would ultimately like to develop system aids for sense-making from ASR-MT outputs. The system would provide context information in result presentation, add more visual clues, and allow users to pull pieces of relevant information together to tell a whole story or get a whole picture. A workspace for post-search notes editing and organization and for outlining and writing the report which is the final result of the task, plays an essential role in connecting search and use of information. It is the central place where the users' sense-making takes place. A user-centered design for a workspace module, a significant enhancement of the User Notes feature of the system, to include, for example, the user's own knowledge organization, concept mapping, outlining and copy and paste with source tracking, would be a major step forward, and we are keenly interested in studying how users work in such an integrated environment.

Information in multiple languages and multiple media is becoming increasingly important not just for analysts but for many users. With RSS feeds, and increasing bandwidth and hardware capacity, systems that allow users to retrieve this information, make sense of it, and use it in their work can soon run on a desktop. This paper informs the design of such systems.

ACKNOWLEDGMENTS

REFERENCES

[1]   Allan, J., Introduction to topic detection and tracking. In James Allan, ed., *Topic Detection and Tracking − Event-based Information Organization*, 1-16, 2002

[2]   Blatz, J., Fitzgerald, E., Foster, G., Gandrabur, S., Goutte, C., Kulesza, A., Sanchis, A., and Ueffing, N., Confidence estimation for machine translation. *COLING '04* (2004), 315-.

[3]   Borlund, P., The concept of relevance in IR. *JASIST 54*, 10 (2003), 913-925.

[4]   Chin, D.N., Empirical evaluation of user models and user-adapted systems. *User Modeling and User-Adapted Interaction* 11, 1-2 (2001), 181-194.

[5]   Chin, J. P., Diehl, V. A., and Norman, K., Development of an instrument measuring user satisfaction of the human-computer interface. *CHI '88*, (1988), 213-218.

[6]   Efthimiadis, E. N., Interactive query expansion: A user-based evaluation in a relevance feedback environment. *JASIS 51*, 11 (Sep 2000), 989-1003.

[7]   Fischer, G., User modeling in human-computer interaction. *User Modeling and User-Adapted Interaction, 11, 1-2,* (2001), 65-86.

[8]   Gonzalo, J., Clough P., and Vallin, A., Overview of the CLEF 2005 interactive track. In *CLEF '05,* 251-262.

[9]   Gotz, D., Zhou, M. X., and Wen, Z., A study of information gathering and result processing in intelligence analysis. *IUI'06,* 2006.

[10] Kim, J., Oard, D. W., and Soergel, D., Searching large collections of recorded speech: A preliminary study. *ASIST '03*, (2003), 330-339.

[11] Knight, K., and Marcu, D. Machine translation in the year 2004. *ICASSP '05* (18-23 March,2005), 965-968.

[12] Koenemann, J., and Belkin, N., A case for interaction: a study of interactive information retrieval behavior and effectiveness, *CHI '96*, (Vancouver, 2005), 205-212.

[13] Krizan, L., Occasional paper number 6: Intelligence essentials for everyone. Joint Military Intelligence Conf., 2000.

[14] Lin, J., Quan, D., Sinha, V., Bakshi, K., Huynh, D., Katz, B., and Karger, D. R., What makes a good answer? The role of context in question answering. *INTERACT '03,* (2003), 25-.

[15] Oard, D. W., and Diekema, A. R., Cross-language information retrieval. *ARIST 33* (1998), 223-56.

[16] Oard, D. W., He D., and Wang, J., User-assisted query translation for cross-language information retrieval. To appear in *Information Processing and Management*.

[17] Oard, D. W., and Powell, A. L. (eds.), *Team TIDES Newsletter*, 2002-2005.

[18] Ogden, W. C., Cowie, J., Davis, M., and Ludovid, S. N., Keizai: An Interactive Cross-Language Text Retrieval System, *Workshop on Machine Translation for Cross Language Info.,* 1999.

[19] Ogden, W. C., and Davis, M. W., Improving cross-language text retrieval with human interactions. In *System Sciences, 2000* (Jan 4-7 2000).

[20] Orengo, V. M., and Huyck, C., Relevance feedback and cross-language information retrieval. *IPM, 42* (2006), 1203-.

[21] Pallett, D. S., A look and NIST's benchmark art tests: past, present, and future. In *Proc. ASRU '03*, (2003), 483-488.

[22] Palmer, D. D., User-centered evaluation for machine translation of spoken language. In *ICASSP'05*, (2005),1012-.

[23] Petrelli, D., Beaulieu, M., Sanderson, M., Demetriou, G., and Herring, P., Observing users, designing clarity: A case study on the user-centered design of a cross-language information retrieval system. *JASIST 55*, 10 (March 2004), 923-934.

[24] Pu, P., and Chen, L., Recommendations I: Trust building with explanation interfaces. *IUI '06,* 2006.

[25] Quirk, C., Training a Sentence-Level Machine Translation Confidence Metric. *LREC '04,* 2004.

[26] Sinha, R., and Swearingen, K., The role of transparency in recommender systems. In *CHI '02*, (2002), 830-831.

[27] Spark Jones, K., and Galliers, J. R., Evaluating Natural Language Processing Systems, *LNAI*, (1996), 1083-.

[28] Yang, J., and Hauptmann, A., Multimodal analysis for person-type classification in news video. In *Storage and Retrieval Methods and Applications for Multimedia*, (San Jose, CA, 2005), 165-172.