Formative Evaluation for Multilingual Multimedia Search and Sense-Making

Douglas W. Oard
Judith Klavans
Dagobert Soergel
Pengyi Zhang
University of Maryland
College Park, MD 20742
{oard,jklavans,dsoergel, pengyi}@umd.edu

Peter Brusilovsky
Daqing He
Tomasz Loboda
University of Pittsburgh
Pittsburgh, PA 15260
{peterb,daqing}@ mail.sis.pitt.edu

Leiming Qian
IBM T.J. Watson Research Center
Yorktown Heights, NY 10598
qianl@us.ibm.com

Abstract

The goal of GALE is to empower the warfighter using language technologies. Component development, and intrinsic evaluation of how well components meet specific criteria, is merely a means to that end. If we are to truly serve the warfighter well, we also need extrinsic measures of how well the components we have built can be used to accomplish our ultimate goal. That is the focus of this paper. Rather than asking only how well we have done ("summative evaluation"), we instead use our process to guide what we do ("formative evaluation"). This paper reports aggregate results from 52 sessions in which we explored how real users actually used real systems that are representative of what can be built today using GALE technologies.

1 Introduction

This paper describes the iterative development of Rosetta, a distillation system for multilingual (Arabic, Chinese, English, Spanish) multimedia (television, Web) streaming content. The process involved close collaboration between the IBM T.J. Watson Research Center (IBM), the University of Maryland (UMD), the University of Pittsburgh (Pitt), and Carnegie Mellon University (CMU). The fundamental challenge was process-system co-design: as new technical capabilities were introduced, new work processes were sometimes needed to best leverage those capabilities; those new work processes in turn help to identify new technical requirements. The result was a virtuous cycle of innovation.

Rosetta integrates six key technologies to support search and sense-making: Automatic Speech Recognition (ASR), Machine Translation (MT), Information Extraction (IE), Information Retrieval (IR), User Modeling (UM), answer pinpointing for Question Answering (QA), and summary generation from structured knowledge representations. The focus of the work reported in this paper was on design innovation and process innovation for integrated architectures (e.g., ASR→MT→IR→ IE→QA) in which GALE technologies are used together to accomplish challenging and realistic tasks.

The remainder of this paper is organized as follows. Section 2 describes the Rosetta system, with an emphasis on the integration of key GALE technologies in four "interaction modes." Section 3 then describes four clusters of user studies that provided insight into how GALE technologies can be used together to accomplish representative tasks. Section 4 concludes the paper by drawing on those experiences to identify key results and implications for future technology development and implementation.

2 The Rosetta System

The Rosetta system consists of three major components: the data collection sub-system, the data processing pipeline, which integrates the ASR, MT and IE components, and the Web application, which integrates IR, UM, QA and summarization components.

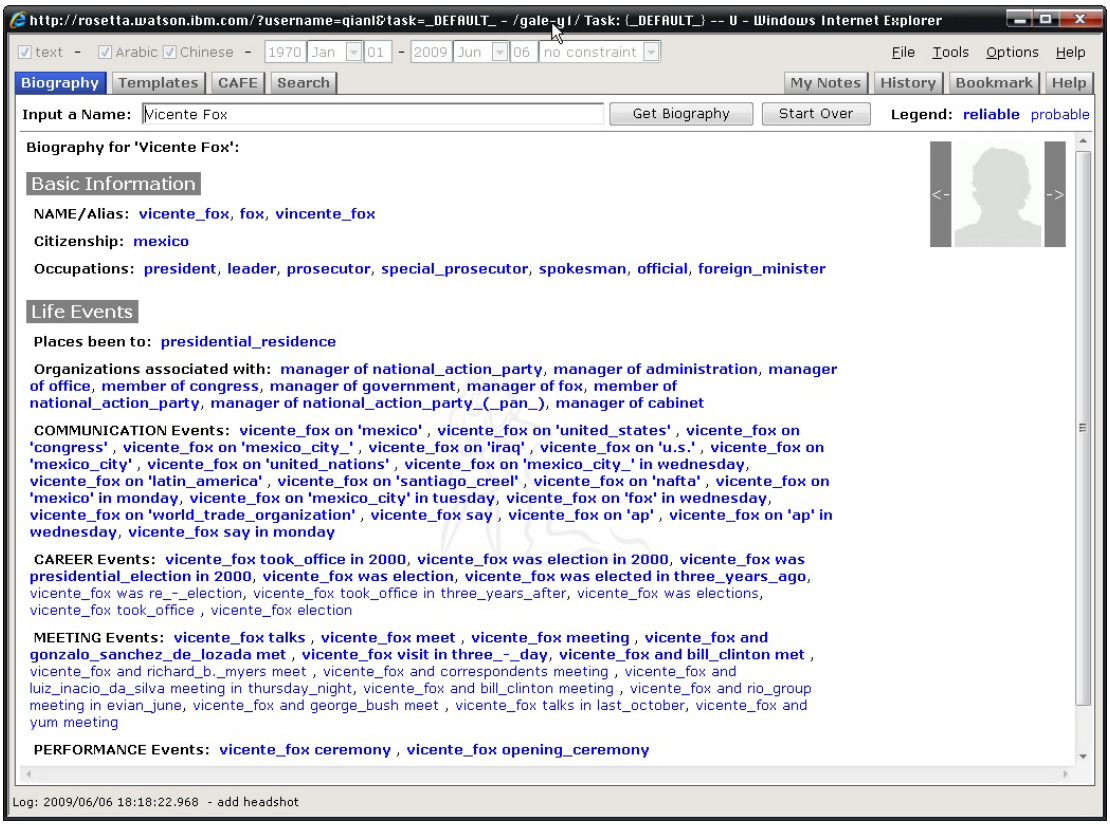

Figure 1. The Biography mode.

Rosetta captures foreign language news broadcasts from Dish Network and performs daily crawl of foreign language Web sites. Rosetta's data processing pipeline is developed on top of IBM's Unstructured Information Management Architecture (UIMA) platform and features a series of data processing components (or "annotators" in UIMA terminology) analyzing input data sequentially. The list of major data annotators in Rosetta includes: ASR, MT, and IE. The Rosetta Web application consists of a browser-based end-user client GUI, and a J2EE back-end server Web application. A browser-based client design was chosen for maximum cross-platform compatibility. Extensive use of Web 2.0 technologies such as AJAX (Asynchronous JavaScript and XML) gives the client a desktop-like feel and enables it to support advanced features not commonly found in conventional Web applications.

Rosetta provides the user with four major modes for information access, each of which can be selected using a tab in the upper left of the screen shots in Figures 1 through 4. These major modes are organized from left to right in an order that we think of as going from most focused to most general.

The **Biography** and **Templates** modes are driven by the full ASR, MT, IE, QA pipeline. As Figure 1 shows, the Biography mode produces an extensive structured display of all information about a person that is known to the system, with drill-down available to a supporting document for each item. In the Template mode, the user can choose from a set of 15 predefined *question templates*, filling in required arguments. The output is a list of snippets that are selected and ordered, based on their likelihood of providing the answer. Figure 2 shows an example template question along with the resulting snippets.

The **CAFÉ** mode exposes the user to GALE user modeling technology (Yang et al., 2007). Rosetta allows each user to create multiple "tasks," each corresponding to a project that the user is working on over time. The search history, document bookmarks, and notes are all task-specific. When users access documents in Rosetta, their actions are tracked and made visible by the system as "footprints" that can act as reminders. For ex-

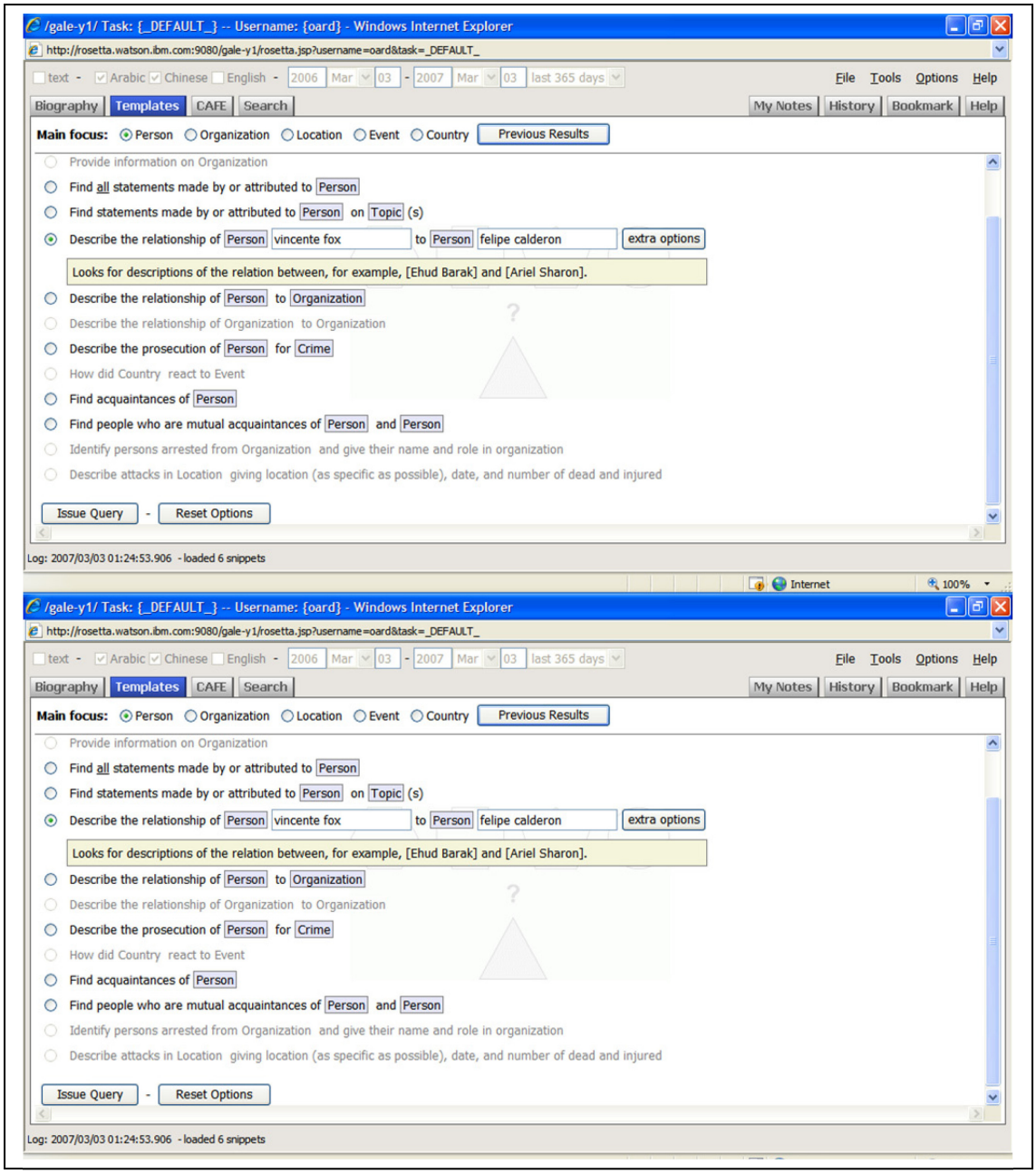

Figure 2. The Templates mode.

ample, a document that has been visited will be highlighted differently when it shows up again, and users can also select a span of text and add it to the "Notes" tab. CMU's Adaptive Filtering Engine (CAFÉ) also uses this evidence to iteratively improve search results. The user first issues a simple query and obtains a list of result snippets from the CAFÉ engine. The user can then provide explicit feedback to the CAFÉ engine by rating individual passages as useful (which results in adding it to the notes), irrelevant (which immediately removes it from the view), or redundant (by marking it as "not new"). The CAFÉ engine also accumulates implicit feedback when the user is working in other modes. For example, if a user finds a document using the Search mode and adds some text to their notes, that will be interpreted as "relevant" feedback by CAFÉ.

This feedback helps CAFÉ refine the system's internal model of the user's task, which can then

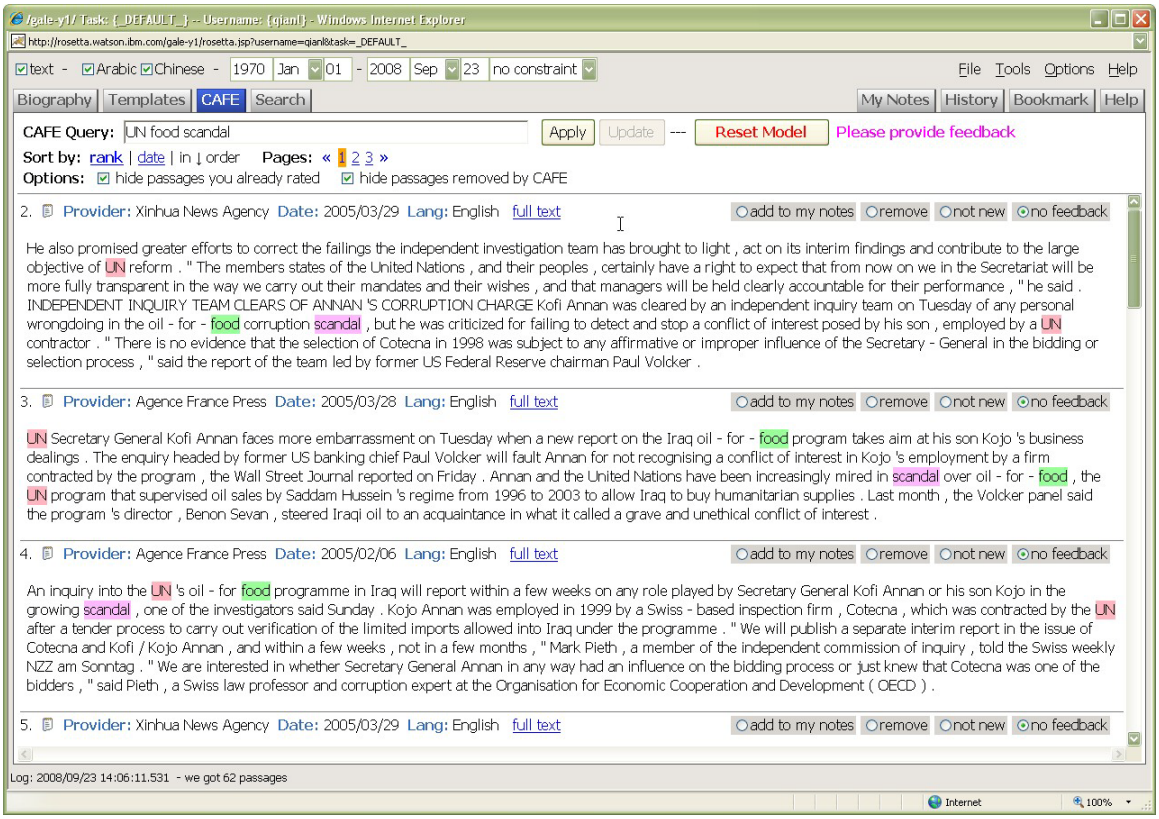

Figure 3. The CAFE mode.

result in identification of additional useful passages. Figure 3 shows the output after just a few feedback iterations. The original query was "UN food scandal;" after adaptive filtering, most of the top passages retrieved are about the food scandal and the role of U.N. secretary Kofi Annan and his son, which in this case are right on target.

The **Search** mode utilizes the GALE ASR, MT, and IR technologies. It allows the user to do conventional keyword search using English queries to find foreign-language multimedia content. Among the four tabs, Search is probably most familiar to most users. For video results, the user can stream the video online with English closed caption, look at the document in a "Storyboard" type of view, and download the video file together with its caption for local playback. For Web results, the user can look at the non-structured textual content, or submit the cached Web page for on-demand translation.

Rosetta has an elaborate event logging mechanism, able to capture nearly every user action (down to mouse movement) and log those events and the associated event context at the server. The events are logged asynchronously in the background so that they do not impact the system's performance. This functionality makes Rosetta well suited for user studies.

One of the most popular features is the ability to create a Microsoft Word report directly from inside Rosetta. The user can select a span of text or an image and click a button to create a new report in Word, he can further add more material to the report and edit the report freely, and each snippet of information contained in the report also links back to the original document. The user never has to leave Rosetta to compile a report.

3 Formative Evaluation

We conducted 52 formative user study sessions at UMD and Pitt between June 2006 and May 2007. In this section we describe those studies in four clusters.

3.1 Initial Eye-Tracking Experiments

In an attempt to gain insight into the way users employed the CAFÉ and Templates modes, we

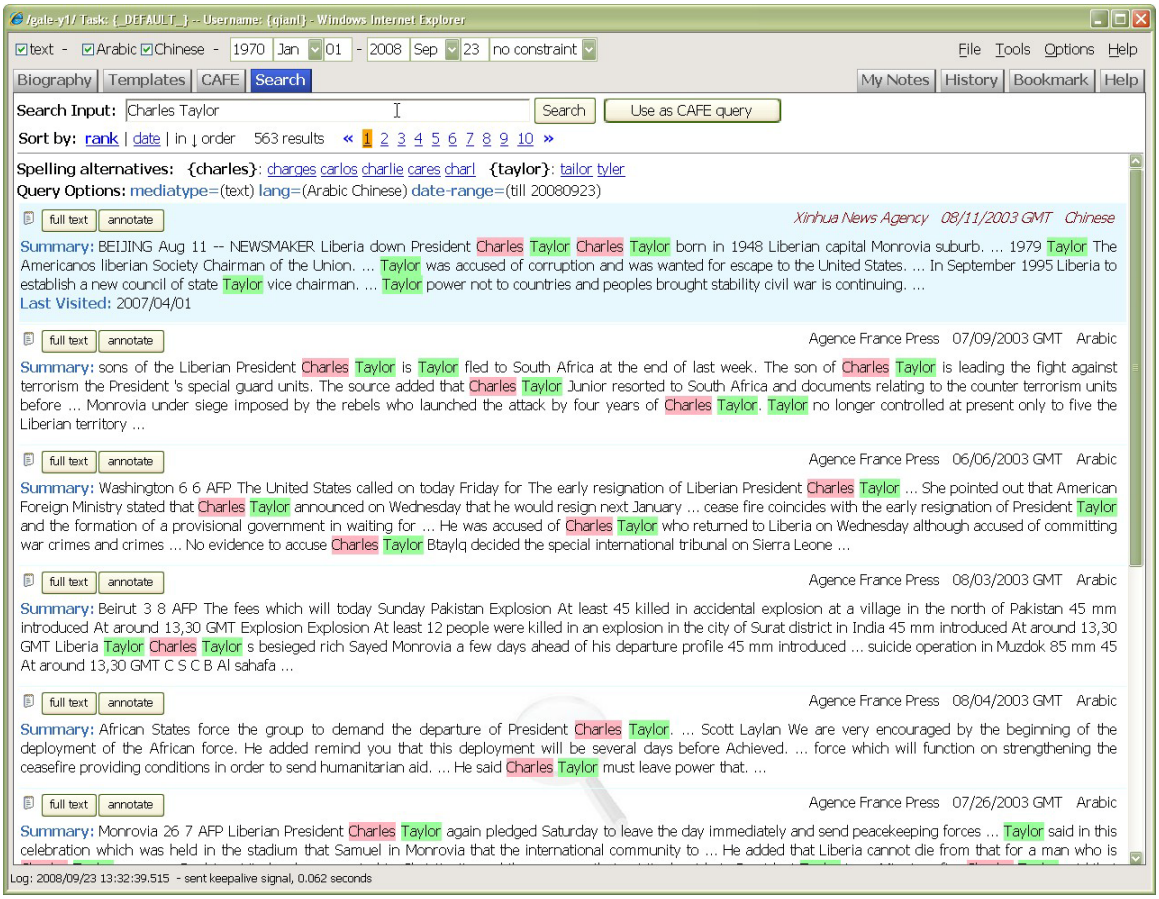

Figure 4. The Search mode.

used eye tracking to collect user gaze data in several user studies. We employed the Tobii 1750 remote eye-tracking device (http://www.tobii.com), which introduced virtually no intrusion in the subject's working environment. The exploratory insights we gained resonated with the more complete and convincing body of evidence described in the following subjections. For example, Figure 5 depicts a fixation-count "heat map" for an early version of Templates mode, illustrating a case in which the user may have paid little attention to the control widgets near the top of the screen (in the dark grey "cold" region) because they were placed too far from the focal region (around the Issue Query button in the lower left, where the circular red "hot" region is).

3.2 Formative Evaluation for Template QA

In order to evaluate the utility of the Templates mode we ran a controlled experiment that compared it with a Search mode based on the Indri retrieval engine. The study was run at the University of Pittsburgh and involved ten graduate students. As surrogates for intelligence analysts, we recruited study participants who were pursuing advanced degrees in Information Sciences, were native English speakers, and had either 1) taken a masters-level course in information retrieval, or 2) had an extensive background in journalism. Each participant was asked to perform two 40-minute search tasks with Rosetta in a counterbalanced design, one task using only the Search mode and one task using only the Templates mode. Prior to starting the tasks, each subject received 15 minutes of training on the Templates mode. All subjects were familiar with generic search systems that are similar to the Search mode. Each task required the participant to interact with a large collection of documents to find answers to several questions that had been provided to them in written form. The outcome of their work was a collection of text passages that they had marked as relevant. An analyst

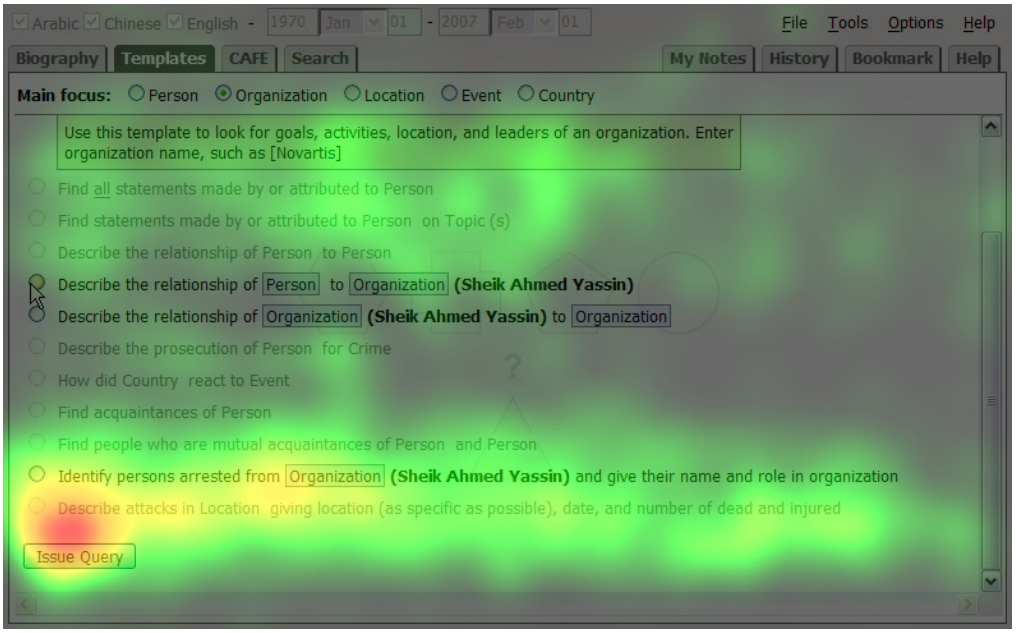

Figure 5. A user gaze fixation heat map for an early version of the Templates mode.

could later use these passages to compile a report, but that was not done as a part of this study.

We found that the rate of accumulation of putatively relevant text passages for subjects using Templates was significantly higher than for subjects using Search. That supports the view that it was easier to find relevant pieces of information when using Templates. Further, we found that subjects issued significantly more queries when using Search. As a result, the inter-query time was significantly longer for the Templates mode, which provides evidence that at average results of Template search brought more interesting information for users to analyze. Query log analysis indicated that this difference reached a statistically reliable level after 20 minutes. Therefore, for similar studies, we recommend not employing short sessions. We hired three annotators to judge the topical relevance of each text passage collected by our participants. Our comparison of the two modes with respect to the relevance of the results also favored the Templates mode.

Our early version of the Templates mode offered a choice between 16 question types. We found that four question types accounted for 77% of all questions (and that six question types accounted for 89% of all questions). This is an instance of the well-known Pareto (80/20) principle. Of course, the specific questions involved surely depended on the specific tasks that we assigned. Nevertheless, this serves to remind us that identifying the most heavily used functions can help to focus optimization effort.

3.3 Formative Evaluation for CAFÉ

A third series of formative evaluation studies at the University of Pittsburgh focused on the CAFÉ mode. The first question to be explored was whether CAFÉ mode could yield better results than Search mode (again, using Indri) in a task-based information exploration context. Realistic task scenarios were used with a static document collection. Eight subjects were recruited to work on eight task scenarios using 3 to 4 search iterations. The studies showed that CAFÉ mode could generate ranked lists of passages with significantly higher precision than Search mode, that it could help users to find a greater quantity of useful information at each stage of an iterative search process, and that it could help users to spend significantly less effort overall to find the same amount of useful information (He et al., 2008).

The second question that we explored was whether the system results should be updated once per session (a between-session update strategy) or whenever there is feedback (an instant update strategy). We recruited 15 participants, each of whom completed the same two task scenarios. Our

results show that the between-session strategy helped to find better quality information, and that it resulted in improved usefulness (i.e., finding useful information) and usability. However, the instant update strategy helped subjects to obtain results more quickly, finding over 98% of all the results that they would eventually find within the first 5 minutes. We believe that the best update strategy may be at some point between the two extremes, balancing adaptability and stability (He et al., 2007), for example, allowing within-session updates by user request only.

3.4 Scenario-Based Integrated Evaluation

Well designed studies with static collections are particularly useful for comparing the utility of alternative components, but for understanding how complete systems will actually be used we wanted to get as close to the real setting as possible. In our case, this called for additional user studies using live content. We therefore chose a study design in which tasks were designed in near-real time and executed simultaneously by all participants (i.e., without counterbalancing). Our principal goal was to explore process-system co-design: the coupled iteration between new system capabilities and new ways of using the resulting system.

A cohort of information studies graduate students at the University of Maryland was trained as surrogates for intelligence analysts. An initial training session was supplemented with brief lectures at the start of each session on aspects of an analyst's tasks and methods that were relevant to that day's scenario. A second cohort of graduate students performed observational data collection and analysis. Observation notes were automatically integrated with system logs in real time; complementary data were collected from post-session interviews, and both structured questionnaires and brief free-style reaction papers from participants. Sessions were generally scheduled about two weeks apart, thus allowing some time to implement system improvements. Summary reports from each session, a requirements tracking database, and weekly teleconferences facilitated information flow between the development and evaluation teams. These studies provided important guidance for user interface refinement, which has been previously reported (Zhang et al., 2007). Here, we focus on what was learned about how users actually employ integrated GALE technologies.

Every session that we ran focused principally on Arabic and/or Chinese content (although during training we did make some use of English content). In some cases, the translations were based on ASR transcripts; in other cases the translations were of Web content. We can, therefore, think holistically of our system as a device for assessing the utility of ASR, MT, or ASR-MT cascades for certain tasks. Consistent with previously reported results, we found that cascading present ASR and MT systems often proved to be adequate for tasks that involved identifying topics and sometimes proved to be adequate for tasks that involved detection of specific factual content, and rarely proved to be adequate for tasks that require detection of nuance. We also found that our participants had more difficulty making use of present Chinese MT systems than present Arabic MT systems, which corresponds well to the reported differences in Translation Error Rate (TER) results in machine transition evaluations.

Remarkably often, our participants proved to be adept at using context and (when present) multiple media (e.g., photos or video) to infer the correct meaning of misrecognized and/or mistranslated terms. For example, several spellings of the same name could be recognized if they appeared next to pictures of the same person. Moreover, our participants learned that using systematically misrecognized and/or mistranslated terms as query terms could sometimes improve results in the Search mode. For example, in a task scenario about Iran's nuclear program, a participant recognized the mistranslated expression "nuclear file" as actually meaning "nuclear program" and used the misrecognized term with good results in subsequent queries. Similarly, a participant recognized "berating women" as the intended meaning of what had been misrecognized as "brainstorming women."

Our participants were repeatedly observed to spontaneously post-edit incorrect translation results, either using the Notes facility or the Microsoft Word editor in Rosetta. Moreover, when asked to suggest system enhancements, they repeatedly requested the ability to provide feedback to the system about mistranslations. Of course, user study participants may be more focused on the system design, so we do not know whether this preference would also be present in operational

users of such systems. Even if the translation system were not adapted, some benefit to future searchers might also accrue from simply memorizing the edits to individual documents. Rosetta now includes this capability.

Earlier user studies had shown that fatigue could become a problem in complex tasks when translations were not easily readable. For this reason, Rosetta includes several features that are designed to help focus the user on specific information (e.g., the snippets in the CAFÉ and Templates mode, and highlighting query terms in Search mode result summaries). These generally proved to be useful, and we believe that investigating additional techniques for helping to guide the user's focus would be useful.

One unexpected and potentially important result was that we repeatedly observed participants engaged in a behavior akin to what journalists might think of as "fact checking from multiple sources." In this case, they were not seeking to verify the correctness of the report but rather the correctness of the translation. Our participants quickly learned that the same pre-translation input typically resulted in the same post-translation output, so they typically dismissed exactly identical output as unhelpful duplication. When they saw the same fact paraphrased in a different way, however, they tended to develop greater confidence in both translations. We are now exploring how we might use this type of cross-source evidence as a basis for unsupervised estimation of translation quality.

We generally left our participants free to use whichever system modes they found most useful, although in some cases we did ask them to use specific modes in order to gain insight into specific usability issues. The Search mode was initially most familiar to our participants, and not surprisingly they initially chose to use it most often. This persisted over time, and an interesting trend emerged: the further to the right the tab was (i.e., the more general the tool), the more often it was used. Such an outcome would not have been predicted from what we had seen earlier in more controlled settings, where Templates and CAFÉ had both shown substantial advantages. Interview and self-report data suggest that the cause was not a reluctance to use new tools; indeed, our participants were all volunteers, and hence could reasonably be described as early adopters of new technologies. Rather, the cause of this effect seems to be twofold. First, more specialized tools are naturally useful in fewer situations. For example, we had a template asking about the relationship between two people (which might be an event), but no template asking about the relationship between two events (which might be a person). Second, our least sophisticated tool (Search) had to expose the most context to the user (for the simple reason that the user had to do more of the job). As our tools became progressively more sophisticated, we adjusted the focus-vs.-context tradeoff more in favor of focus (although with the full context still available on demand). This became most extreme in the biography mode, where the initial display included a large number of very short results. With the context less easily accessible, our participants seemed to have more trouble assessing the quality of the systems results.

Because information seeking is often an iterative process, this difficulty with assessing results may have had the additional effect of making exploratory use of new capabilities somewhat more challenging as well. Indeed, we saw some evidence of that from the way in which participants used CAFÉ. Positive feedback was often provided, and our participants often seemed to be able to develop some degree of confidence from seeing the results of positive feedback. Negative feedback was used far less often, and our users expressed serious reservations about using it in interviews and self-report data. Their reasoning was generally that when they used positive feedback they could see what they were then getting as future results, but when they used negative feedback they had no way of seeing what they then did not get as future results. In this case, the context that they would need extends beyond individual documents – they would also need some way of skimming documents that would have been displayed to them had the negative feedback not been provided.

4 Conclusion and Future Work

Looking back over our full set of studies, we can draw several broad conclusions. Most obviously, and most importantly, there are things that we could learn through user studies that we could not have seen as easily (if at all) in "batch-mode" experiments. Just as we use TER as a predictor of

Human-assessed Translation Error Rate (HTER) during development, we should be using HTER as a predictor of actual utility during development, but then we should check that prediction from time to time using actual user studies. User studies are more expensive than HTER (which is in turn far more expensive than automatic TER scoring), so we need not do user studies every year. But unless we do user studies periodically, it would be hard to know for sure whether we're headed in the right direction.

Another conclusion that we can draw is that all user studies are not created equal. Highly structured user studies are useful early in a development process when the questions focus on capability, but studies situated in setting that are as representative as possible of the envisioned application are also needed at some point if we are to iterate between system design and discovery of the most effective way to use the resulting systems. Indeed, a sequence of increasingly realistic studies may be needed if we are to optimally balance cost and insight. For example, when actual intelligence analysts later used our systems, we learned that some of the behaviors that we had observed with students (e.g., seeking background information) were less common with actual analysis (who of course started with more background).

Ultimately, the most important benefit of our formative evaluation process may have been the bridges that we built between research communities. Just as GALE has brought speech and translation researchers together to study how best to translate spoken content, we have brought component developers together with system developers, and system developers together with process developers (i.e., real users). After all, the ultimate cascade does not end with transcription, translation, retrieval, extraction, or summarization; it ends with use.

References

Daqing He, Dina Demner-Fushman, "HARD Experiment at Maryland: From Need Negotiation to Automated HARD Process," in Proceedings of the Text Retrieval Conference, pp. 707-714, Gaithersburg, 2003.

Daqing He, Peter Brusiloviksy, Jonathan Grady, Qi Li, Jae-wook Ahn, "How Up-to-date Should It Be?: The Value of Instant Profiling and Adaptation in Information Filtering," in Proceedings of the International Conference on Web Intelligence, pp. 699-705, Silicon Val-Valley, 2007.

Daqing He, Peter Brusilovsky, Jae-wook Ahn, Jonathan Grady, Rosta Farzan, Yefei Peng, Yiming Yang, Monica Rogati, "An Evaluation of Adaptive Filtering in the Context of Realistic Task-Based Information Exploration," *Information Processing and Management*, 44(2)511-533, 2008.

Douglas A. Jones and Wade Shen, "Two New Experiments for ILR-Based MT Evaluation**,"** in Proceedings of the Association for Machine Translation in the Americas, Boston, 2006.

Yiming Yang, Abhimanyu Lad, Ni Lao, Abhay Harpale, Bryan Kisiel, Monica Rogati, "Utility-based information distillation over temporally sequenced documents," Proceedings of SIGIR 2007, pp. 31-38, Amsterdam, 2007.

Pengyi Zhang, E. Lynne Plettenberg, Judith L. Klavans, Douglas W. Oard, and Dagobert Soergel, "Task-based Interaction with an Integrated Multilingual, Multimedia Information System: A Formative Evaluation," in Proceedings of the Joint Conference on Digital Libraries, Vancouver, 2007.