SIGIR 2004 Workshop

New Directions For IR Evaluation: Online Conversations

Sheffield, UK, July 29 2004

Organizers

Anton Leuski (co-chair)
USC Institute for Creative Technologies
Douglas W. Oard (co-chair)
University of Maryland, College Park
Abdur Chowdhury
America Online
David Evans
Clairvoyance
Jennifer Preece
University of Maryland, Baltimore County

Online Conversations

Anton Leuski
Institute for Creative Technologies,
University of Southern California
Marina del Rey, CA, 90292
leuski@ict.usc.edu

Douglas W. Oard
College of Information Studies and
Institute for Advanced Computer Studies
College Park, MD 20742
oard@glue.umd.edu

Background and Theme

An online conversation is not very different from the conversations people have been having for thousands of years. Topics are introduced, ideas are shared, and sometimes enlightenment is forthcoming. The difference is that these conversations are sometimes recorded and archived for future use. The knowledge that is created and shared in conversations can therefore be preserved and later accessed by people who did not participate in the original discussion.

Another difference is that many non-verbal cues that we are used to interpreting in face-to-face conversations are missing. Contextual information about where conversation partners are located and what they are doing is also reduced. Consequently, the knowledge that we exchange via online textual conversations is primarily explicit; tacit knowledge tends to be thin if present at all.

Examples of "on-line conversations" include personal electronic mail, mailing lists, instant messaging (IM), Short Message Service (SMS) notes, chat rooms, Usenet newsgroups, threaded Web-based discussion lists, and massive multi-player on-line role playing games (MMORPG). Conversational content poses a number of interesting challenges to systems designed to support access, including exploitation of discourse and dialog structure (e.g., to support thread-based access), the prevalence of informal language and emergent sub-languages, and the importance of establishing adequate context to interpret retrieved materials. Such collections may also allow for non-factual questions. One may explore how a particular topic of discussion cam into existence. The information in these collection allow for exploring the relationships between people, discovering communities, the internal structure within communities.

Online conversations have three distinct properties that set them aside from the traditional document-based dissemination of information and provide us with a fascinating opportunity to study the immediate connections between peoples, their actions, behavior, and language:

Authorship: Every part of a conversation has a unique, and often identifiable, author. Conversations link different people together, and people link different conversations together. If a single conversation can be represented as a graph of message exchanges, then a collection of conversations creates a meta-

network on top of the multiple text fragments that could explored and exploited.

Interactivity: Imagine a Web-based search engine that always returns the most relevant Web page at the top of the ranked list. How should we study what happens next? Online conversations may provide more insight into the evolutionary nature of information seeking. An on-line conversation's initiation and existence is sometimes tightly linked real-life information needs. Responses that are immediate and on-topic, provided by other participants in the conversation, may help us to better understand how the information need evolves with each response and how to determine when that need is satisfied.

Outcome: a conversation may result in a transfer of information or in actions taken by the participants after the conversation. In some cases (e.g., email) the outcome of a conversation might be indirectly traced from followup discussions; in other situations (e.g. on-line games) the result of the discussion will be readily available from system logs.

Goal

Our goal for this workshop is to focus on the domain of on-line conversations, bring together researchers from information retrieval and related research communities (e.g., recommender systems, text data mining, computer-supported cooperative work, and online communities) to see whether there is a sufficient interest in the IR community to study the genre. We plan to organize this workshop around two key questions:

What are the unique information seeking tasks that exist in the domain of online conversations

What opportunities exist to foster important new research through the creation of test collections for genre that have not previously been available?

One possible set of dimensions of the online conversations that can be explored is the following:

Direction: One-way vs. two-ways vs. group discussion. A news paper article is an example of one-way conversation -- a monolog. A two-way dialog such as a record of an IM session assumes a direct response from the other party and a lot of implicit context is assumed. An interview can be placed somewhere between those two extrems: there are two participants but the information primarily flows in one directions. Finally, a chat room discussion provides a medium for several people participating at once.

Timing: Asynchronous vs. asynchronous. Some of the conversation media such as IM and chat rooms presume an immediate response from the participants, while others such as email lack this condition.

Channel: Peer-to-peer (P2P) vs. client-server. IM serves as a communication between two individuals, while a chat room is generally designed to support multiple people interactive at the same time. On the other hand email can serve both functions.

Context: Social vs. organization vs. individual. The context of the conversation has a significant influence both on the form and content of the conversation. For example, an internal company web board will be more formal and focused on work-related topics than a web board appearing on a public web site.

Content: Structured vs. free-form. For exmaple, emails have well-defined fields such as body, subject, return address, etc.

We hope that this workshop would result in the development of at least one specific proposal for creation of a new track at TREC or some similar venue.

Our goal for this workshop is to focus on the domain of on-line conversations, bring together researchers from information retrieval and related research communities (e.g., recommender systems, text data mining, computer-supported cooperative work, and online communities) to see whether there is a sufficient interest in the IR community to study the genre, and propose the ways in which such a study could be facilitated through the creation of standard test collections.

Beyond News Retrieval: Next Steps for IR Evaluation

Douglas W. Oard

College of Information Studies and Institute for Advanced Computer Studies
University of Maryland, College Park, MD, USA

What we typically refer to as "information retrieval" might more properly be called "news retrieval." The reasons for this are simple: (1) news is representative of a broad class of documents that are carefully written, information rich, and therefore valuable, and (2) negotiating rights to use written and broadcast news for research purposes has proven to be feasible. There have, of course, been excursions beyond news in the context of IR evaluation; notably, to the Federal Register in TREC, scientific paper abstracts and patents at NTCIR and TREC, and documentary video in TRECVID. But, to the best of my knowledge, the Congressional Record used in TREC-5 and TREC-6 is presently the only sizable IR test collection that contains anything other than carefully written, information rich documents that can reasonably be treated as self-contained units for retrieval purposes (i.e., the well-known "document independence" assumption).

Biasing our research in favor of these "formal" communications is not necessarily a bad thing, of course. Some degree of focus is essential if we are to make progress, searching news is an important problem in its own right, and much of what we have learned from written news (e.g., Okapi weights) seem to carry over well to other contexts (e.g., broadcast news and scientific paper abstracts). So my concern is not that we have been doing the wrong thing. Rather, I am concerned that if we continue to focus on formal communications, we would miss an equally important emerging genre of informal communications.

It may seem somewhat audacious to claim that searching informal communications could be as important as the more formal documents that have consumed our attention for the past four decades. A few examples might help to illustrate the potential. Electronic mail is now used within organizations in a manner similar to the way written memos were used in the past; this has inspired the use of email as electronic records that shed light on organizational processes, and reliance on email in legal proceedings is becoming increasingly common (prominent examples include Enron, Microsoft, and the tobacco settlement). Similarly, personal email is sure to be used in ways similar to the way in which historians now make use of personal letters that have been preserved to get a glimpse of how people in the past saw their own lives. Moreover, online communities that leverage conversational media (e.g., mailing lists, USENET news, and Web-based threaded discussions lists) can offer new sources of insight into social behavior.

As important as conversational text may seem, the ultimate importance of conversational speech will almost certainly be far greater. Only in the past few years has it become

possible to automatically transcribe conversational speech with sufficient accuracy to support information retrieval. Transcription of conversational telephone speech at a word error rate below 20% has already been demonstrated, and oral history interviews with accented, emotional and elderly speech have been automatically transcribed with error rates below 35%. This capability opens up a vast array of important materials, including recordings of meetings, teleconferences, interviews, survey responses, talk shows, and even personal conversations. Realizing the full potential of searching conversational speech will require further progress on transcription accuracy, speed, and robustness, but the present state of the art is already sufficient for us to begin to work with these types of materials.

Conversational text and speech generally pose four broad types of challenges for information retrieval systems: (1) dialog structure, (2) errors, (3) informality, and (4) information density. Dialog structure is a fundamental challenge whenever more than one participant contributes to the construction of meaning; it essentially invalidates the core assumption that what we wish to retrieve is a "document." Rather, conversational text and speech drive us to aggregate some things (e.g., reply chains in email collections) and to disaggregate others (e.g., passages selected from long interviews). Errors are an obvious consequence of automatic transcription of conversational speech, but they are common in many types of conversational text as well (e.g., letter transpositions in email). Informality introduces additional challenges through the colloquial use of language, and from introduction of non-lexical cues to meaning (e.g., emoticons and font changes in instant messaging). Finally, a few hours in almost any chat room would likely leave you with the clear impression that much conversational content is of relatively little consequence, and therefore not worth finding. This poses challenges not unlike Web search; the density of high-value information can be expected to be far lower in many collections of conversational text and speech that we would expect to find in collections of carefully edited and vetted materials.

These characteristics have implications for the ways in which people will search for conversational text and speech; that, in turn, has implications for the design of IR test collections that model those search processes in ways that can support system development. There are at least three fundamental challenges that we need to address: (1) gaining access to representative collections, (2) defining what we mean by a "document" in this context, and (3) crafting descriptions of representative information needs. Many online communities make discussion histories publicly available, so those would be one natural place to look. Another option would be to use email collections that have been released to the public as a consequence of legal proceedings (e.g., Iran-Contra emails from the U.S. National Security Council). Automatically transcribed conversational speech from some sources (e.g., radio talk shows) might also be a viable option. We traditionally call the unit of retrieval a "document." Different types of collections are likely to demand different definitions of a document, however. In email, we may wish to find conversational threads that include many related reply chains; in interviews, we might seek question-answer pairs. The right choices here iare intimately tied up with our understanding of how the systems we build will be used, but guesses that we make before we see people actually using our systems are likely to be imperfect.

Anticipating representative information needs may prove to be even more challenging. I know, for example, that I would love to be able to search my own conversations for people's names. But I would have more difficulty guessing the kinds of questions a sociologist would want to explore in a collection of radio talk shows that spanned the first decade of the 21st century.

We must also grapple with some broader questions about how our efforts should be organized. Should we propose a special-interest track at TREC, CLEF, or NTCIR? Perhaps we might propose that conversational content be used in some existing venue (e.g., the CLEF spoken document retrieval track). Should we create some separate evaluation venue, as has been done for TDT and SENSEVAL? Or is it too early to focus on a single collection; perhaps more can be learned from many teams working independently with many types of conversational text and speech? If we do choose to pool our efforts in some venue, how should the development of evaluation resources be managed and financed? What other research communities should we reach out to? What kinds of support would be needed to make the barriers to entry sufficiently low that we would attract a critical mass of participants?

We certainly won't be able to answer all of these questions in a single day, but I'm looking forward to our discussions. After all, it is often said that a journey of a thousand li (a measure of distance in ancient China) begins with a single step. When we meet in Sheffield, we will take that first step!

Statement of Interest
SIGIR 2004 Workshop on New Directions for IR Evaluation

Ian Soboroff
National Institute of Standards and Technology
Gaithersburg, MD, USA
ian.soboroff@nist.gov

I have been involved in IR evaluations for several years. As a member of the Retrieval Group at NIST, I have helped design and coordinate TREC evaluations for retrieval, text filtering, web search, and novelty detection tasks. Most of these evaluations require innovative test collection design [Soboroff and Robertson, 2003, Soboroff and Harman, 2003]. Recently, I have been involved in extending IR evaluations to collections in the terabyte range and beyond [Soboroff et al., 2003]. My research interests in evaluation extend to advanced web search, collaborative filtering, and other domains where traditional evaluation methodologies break down.

Thus, I am quite eager to participate in this workshop. The domain of online conversations is an entirely new one for IR evaluation, but nevertheless is quite close to other evaluations that I have been involved in. One can consider informational tasks such as "passage" retrieval of threads or conversation snippets; filtering out noise in the form of off-topic comments, flames, advertising, etc.; and detecting novel topics or information. Additionally, users are likely to be interested in the community where the conversation takes place, and thus want to know who knows whom; who are the domain experts and information gatekeepers; and what are the key information resources that are contextually part of the conversation but which reside outside it, such as web sites, documents, and other conversational areas.

The key question for this workshop, as I see it, is to focus on a task which is at once relevant in the real world, interesting to the research community, and which can be operationalized into a functional evaluation. Many tasks by their nature are impossible to evaluate without a user study. Other tasks need to be simplified and abstracted from the real-world setting so that clear measures can be defined and a reusable test collection can be produced.

Building test collections in this domain is likely to be further complicated because of the privacy and copyright issues involved. However, it would be very interesting to see a collection which contained multiple interrelated resources. For example, an online community might have multiple web sites, an IRC channel, several mailing lists, and a USENET newsgroup.

References

Ian Soboroff and Donna Harman. Overview of the TREC 2003 novelty track. In *Proceedings of the Twelfth Text REtrieval Conference (TREC 2003)*, NIST Special Publication xxx-xxx, Gaithersburg, MD, November 2003. URL http://trec.nist.gov/pubs/trec12/t12_proceedings.html.

Ian Soboroff and Stephen Robertson. Building a filtering test collection for TREC 2002. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2003)*, pages 243–250, Toronto, Canada, July 2003. ACM Press.

Ian Soboroff, Ellen Voorhees, and Nick Craswell. Summary of the SIGIR 2003 workshop on defining evaluation methodologies for terabyte-scale test collections. *SIGIR Forum*, 37(2), Fall 2003. URL http://www.sigir.org/forum/2003F/sigir03_soboroff.pdf.

Cross-Language Collaboration Between Distributed Partners Using Multilingual Chat Messaging.

John Warner
US Army Research Lab
Human Research &
Engineering Directorate
Ft. Huachuca, AZ, USA

Bill Ogden
Computing Research Lab
New Mexico State University
Las Cruces, NM, USA

Melissa Holland
US Army Research Lab
Computational &
Information Sciences
Directorate
Adelphi, MD, USA

1.0 Introduction.

On-line conversations have been growing in use for a number of reasons, but an important one is that they facilitate collaboration among distributed partners, workgroups and teams. The internet, in fact, places the world at your doorstep, allowing collaboration across languages and cultures. This fact points to significant challenges for distributed collaboration. The first challenge is the "machine translation" challenge. Effective collaboration amongst parties that speak different languages requires effective machine translation in many directions. That is, if we have three speakers of language X, Y and Z, for them to collaborate there must effective translation from language X to language Y, from Y to X, from X to Z, from Z to X, from Y to Z and from Z to Y. Each of these paths represent a separate problem with a great deal of variability in language engine translation performance. A machine translation may provide an 80% translation for X to Y but only a 40% translation for Y to X.

The second problem is that even if you can effectively translate words, even if you have a domain limited vocabulary, you may lose valuable affective and cultural cues because you are not co-located. This can be an issue even with same-language on-line conversations among people whose background, experience and milieu are different. However, when languages and culture differ, this factor may take on much greater weight.

In the military, distributed collaboration amongst multinational military and civilian partners has become a reality even though the tools needed to support such collaboration may not yet be in place. This has led to a number of efforts to look at the technologies that are available now in order to see if they can give the military "good enough" tools to help them until more robust and advanced tools can be developed down the road. One thing is clear. There are never going to be enough of the right kind of linguists to meet the operational translation demands of the military. There is a need for looking at promising technologies that can support the non-linguist soldier in his or her ability to communicate effectively with non-English speaking soldiers, coalition partners and civilians. This technology is needed now, even if it does not present a 100% solution. In some cases, even a 40% solution may be welcome.

The purpose of this paper is to present work being done by the Army Research Laboratory and the Computing Research Lab at New Mexico State University to evaluate

the Translingual Instant Messenger (TrIM) software developed by Mitre Corp. as a possible tool for multilingual staff level collaboration. The key questions involve how well the software and its text-based chat/instant messaging paradigm work for collaboration. Since we are evaluating the chat-based tool and not the "quality" of the language engines per se, we are particularly interested in the strategies that people use to deal with less-than-perfect translation. In addition, we are interested in the usability of the software in terms of what functions might be needed in the interface to help overcome mistranslation difficulties.

2.0 A functional evaluation of TrIM for distributed collaboration.

2.1 A brief description of TrIM

Mitre's (www.mitre.org) TrIM software is aimed at the integration of machine translation (MT) and instant messaging. It is based on the Simple Instant Messaging and Presence (SIMP) service, a distributed instant messaging architecture and the Cybertrans machine translation framework, both earlier research projects at Mitre. The Cybertrans framework provides the means for routing messages through a language translation engine before being passed on, translated into the appropriate target language, to the intended receiver of the information. The user interface (UI) provides a dialog window that shows both the original message as typed (in whatever language) and the translated message (in the language that the receiver understands). In addition, logs can be made of the translations for more detailed analysis later.

Figure 1 shows the TrIM UI. There are three main elements. The first is a "Buddy List Window" listing the people available for "messaging." Next is the actual "Chat Window" in which you type and see your messages, their translations as well as what the any other person types and the translations of those messages. In addition, this window lists the people from the buddy list that have entered the chat room and therefore can participate in the conversation. Finally, there is a "Messaging Window" that can be used for a private, instant messaging type conversation between just two of the conversants. This is the mode you use when there are only two people, or it can be used to have a private sidebar between two conversants when there is more than two people using the chat. This interface is based on common Windows instant messaging and chat interface paradigms that people are already familiar with, such as AOL Instant messenger. The program is actually written in Java and can be run under either Windows or UNIX.

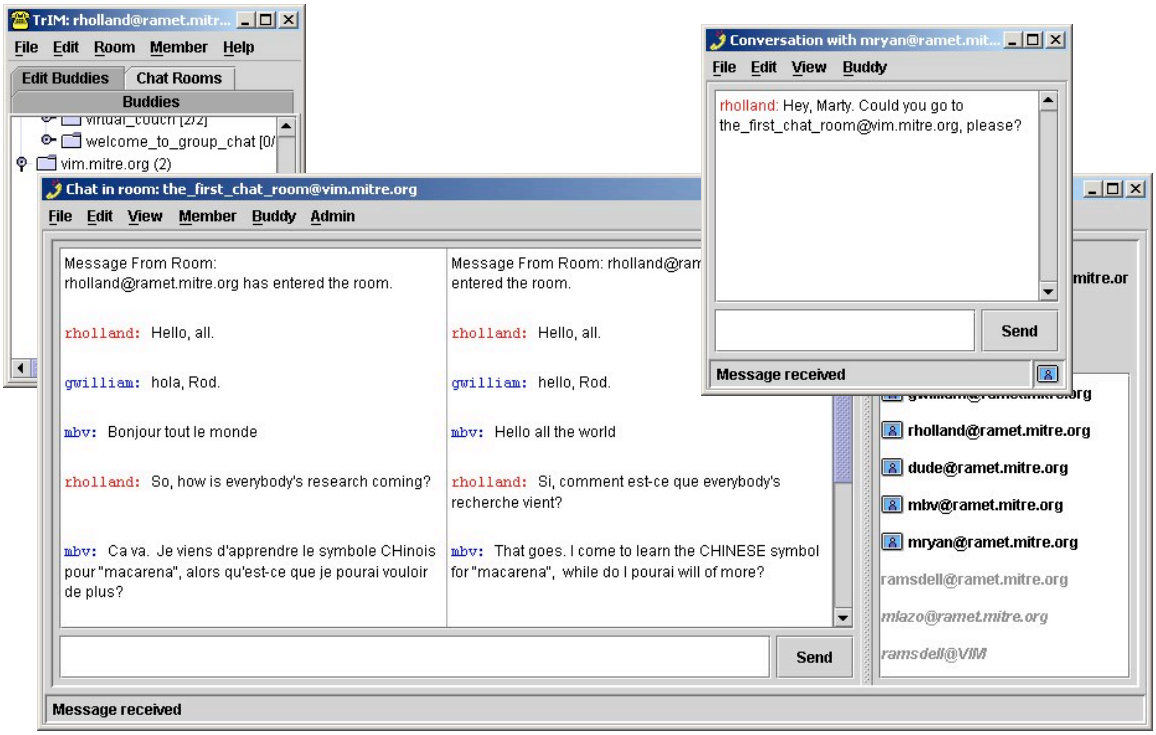

Figure 1: The TrIM user interface.

Thus you can see that even this simple interface and functionality is capable of supporting quite complex and interesting text-based communicative behaviors while providing near-real-time translation in multiple languages. That the translations will vary in quality is a given. We are interested in how people can communicate their intent in a shared knowledge domain, given (or, in spite of) the quality of the translation. Is the tool something that we could put to good operational use now, as is?

2.2 The approach

Our immediate interest was to develop conversation-based tasks that would begin to tell us about the usefulness of the software for collaborative information sharing and shared decision-making. In a previous pilot study, we had looked at a simple fill-in-the-table information sharing task. In this task, we paired native English speakers with either native Chinese, Korean, Japanese, Spanish or English (baseline) speakers. Each member of the pairs were in different rooms (distributed) and could only communicate via TrIM. Each were given a partially filled in matrix of information in which the other partner had the missing information. Thus the object of the task was to share missing information with each other. This very simple task showed that they could accomplish the task with TrIM, but the matrix format and lack of a real context encouraged very simple strategies such as going row by row and indicating what they had. This doesn't really capture what we mean when we say collaboration.

In an attempt to develop more operationally realistic tasks that involve a certain amount of shared context, we came up with two map-and-scenario tasks. Each of the two tasks were supported by a logistics map showing countries, cities and the locations of supplies. Other information included terrain, weather, road conditions and security status. The complete map is shown in Figure 2.

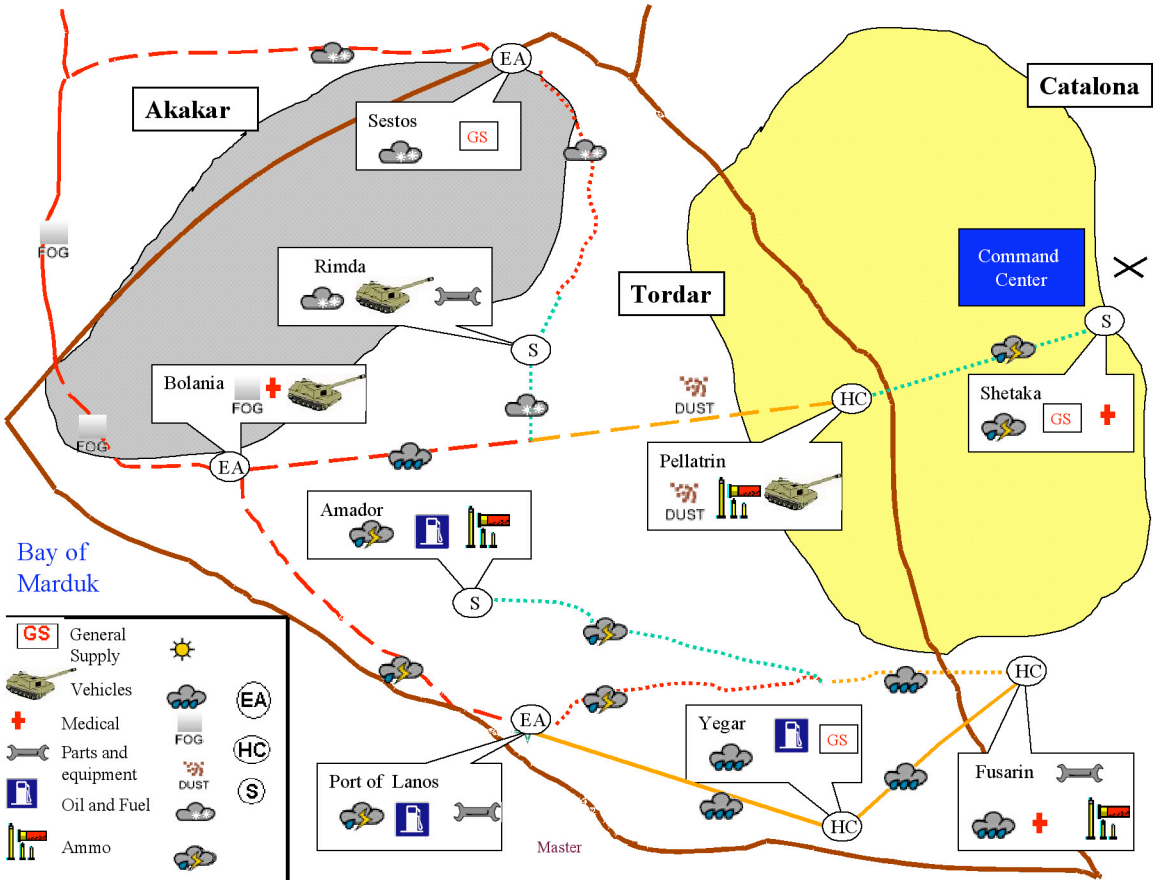

Figure 2: Map of fictional territories used in both the information sharing and decision making tasks.

2.2.1 The information sharing task.

In this task, as in the previous study, we paired two users who were placed in different rooms. Except for the baseline pairs in which we used two native English speakers, all pairs consisted of one native English speaker using English and one non-native English speaker using their native language. Most of the participants were graduate students at New Mexico State University. We had 5 English-Chinese pairs, 4 Korean-English, 4 Japanese-English and 4 English-English pairs.

To give the task a richer, more meaningful context, each participant was given a version of the map in Figure 2 with half the iconic information missing and a scenario that

provided a background to the map and an explanation of what information they needed. The scenario they were presented was:

> *The formerly peaceful but autocratic nation of Tordar has been under attack by various rebel groups operating with the support and protection of the neighboring country of Akakar. The reported goal of the rebels is to claim the oil-rich coast of the Bay of Marduk, which is where Tordar's Akakarian minority is concentrated.*
>
> *The rebels operate in platoon size units and rely on guerilla and terrorist tactics. They have not been successful in toppling even local governments within Tordar, but they have weakened the government by attacking infrastructure and inciting unrest in the Akakarian minorities.*
>
> *Because of it's strategic importance, the US, supported by coalition partners, has sent in an initial fighting force to push back the rebels. Another neighboring ally, Catalona, worried about spill-over of the fighting, has agreed to let US forces use Catalona as a staging area. However, due to urgency, there will be a delay before logistic supply lines can catch up with the deployment. Therefore, initially US forces will have to depend on local sources for most supply needs. Tordar is a heavy purchaser of US Military equipment and supplies. It has reasonably good hospitals and oil reserves, although most of the oil is in disputed territories.*
>
> *The Commander needs to understand the supply situation within Tordar completely, i.e.*
>
> 1. *What supplies are available?*
> 2. *Where all supplies are located at?*
> 3. *Current weather in the towns.*
> 4. *Security conditions for the towns.*
> 5. *Current weather on the roads.*
> 6. *Security conditions for the roads between towns.*

To accomplish their goals, each partner had to find out what information was missing from their map by querying the other using TrIM. They then dragged the appropriate icon to the proper location to complete the map, hopefully ending up, each of them, with the same map, identical to the master map in Figure 2.

2.2.2 The collaborative decision task.

This task was intended to involve a higher level of collaboration than simply information sharing and came when the pairs had completed the first task. They were asked to consider the following problem and to arrive, via TrIM, at a mutually agreed upon solution and justify it:

> *You need to get new reserves of oil and fuel to Rimda. Any pipelines have been shut down and, therefore, oil and fuel needs to be trucked. Consider road*

conditions, the effect of weather on those roads and security issues at a particular supply depot and along the roads in between. Also consider that larger trucks have a harder time with primitive roads than smaller trucks, but smaller trucks will require longer convoys to get the needed amount. What source city and route would be the most effective, balancing risk and time (you would like to get the supplies a soon as possible)?

2.3 Summary of findings.

To get an overall rating of what we call "Translation Efficiency," we sent the logs of the conversations to native speakers who rated the translations as being "Good" (the message was translated with correct syntax and it makes sense), "Acceptable" (the syntax is poor but it nonetheless makes sense to the native speaker) and "Poor" (the message is garbled and has no sensible meaning). Translation efficiency is then determined adding up the number of good and acceptable messages and dividing by the total number of messages. In general, translation efficiency was high with Chinese at 89%, Korean at 82%, Japanese at 80 and Spanish at 84%..

When they did run into translation difficulty, a number of strategies were observed, including:

- **Using the context provided in the problem**. For example, when discussing a route in the decision task, the inquiry "why not Port of Lanos?" was translated in Japanese as "why order to take the left side of the ship of Lanos?" However, the Japanese user was able to figure out what was meant sufficiently to reply, appropriately, "There is enemy in Lanos, isn't there. Let's choose a safe road."
- **Working systematically.** In the information sharing task, many pairs completed their maps by moving city by city, completing all the information for one city and then moving on. Some few did this icon-by-icon (i.e., find all the oil, then all the medical supplies, etc.). While this is similar to the row-by-row strategy in the table versions, it still seems to generate a bit more real conversation with fewer instances of telegraphic lists. Nonetheless, it allows the context to narrow down the possible utterances, limiting possible mistranslations.
- **Rephrasing and meta-statements.** A common way that users found for handling incomprehensible translations was to either attempt to mirror a portion of that phrase as a question, to attempt to rephrase what they thought might have been meant or to make a meta-comment to let the other know that they were not understood. Meta-comments were also used at various points to manage the conversation, such as letting each other know that each felt the information for a particular city was complete and they should move on.
- **Short expressions.** Relying on the context of the map and the scenario, many users found that short expressions were sometimes more efficient and were more successfully translated as intended. This, of course, overlaps with the strategy of working systematically, above.

This study demonstrated that TrIM is an effective tool for distributed multiligual collaboration within a well-defined context. It shows that even though translations are not

always perfect (up to 20% of translations were unacceptable), users are able to find strategies to overcome such problems, at least if they are given sufficient and mutually-shared context.

This study, however, is still a preliminary evaluation. We need to further explore more complex tasks using more than two collaborators. We also need a way to game authorization and rank hierarchies, cultural differences, and whether differences in perception or interpretation can either be limited by a context or overcome despite any shortcomings in translation. Also, many of the non-native speakers here had at least some command of English (being in a graduate program taught in English)—that may or may not be true of all distributed coalition partners in the military.