INFORMATION EXTRACTION FOR THE FUTURE

Paul S. Jacobs
Artificial Intelligence Laboratory
GE Research and Development
Schenectady, NY 12301 USA
psjacobs@crd.ge.com

This broad topic session covered a range of ideas on challenges for the future, along with approaches to meeting these challenges. Although some discussion of matters of scale was common to all the presentations, the speakers offered different perspectives on how to achieve scale-up.

Two of the presentations, "A Unified View of Different Understanding Tasks" by Dekang Lin, and "Applications of Extraction Technology: Today and Tomorrow" by Christine Montgomery, focused on linguistic principles and processing as ways of meeting future challenges. Lin emphasized that a unifying theoretical framework for understanding could lead to "graceful degradation", helping systems to produce better results. Montgomery, while covering a number of practical applications of data extraction, stressed that there are many areas in which systems still perform poorly.

Robert Gaizauskas' talk "Applications of Text Extraction in Police Command and Control Systems: From POETIC to GENIE" presented the "ideal", generic information extraction system as an extension of a current message interpretation framework, which is robust but handles only limited messages in limited domains.

My own view, "Making Knowledge Acquisition Work" questioned the incremental approach to extending data extraction systems, claiming that what systems need to make further leaps is not more of what they have now but more of what they *don't* have. For example, lexical acquisition work has focused on learning features of words that are usually found in computational lexicons, while scaling up seems to emphasize world knowledge and discourse-level knowledge, which exist only in very limited forms today.

John Burger's presentation, "Information Extraction and Text Enrichment" emphasized progress through the integration of data extraction with other technologies and standards rather than the development of new methods. Burger stressed the importance of the SGML standard for document markup, explaining how data extraction and document enrichment work together, creating structured texts, hypertext, and serving a text database. In addition to showing how data extraction could help in a document management system, Burger pointed out that conforming to standards makes it easy to use off-the-shelf tools instead of developing custom research software for many functions.

While there are certain areas of agreement about what systems currently can do and what they can't do, the differences among these brief talks shows some variation in opinion, not only about which problems are most crucial, but also about how to address them. In the discussions at MUC, there were substantial disputes among participants even about whether more progress is likely in the near future, and whether that progress will come from more research, as most of the presenters in this session suggested, or from "fine tuning" and engineering. It seems that the only way to find out is to continue with task-oriented evaluations like MUC, but the rate of progress that has been made in the MUC series does not seem to be slowing. Hence the reading for the future may be "we will do better"—the open questions are *how* we will do better, and *how much* better we will do.