Editor: Steffen Staab
University of Karlsruhe
sst@aifb.uni-karlsruhe.de

Human Language Technologies for Knowledge Management

Knowledge management has changed the way we look at knowledge in the current economy; it is a key factor in an enterprise's success or failure. In contrast to what we as engineers typically love, KM puts people first, organizational issues second, and technology third. Let me explain this triad.

Company leaders want their employees to share knowledge in a way that supports creativity and increases profitability. Thus, the employees must quickly recognize KM's benefits, understanding its ability to offer more capabilities and less work and to better benefit the company. If employees don't recognize KM's value, they won't engage in the KM undertaking, possibly causing it to fail.

Organizational issues comprise several dimensions germane to the organization's objectives. The company's goal is to succeed in the market—otherwise, it wouldn't exist. It must deal with its value chain, taking care of core processes, core products or services, and its customers. KM that does not deal with these issues fails to contribute to the organization's success.

Finally, technology can rescue people from boring, redundant tasks. The question KM asks is, "What are these tasks?" Consider current KM systems—typically knowledge repositories—and how they carry knowledge and the tools they use to access it. These systems easily fall into one of two extremes. Systems at the one extreme tend toward *rigid substance*—databases or knowledge bases in which algebras or logics carry the meaning of data bits and pieces. The corresponding tools are carefully crafted but for a surgeon rather than for a layman. Such systems are valuable but extremely hard to build and maintain. They are easy to destroy, and it isn't always easy to find knowledge in them or reuse knowledge from them. At the other extreme are systems that lean toward *fluid substance*—text content that cannot be gripped unless with a large bucket (the document), making it impossible to find the nuggets of valuable knowledge. Thus, current systems are bound to fail in many practical applications.

Seriously considering these issues to produce a successful KM system leads to at least three requirements. We must

- encourage employees to participate,
- integrate KM with current organizational practice, and
- provide the natural tools such that people can easily recognize the benefits, align with current organizational practices, and use the system.

The natural choice of "substance" for such a KM system is human language, and the required tools are based on human language understanding. Motivating people to use human language for knowledge sharing is easy, because they are used to it. Similarly, aligning human language with many organizational practices is easy, because much of the latter come with documents written in or entangled with human language.

Here is where reality strikes. As we all know, comprehensive human language understanding is out of reach for the foreseeable future. Nevertheless, although the knowledge system's substance is language and complete human language understanding is out of reach, the system need not be restricted to text nor the tools restricted to a keyword-based search. Human language technology can and should do better.

The following essays elaborate on the interaction between human language technology and KM to achieve this goal. Mark Maybury surveys many techniques that have just found—or are currently finding—their way into KM applications. Fabio Ciravegna elaborates on the challenges KM faces regarding information extraction and some of the techniques being explored. Dan Moldovan explains the role that question-answering systems might play in future KM systems. Finally, Georg Niklfeld talks about lifting the restriction of text input and moving toward voice and mobile KM devices.

The essays consider a variety of topics, but share quite a number of common themes. First, knowledge system builders should not have to be human language technology specialists. However, they do need solid human language technology blocks on which to build. Second, human language technology should not presuppose any extremely specific input. Rather, it must adapt to text and speech input types and to its users in a flexible and efficient way. Although many of the building blocks of human language technology are ready to use, others still must find their way into overarching KM systems.

—Steffen Staab

Acknowledgments

This installment of Trends and Controversies was inspired by the corresponding workshop on human language technology and knowledge management at the ACL (Association for Computational Linguistics) 2001 Conference (www.elsnet.org/acl2001hlt+km.html).

Steffen Staab is an assistant professor at the University of Karlsruhe and cofounder of Ontoprise GmbH. His research interests include computational linguistics, text mining, knowledge management, ontologies, and the Semantic Web. He received an MSE from the University of Pennsylvania and a Dr. rer. nat. (PhD in the sciences) from the University of Freiburg, both in informatics. He has cochaired several national and international conferences and workshops, including the 1st German Conference on Knowledge Management (WM 2001). Contact him at the Inst. AIFB, Univ. of Karlsruhe, 76128 Karlsruhe, Germany; sst@aifb.uni-karlsruhe.de.

1094-7167/01/$10.00 © 2001 IEEE

Challenges and Opportunities

Mark Maybury, MITRE

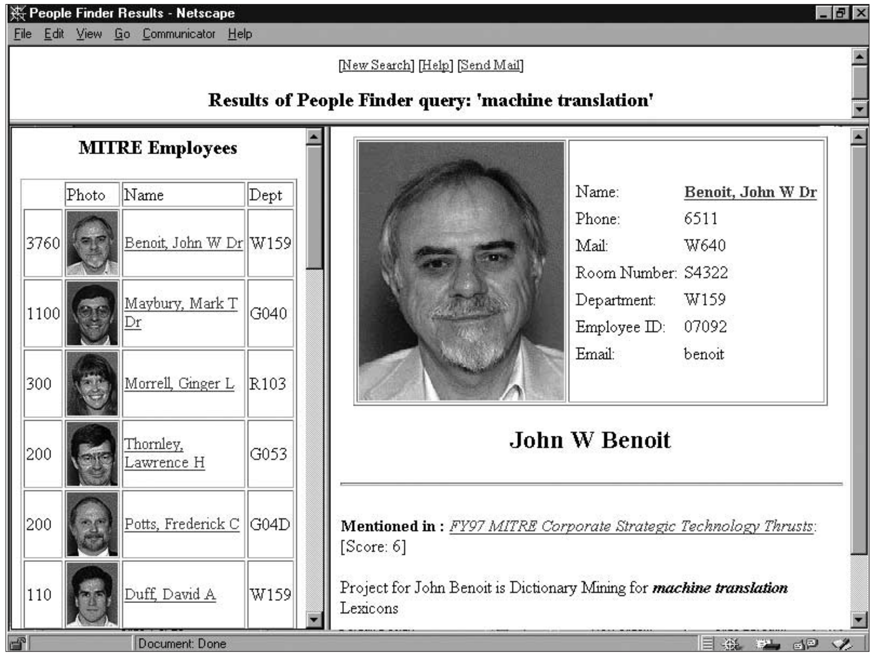


Figure 1. Expert Finder.

In the past few years, knowledge management has received increasing attention from industry, academia, and government. Effective KM is often cited as a key capability for gaining a competitive advantage in global enterprises, and human language technology plays a central role in KM.

KM challenges

KM enhances organizational performance through organizational knowledge sharing, learning, and application of expertise. Indicating KM's importance, many corporations that traditionally measured only the financial aspects of value are beginning to measure human and intellectual value as well.

A range of human language technologies can enable KM, including enhanced information retrieval, extraction, summarization, presentation, and generation. Moreover, human language technologies promise to enhance human access to information and human interaction by increasing our awareness of knowledge artifacts or activities intersecting our interests. Key KM elements include mapping existing knowledge, discovering expertise, and discovering new knowledge.

Knowledge mapping

A primary issue for many organizations is recognizing what they know. Even providing easy access to explicitly captured knowledge in artifacts such as written policies, strategies, documents, and presentations can provide individuals in organizations with tremendous power and efficiency. Often, however, an organization creates so much material that effectively organizing it is a daunting task. We need tools that can automatically generate classifications or taxonomies of explicit corporate and world knowledge. The success of services such as Yahoo, Northern Light, and Quiver illustrate the value (and limitations) of current classification-based collections and collaborative filtering approaches.

Expert and community discovery

Knowing whom to call, who knows a key fact, or who has the know-how or skill to analyze, diagnose, or recommend solutions in a particular domain is a challenge. Manually created corporate-skills databases are costly and inconsistent across individuals and disciplines, and they quickly become obsolete. Finding experts or communities of experts rapidly can be a competitive advantage for a company. Indeed, it is an important function of new business models of virtual corporations.

Knowledge discovery

Having end users or communities of experts (individually or collectively) on hand to answer questions is advantageous. They can accomplish this through their own expertise, machine learning, or data mining, ultimately learning new knowledge, including *ontology induction* (learning new classes of knowledge and meaning representations).

Preliminary results

Researchers have already applied human language technologies to some key KM areas.[1] For example, Irma Becerra-Fernandez created the Searchable Answer Generated Environment (http://sage.fiu.edu) as a repository of experts in Florida's state university system and an expert finder for NASA.[2] In our research at MITRE, we created systems that automatically extract and correlate information from human-created artifacts to assess human expertise.[3] Figure 1 illustrates the screen that MITRE's Expert Finder (also called People Finder) generates after a user types in the keywords "machine translation." Expert Finder uses named-entity extraction to process employee-published resumes and documents as well as corporate newsletters that mention individuals' names in this topic's context in order to automatically create expertise profiles for each employee. Expert Finder then presents a rank-ordered list of employees whose expertise profile best matches this query.

An empirical evaluation comparing 10 technical human resource managers' performance with Expert Finder on five specialty areas (data mining, chemicals, human–computer interaction, network security, and collaboration) demonstrated that Expert Finder performed at approximately 60 percent precision and 40 percent recall when appropriate data was available. This is sufficient performance for finding an expert within one phone call—the original KM objective.

Earlier I mentioned the value of taxonomic search engines, but another area of preliminary success is in knowledge discovery. Several research groups are working to create more effective means to access multimedia information sources.[4] Figure 2 illustrates MITRE's Broadcast News Nagivator, the culmination of many years of research and an integration of multiple human language and other technologies.[5] BNN applies speech, language, and image processing methods to segment, extract, and summarize broadcast news sources to enable personalized and targeted news searches. Figure 2 shows the user querying for stories from all sources for 19 April through 3 May 2001 containing the keyword "Aegis" and the location "Taiwan." Using entities extracted from the retrieved documents, BNN dynamically gen-

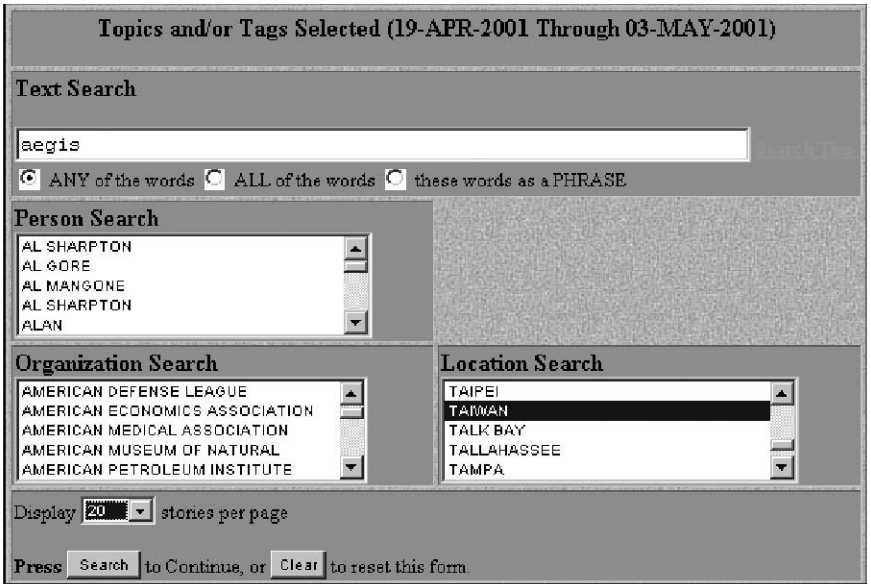


Figure 2. Broadcast News Navigator.

Figure 3. Results from a Broadcast News Navigator search.

erates menus listing people, organizations, and locations.

Figure 3 shows the results from the search in Figure 2. The results include 52 stories mentioning Taiwan and Aegis war ships on multiple programs (for example, C-SPAN, Fox News, CNN Headline News, CNN Morning Headline, CNN World Today, CNN World View, and CNN Moneyline). As Figure 3 shows, BNN presents a quick skim including a keyframe and the top three named entities for each retrieved story. Clicking on any of the keyframes brings the user to that story. Clicking on any of the named entities (people, places, or organizations) brings the user to all stories mentioning that name. Using document-clustering techniques, BNN further provides users with quick access to related stories. An extension of this system automatically mines correlations among named entities that appear across stories to detect and track novel topics.

Human language technology for KM

Having considered KM needs and some preliminary promise of human language technology to provide solutions to those needs, let's look at how human language technology can contribute to KM. Here I outline a range of functional areas of human language technology that offer potential solutions to some required KM elements.

Input analysis

Analyzing user-spoken language and natural input is key to knowledge access. This is essential for applications such as natural language interfaces to databases, question and answering, and multimedia interfaces. In today's conventional interfaces, users are allowed to sequentially input mouse, keyboard, and speech input and perform limited natural language processing—for example, stemming, morphological analysis, and query expansion.

Challenges include dealing with imprecise, ambiguous, or partial input. Addressing these issues in multimodal (that is, text, speech, or gesture) and multiplatform (desktop, kiosk, or mobile) interfaces provides additional challenges, including the need to use potentially uncertain inputs from individual recognizers. Input mechanisms that are intuitive and can adapt to different users and situations (or automatically adapt) promise to mitigate access complexity and user training, increasing broad availability of knowledge access.

Retrieval

Retrieval technology has achieved approximately 80 percent precision with low recall (or 80 percent recall with low precision); by using relevance feedback, systems approach human-crafted queries.

The ability to leverage the advances in input processing—especially query processing—together with advances in content-based access to multimedia artifacts (text, audio, imagery, and video) promises to enhance the richness and breadth of accessible material while improving retrieval precision and recall. Dealing with noisy, large-scale, and multimedia data from sources as diverse as radio, television, documents, Web pages, and human conversations will also offer challenges, but advances in this area will enhance document retrieval precision and recall, ease navigational burden for users, and reduce search time.

Extraction

Extraction is the ability to identify and cull objects and events from multimedia sources. We currently can achieve 90 percent precision and recall when extracting named entities (people, organizations, or locations), 70 percent for relations among named entities (such as "father-of"), and 60 percent for events from text.

One challenge includes extracting entities within media and correlating those across it. This might include extracting names or locations from written or spoken sources and correlating them with associated elements within images. Whereas commercial products exist to extract named entities from text with precision and recall in the 90th percentile, domain-independent event extractors work at best in the 50th percentile, and performance degrades further with noisy, corrupted, or idiosyncratic data. Achieving better extraction will provide direct access to information or knowledge elements—including specific types that might be user preferred—and will let us reuse media elements, enabling user-tailored selection or presentations.

Question answering

The single best performing system today provides approximately 75 percent precision and recall for a small question-and-answer corpus. Drawing on techniques from query processing, retrieval, and presentation, this important new class of systems is moving us from our current form of Web searches (type in keywords to retrieve documents) to more direct natural language searches that are answered directly by an answer extracted from the source.

Challenges will include question analysis, response discovery, source selection, multi-perspectives, source segmentation, extraction, and semantic integration across heterogeneous sources of unstructured, structured, and semistructured data. Eventually, by providing direct answers to questions, we'll be able to overcome the time, memory, and attention limitations currently required to sift through many returned Web pages.

Translation

Last year, for the first time, English was estimated as constituting less than half the material on the Web. Some predict that Chinese will be the Web's primary language by 2007. Given that information will increasingly appear in foreign languages, there will be a need for systems to gist or skim content for relevance assessment—beyond the current ability to translate approximately 40 languages. We also need to improve current human-assisted machine translation to provide higher-quality translation for deeper understanding.

New innovative applications include the translation of multilingual conversations, rapid creation of translingual corpora, and effective translingual retrieval, summarization, and translation. Other applications will involve verbalizing graphics and visualizing text. Cross-media or cross-mode information and knowledge access will enable broader access to global information sources using methods such as translingual information retrieval.

Dialogue management

We currently can perform simple fact-seeking dialogues in specific domains such as weather, travel planning, and inventory. However, knowledge workers will require systems that can support natural, mixed-initiative human–computer interaction that deals robustly with context shift, interruptions, feedback, and shift of locus of control. Open research challenges include tailoring the flow and control of interactions and facilitating interactions such as error detection and correction tailored to individual physical, perceptual, and cognitive differences. Motivational and engaging lifelike agents also offer promising opportunities for innovation.

Agent and user modeling

Computers can construct models of user beliefs, goals, and plans. They can also model users' individual and collective skills by processing materials such as documents or user interactions and conversations.

While raising important privacy issues, unobtrusively modeling users (or groups of users) from public materials or conversations can enable a range of important KM capabilities. For example, this might include expertise databases that can enhance organizational awareness and efficiency or track user characteristics, skills, and goals to increase interaction and help users or agents find experts.

Summarization

Summarization aims to select content and condense it to present a compact form of the original source. Summaries can compress their content by 50 percent without losing information and can contain extracted information from or an abstract of original source material. They can be informative, indicative, or evaluative, and they offer knowledge workers access to larger amounts of material with less required reading.

Some related challenges include multimedia, multilingual, and cross-document summarization. Addressing scaleability to large collections and user- or purpose-tailored summaries is also an active research area.

Presentation

Effective presentations require selecting appropriate content, allocating content to appropriate media, and ensuring fine-grained coordination and realization in time and space. Discovering and presenting knowledge might require mixed media and mixed-mode displays tailored to the user and context. This could include tailoring content and form to the user's specific physical, perceptual, or cognitive characteristics. It also might lead to new visualization and browsing paradigms for massive multimedia and multilingual repositories that reduce cognitive load or task time, increase analytic depth and breadth, or simply increase user satisfaction.

We currently have highly complex systems with typically knowledge-rich methods of presentation planning and realization, but a grand challenge is to automatically generate coordinated speech, natural language, gestures, animation, and nonspeech audio, possibly delivered through interactive, animated, lifelike agents. Preliminary experiments suggest that independent of task performance, agents might be more engaging to younger or less experienced users.[6]

Awareness and collaboration

Our global Web provides unprecedented opportunity for worldwide collaboration, both asynchronously and synchronously. Users can use instant messaging and interact in place-based collaboration environments, but in the future they will need enhanced awareness of both emerging knowledge and one another's expertise. One such aid is detection and tracking of topics of interest to facilitate discovery and connection among communities. Another is creating expertise profiles based on publicly available information such as publications, interviews, or public conversations.

Human language technology promises to deliver great value to the KM challenge, but we need to address many fundamental scientific and technical challenges to ensure that this value accrues:

- *Heterogeneity*—dealing with the diverse nature of human language artifacts, both in form and semantic content.
- *Scalability*—addressing the size of corporate collections and global content.
- *Portability*—creating adaptive methods (such as corpus-based machine-learning approaches) that enable rapid retargeting of algorithms to new languages and media.
- *Complexity*—ensuring that the many content forms and presentational methods and devices do not overwhelm end users.
- *Security*—ensuring authentication to control access to source materials or ensuring the identity and integrity of source materials.
- *Privacy*—addressing the legal and social issues of maintaining privacy and control of a user's model extracted from public materials or interactions.

Overcoming these human language technology challenges is essential for KM to advance.

References

1. K. Bontcheva et al., "Using HLT for Acquiring, Retrieving and Publishing Knowledge in AKT," *ACL '01 Workshop on Human Language Technology and Knowledge Management*, ACM Press, New York, 2001, pp. 17–24.
2. I. Becerra-Fernandez, "Searching for Experts with Expertise-Locator Knowledge Management Systems," *ACL 01 Workshop Human Language Technology and Knowledge Management*, ACM Press, New York, 2001, pp. 9–16.
3. M. Maybury, R. D'Amore, and D. House, "Automating the Finding of Experts," *Int'l J. Research Technology Management*, vol. 43, no. 6, Nov./Dec. 2000, pp. 12–15.
4. M. Maybury, ed., "News on Demand," *Comm. ACM*, vol. 43, no. 2, Feb. 2000, pp. 33–34; www.acm.org/cacm/0200/0200toc.html (current 18 Oct. 2001).
5. A. Merlino, M. Maybury, and D. Morey, "Broadcast News Navigation Using Story Segments," *Proc. ACM Int'l Multimedia Conf.*, ACM Press, New York, 1997, pp. 381–391; www.acm.org/sigmm/MM97/papers/morey (current 18 Oct. 2001).
6. E. André and T. Rist, "Presenting Through Performing: On the Use of Multiple Lifelike Characters in Knowledge-Based Presentation Systems," *Proc. 2nd Int'l Conf. Intelligent User Interfaces* (IUI 2000), ACM, New York, 2000, pp. 1–8.

Acknowledgments

The efforts of several Mitre human language and knowledge management researchers and groups inspired the ideas described here, including Eric Breck, Ray D'Amore, David Day, Chris Elsaesser, Lynette Hirschman, David House, Manu Kanchady, Marc Light, Inderjeet Mani, Harry Sleeper, Cynthia Small, Jean Tatalias, and Alex Yeh. I also benefited from ideas from Wolfgang Wahlster, Elisabeth André, Thomas Rist, and Christoph Thomas.

Challenges in Information Extraction from Text for Knowledge Management

Fabio Ciravegna, *University of Sheffield*

Nowadays, most knowledge is stored in an unstructured textual format. We can't query it in simple ways, thus automatic systems can't use the contained knowledge and humans can't easily manage it. The traditional knowledge management process for knowledge engineers has been to manually identify and extract knowledge—a complex and time-consuming process that requires a great deal of manual input. As an example, consider the collection of interviews with experts (*protocols*) and their analysis by knowledge engineers to codify, model, and extract the knowledge of an expert in a particular domain. In this context, *information extraction* from texts is one of the most promising areas of human language technology for KM applications.

Information extraction

IE is an automatic method for locating important facts in electronic documents—for example, information highlighting for enriching a document or storing information for further use (such as populating an ontology with instances). IE thus offers the perfect support for knowledge identification and extraction, because it can, for example, provide support in protocol analysis in either an automatic (unsupervised extraction of information) or semiautomatic way (helping knowledge engineers locate the important facts in protocols through information highlighting).

It is widely agreed that the main barrier to using IE is the difficulty in adapting IE systems to new scenarios and tasks. Most of the current technology still requires the intervention of IE experts. This makes IE difficult to apply, because personnel skilled in IE are difficult to find in industry, especially in small-to-medium-size enterprises.[1] A main challenge is to enable personnel with knowledge of AI (for example, knowledge engineers) who have no or scarce preparation in IE and computational linguistics to build new applications and cover new domains. This is particularly important for KM. IE is just one of the many technologies for building complex applications: wider acceptance of IE will come only when IE tools don't require any specific skill apart from notions of KM.

Several machine learning-based tools and methodologies are emerging,[2,3] but the road to fully adaptable and effective IE systems is still long. Here, I focus on two main challenges for IE adaptivity in KM that are paramount in the current scenario: automatic adaptation to different text types and human-centered issues in coping with real users.

Adaptivity to text types

Porting IE systems means coping with four (often overlapping) main tasks:

1. *Adapting to the new domain information*—implementing system resources such as lexica, knowledge bases, and so forth, and designing new templates so that the system can manipulate domain-specific concepts.

2. *Adapting to different sublanguage fea tures*—modifying grammars and lexica to enable the system to cope with specific linguistic constructions that are typical of the application or domain.
3. *Adapting to different text genres*—specific text genres such as medical abstracts, scientific papers, and police reports might have their own lexis, grammar, and discourse structure.
4. *Adapting to different types*—Web-based documents can radically differ from newspaper-like texts. We need to be able to adapt to different situations.

Most of the literature on IE has focused on issues 1, 2, and 3, with limited attention to text types, focusing mainly on free newspaper-like texts.[4] This is a serious limitation for portability, especially for KM, where an increase in the use of Internet and intranet technologies has moved the focus from free-texts-only scenarios (based on, for example, reports and protocols) to more composite scenarios including semistructured and structured texts such as highly structured Web pages produced by databases. In classical natural language processing (NLP), adapting to new text types is generally considered a task of porting across different types of free texts.

Using IE for KM requires extending the concept of text types to new, unexplored dimensions. Linguistically based methodologies used for free texts can be difficult to apply or even ineffective on highly structured texts, such as the Web pages databases produce. They can't cope with the variety of extralinguistic structures such as HTML tags, document formatting, and stereotypical language that convey information in such documents. On the other hand, wrapper-like algorithms designed for highly structured HTML pages are largely ineffective on unstructured texts (for example, free texts). This is because such methodologies make scarce or no use of NLP, usually avoiding any generalization over the flat word sequence and tending to be ineffective on free texts, because of, for example, data sparseness.[5]

The challenge is developing methodologies that can fill the gap between the two approaches and cope with different text types. This is particularly important for KM with its composite Web-based scenarios, because Web pages can contain documents of any type and even a mix of text types—an HTML page, for instance, can contain both free and structured texts. Work on this topic has just started.

Wrapper induction systems based on lazy NLP[5] try to learn the best and most reliable level of language analysis useful for a specific IE task by mixing deep linguistic and shallow strategies. The learner starts inducing rules that make no use of linguistic information, such as in wrapper-like systems. It then progressively adds linguistic information to its rules, stopping when the use of NLP information becomes unreliable or ineffective. Generic NLP modules and resources provide linguistic information that is defined once and is not to be modified to specific application needs by users. Pragmatically, the measure of reliability here is not linguistic correctness (immeasurable by incompetent users) but effectiveness in extracting information using linguistic information as opposed to using shallower approaches.

Unlike previous approaches in which different algorithm versions with different linguistic competence were tested in parallel and the most effective version was chosen,[6] lazy NLP-based learners learn which is the best strategy for each information or context separately. For example, they might decide that parsing is the best strategy for recognizing the speaker in a specific application on seminar announcements but not the best strategy to spot the seminar location or starting time. This is promising for analyzing documents with mixed genres.

Coping with non-IE experts

The second main task in adaptive IE concerns human–computer interaction during application development. Nonexpert users must be supported during the entire adaptation process to maximize the final application's effectiveness and appropriateness. A typical IE application's life cycle is composed of scenario design, system adaptation and results validation, and application delivery.[7]

Scenario design

Scenario design defines the information to extract. Many potential users need specific support, because they might find it difficult to manipulate IE-related concepts such as templates. Moreover, there might be a gap between the information the user needs, the information the texts contain, and what the system can actually extract. It is thus important to help users recognize such discrepancies, forcing them into the right paradigm of scenario design. Highlighting information in different colors is generally a good approach. Tag-based interfaces, such as MITRE's Alembic, have proven to be effective and have become a standard in adaptive IE.

Selecting the corpus to be tagged for training is also a delicate issue. Nonlinguistically aware users tend to focus on text content rather than on linguistic variety. Unfortunately, IE systems learn from both. Provided corpora might be unbalanced with respect to types or genres (emails could be underrepresented with respect to free texts) or might show peculiar regularities because of wrong selection criteria. For example, in designing an application on IE from professional resumes, our user selected the corpus by using the names of US cities as keywords. When the trained system was tested, it became clear that most of the resumes actually originated from Europe, where addresses, titles of study, and even text style can significantly differ from US styles. The resulting system was therefore largely ineffective and left the user dissatisfied with the final application.

A number of methodologies can be used to validate the training corpus with respect to a (hopefully big) untagged corpus. One possible validation concerns the formal comparison of training and untagged corpus. Adam Kilgarriff proposes heuristics for discovering differences in text types among corpora.[8] Average text length, distributions of HTML tags and hyperlinks in Web pages, average frequency of lexical classes in texts (such as nouns), and so forth can be relevant indicators of corpus representativeness and can warn inexperienced users that some training corpora might not sufficiently represent the whole corpus. Even detecting an excess of regularity in the training corpus can indicate an unbalanced corpus selection. For example, if a high percentage of fillers for some slots is the same string (such as "ACME Inc."), there is the concrete risk that the corpus contains some unwanted regularity that could influence the learner in an unpredictable way.

System adaptation and results validation

With a corpus reasonable in size and quality, the IE system can then be trained. Unfortunately, even the best algorithm is unlikely to provide optimized results for specific use. This is because a 100 percent accurate system is out of reach for current IE technology, and therefore we must balance recall (the ability to retrieve information when present) and precision (the ratio of correct information on the total of information extracted) to produce the optimal results for the task and users at hand. Different uses will require different types of results—higher recall in some cases, higher precision in others. Users must be enabled to evaluate results from both a quantitative and qualitative point of view and to change the system behavior if necessary.

Most of the current technology provides satisfying tools for results inspection: tools such as the MUC scorer let users understand the system effectiveness in detail.[9] The challenging step now is to let users change system behavior. In case of occasional or inexperienced users, the issue of avoiding technical or numerical concepts such as precision and recall arises. This requires the IE system to bridge the user's qualitative vision ("you are not capturing enough information") with the numerical concepts the learner can manipulate—for example, moving error thresholds to obtain higher recall.

Application delivery

When the application is tuned to specific user needs, it can be delivered and used in the application environment. Corpus monitoring should be enabled even after delivery, though. One of the risks in highly changing environments such as the Internet is that information such as Web pages can change format in a short period of time, and the system must be able to detect such changes.[10] The same techniques mentioned earlier for testing corpus representativeness can identify changes in the information structure or test type.

Adaptive IE is already providing useful results and technology for KM. Fully integrated user-driven solutions are still to come, but current research results are promising.

Acknowledgments

The author's work on adaptive IE is supported under the Advanced Knowledge Technologies (AKT) Interdisciplinary Research Collaboration (IRC), sponsored by the UK Engineering and Physical Sciences Research Council (grant GR/N15764/01). AKT comprises the Universities of Aberdeen, Edinburgh, Sheffield, and Southampton, and the Open University (www.aktors.org).

References

1. F. Ciravegna, A. Lavelli, and G. Satta, "Bringing Information Extraction Out of the Labs: The Pinocchio Environment," *Proc. 14th European Conf. Artificial Intelligence*, IOS Press, Berlin, 2000, pp. 416–421.
2. D. Freitag, "Information Extraction from HTML: Application of a General Learning Approach," *Proc. 15th Nat'l Conf. Artificial Intelligence* (AAAI-98), AAAI Press, Menlo Park, 1998, pp. 517–523.
3. F. Ciravegna, "Adaptive Information Extraction from Text by Rule Induction and Generalisation," *Proc. 17th Int'l Joint Conf. Artificial Intelligence* (IJCAI 2000), Morgan Kaufmann, San Francisco, 2001, pp. 1241–1256.
4. C. Cardie, "Empirical Methods in Information Extraction," *AI Journal*, vol. 18, no. 4, Winter, 1997, pp. 65–79.
5. F. Ciravegna: "(LP)2, an Adaptive Algorithm for Information Extraction from Web-Related Texts," *Proc. IJCAI 2001 Workshop Adaptive Text Extraction and Mining*, AAAI Press, Menlo Park, Calif., 2001, pp. 9–17.
6. S. Soderland, "Learning Information Extraction Rules for Semi-Structured and Free Text," *Machine Learning*, vol. 34, no. 1, Feb. 1999, pp. 1–44.
7. F. Ciravegna and D. Petrelli, "User Involvement in Adaptive Information Extraction: Position Paper," *Proc. IJCAI 2001 Workshop Adaptive Text Extraction and Mining*, AAAI Press, Menlo Park, Calif., 2001, pp. 71–73.
8. A. Kilgarriff, "Comparing Corpora," to be published in *Int'l J. Corpus Linguistics.*
9. A. Douthat, *The Message Understanding Conference Scoring Software User's Manual*, www.itl.nist.gov/iaui/894.02/related_projects/muc/muc_sw/muc_sw_manual.html (current 18 Oct. 2001).
10. I. Muslea, S. Minton, and C. Knoblock, "Co-testing: Selective Sampling with Redundant Views," *Proc. 17th Nat'l Conf. Artificial Intelligence* (AAAI 2000) AAAI Press, Menlo Park, Calif., 2000.

Question-Answering Systems in Knowledge Management

Dan Moldovan, *University of Texas at Dallas*

At no time in history have so many people been able to access so much information. Businesses rely on unstructured information available over the Internet, intranets, email, press releases, online newspapers, digital libraries, and other sources. Companies accumulate large quantities of written information with customer comments, trade publications, internal reports, competitor Web sites, and much more. Making sense of all this information and leveraging its advantages is crucial.

New companies were formed with the purpose of helping other companies cope with this huge volume of information. These knowledge-supporting companies provide skills for integrating document management, workflow, workgroups, intranets, and knowledge portals. They might be general purpose or industry specialized, serving vertical markets such as financial, IT, telecommunications, travel, or media. Other companies have extended or created their own internal information management departments.

Yet, the technology to access and use information has dramatically lagged behind its growth rate. People have questions, and they need answers. Current Internet search engines let us locate documents that might have relevant information, but often the documents returned are too numerous to inspect or the answer simply isn't there.

This has motivated the renewed interest in question-answering technology and natural language processing. The task in QA is to find correct answers to open-domain questions expressed in English or other natural languages by searching large collections of documents. These documents can come from a text collection, the Web, databases, digital libraries, or any other electronic source.

QA technology's potential

QA technology will undoubtedly play a major role in knowledge management. Users can include casual questioners who ask simple factual questions; consumers who look for specific product features or prices; research analysts who are in the business of collecting specific information about market, competitors or business events; or professional information analysts such as police detectives, law enforcement officials, financial analysts, or intelligence analysts.

The main difficulty QA technologists face is the broad range of questions a system must answer. Simple questions are relatively easy to answer. These are often the *Who*, *When*, *Where* types of questions:

- Who was the first American in space?
- When did the Neanderthal man live?
- Where is John Wayne airport?

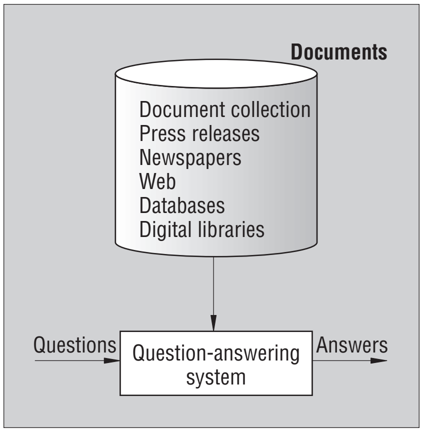


Figure 4. A high-level view of a QA system.

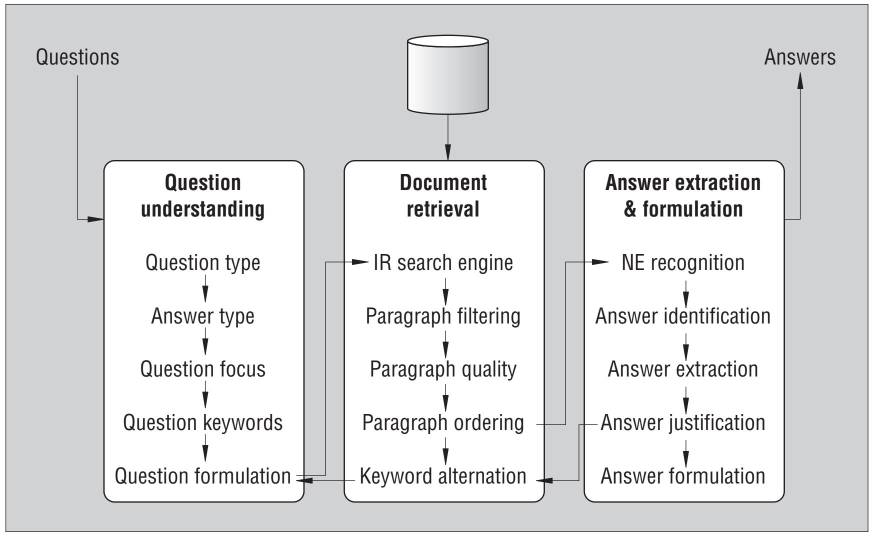


Figure 5. A QA system's architecture.

Other questions require reasoning on knowledge bases, collecting pieces of evidence from multiple documents, and then combining together the answer, or other advanced AI techniques. Examples include (where A and B can be concepts or relations):

- How is A related to B?
- What causes A?
- What are the effects of A?
- What can damage A?
- How do you prevent A?

Even more difficult questions exist, which require considerable world knowledge and powerful reasoning capability. For example, to answer, "How likely is it that the federal government will lower interest rates next month?" the system would first have to find out what usually influences the government's decision and then compare the status of such parameters with previous situations for which the outcome is known.

QA technology alone is not enough to provide solutions to information-management problems. It is merely a powerful tool that, when embedded into other larger software systems currently used for information management, can let companies deliver fast, effective, and affordable results. It also lets companies analyze textual documents, extract desired information, rank and link important concepts, and personalize information. Furthermore, it might help with automation within specific department, marketing, or customer support. Marketing departments might use QA to learn trends, discover customer interests, or identify customers most likely to buy specific products and services. Customer-support departments might use QA to give consumers better and faster product information and services, browse through technical information, or extract customer profiles.

State-of-the-art QA systems

The US government has started a QA competition under TREC (the *T*ext *Re*trieval *C*onference, http://trec.nist.gov) and recently launched a new research initiative called AQUAINT (*A*dvanced *Qu*estion *A*nswering for *Int*elligence). The state of the art in QA technology is clear from the best systems that participated in these competitions.

A QA system receives questions, searches for answers over a large variety of document sources, and extracts and formulates concise answers. Figure 4 lists the type of documents containing answers.

There are three essential modules in almost all QA systems: *question understanding*, *document retrieval*, and *answer extraction and formulation* (see Figure 5).

The question understanding module determines the type of question and the type of answer expected, builds a focus for the answer, and transforms the question into queries for the search engine. The better the system understands the question intention, the easier it is to extract the answer. To find the right answer from a large collection of texts, first we must know what to look for. The answer type can usually be determined from the question. To better detect the answer, the system first classifies the question type: what, why, who, how, where, and when.[1]

However, the question type is not sufficient for finding answers. For example, with the question, "Who was the first American in space?" the answer type is obvious: *person*. However, this does not apply for questions asking "what," because those questions are more ambiguous. The same applies to many other question types; we solve this problem by defining a concept called a *focus*.

A focus is a word or a sequence of words that define the question and disambiguate it by indicating what the question is about. For example, for "What is the largest city in Germany?" the focus is largest city. Knowing the focus and the question type helps determine the type of answer sought—namely, the name of the largest city in Germany.

The document retrieval module is a search engine that extracts relevant documents from a collection of documents. After extracting relevant documents, it identifies paragraphs containing potential answers. This decreases the amount of text it must parse and analyze when extracting answers. Some systems can measure the quality of the paragraphs, discard or add some keywords as needed, and prioritize them.

The answer extraction and formulation module finds one or more pieces of information that eventually are used to formulate the answer. We must rely on lexico-semantic information, provided by a parser that identifies named entities, monetary units, dates and temporal or locative expressions, and products. Recognizing the answer type, through the semantic tag returned by the parser, creates a candidate answer. The module bases its answer extraction and evaluation on a set of heuristics. The better the search engine can narrow down the amount of text with answers, the less work the module must perform. Some systems implement elaborate answer justification modules that perform logic proofs.[2]

Technical challenges

Open-domain QA is a complex task that encompasses many natural language processing, information retrieval, and AI techniques. The inherent difficulties in finding answers to open-domain questions pose many serious technical challenges. The Roadmap Research in Question Answering document has identified these challenges (see www.nlpir.nist.gov/projects/duc/roadmapping.html), some of the most important being

- understanding questions including ambiguities and implicatures,
- understanding questions and finding answers within a given context,
- extracting distributed answers that require answer fusion,
- providing answer justification and proof of correctness,
- offering interactive question answering,
- offering real-time question answering, and
- extracting answers from a wide range of document formats.

QA has attracted considerable interest in the last few years as government research initiatives have been launched in this area. The most performant QA systems today can extract single facts from a large collection of documents but can't answer questions that require answer fusion. Advanced QA systems need richer semantic resources and the capability of online ontology development. There is a strong interrelation between QA and text mining, because one can benefit the other.

To answer, "What software products does Microsoft sell?" the system must first find out what constitutes software products and then check whether Microsoft sells such products. Unless an ontology of software products exists in the knowledge base—which is highly unlikely—the system must first acquire from the document collection a definition of software products.

Dynamic ontologies built ad hoc from the text collection, coupled with existent ontologies, are the best path to follow in answering more difficult questions. To address questions of higher degrees of difficulty, we need real-time knowledge acquisition and classification for different domains.

QA and other technologies

Adaptive IE does not have much in common with QA technology. Although there were attempts to build QA systems around IE engines, open-domain QA is quite different from IE. However, hyperlinking enables QA on the Web—that is, extracting the answer from Web documents.

More importantly, summarization plays a key role in answer formulation. Consider, for example, a question that asks for an explanation. The QA system might extract pieces of evidence from multiple documents, after which it must formulate a coherent answer. This is performed using planning and text generation—techniques frequently used in summarization. Summarization technology complements QA technology, and by combining them, we can build better tools for delivering quality information.

A natural extension of QA technology is voice-activated QA. The advantage is relatively few voice-to-text and text-to-voice conversions needed, because QA operating on text documents will continue to perform the bulk of the processing. There are, however, inherent difficulties in transforming voice questions into text. One is the recognition of proper names. Because voice recognition systems have a limited vocabulary, proper names in open-domain questions are not well recognized. Thus, important keywords are missing and the QA system's accuracy significantly decreases. However, integrating voice recognition in the QA technology is extremely promising, as it opens new application domains.

References

1. D. Moldovan et al., "The Structure and Performance of an Open-Domain Question Answering System," *Proc. 38th Ann. Meeting of the Assoc. Computational Linguistics* (ACL 2000), Morgan Kaufman, San Francisco, 2000.
2. S. Harabagiu et al., "FALCON: Boosting Knowledge for Answer Engines," *Proc. 9th Text Retrieval Conf.* (TREC-9), NIST, Gaithersburg, Md., 2000, pp. 50–59.

Knowledge Management and Mobile Voice Interfaces

Georg Niklfeld, *Telecommunications Research Center, Vienna*

Information technology has already made quick and easy access to remote data and information possible. An individual sitting in an office in front of a desktop PC can cross boundaries of space over a network of computers to retrieve data from another computer or communicate electronically with another person sitting in an office across the globe. At first, this possibility provided a quantitative improvement of information access, but it has also triggered qualitatively new patterns of work and knowledge exchange, as in virtual teams.

When the entry-points to the networks become mobile and independent of the office environment, spatial restrictions of data access relax. This enables mobile-data applications—for example, in the form of mobile-data services in the upcoming third-generation telecommunication networks. Existing and planned end-user terminals for mobile application scenarios vary in size and available input and output channels, ranging from notebooks, PDAs, smart phones, and WAP phones to plain telephones without display. In general, the smaller and simpler the terminal, the more impoverished the visual input and output. Many small terminals also lack a comfortable keyboard for alphanumeric input.

These restrictions of conventional visual-interface techniques raise the attractiveness of voice communication—not only of person-to-person voice telephony, which will remain preeminent as a mobile communication type, but also of voice interfaces that use automatic speech recognition (ASR) or text-to-speech (TTS) to provide access to data applications. We can use such voice-enabled mobile data applications even when our hands or eyes are busy—for example, when driving a car. Mobile data services for use in cars (referred to as *telematics*) have emerged as one of the most promising markets for the wireless Internet. This is a real economic opportunity to apply speech technology, and general human language technology, to the potentially huge markets of telematics, mobile data applications, and 3G data services.

Some of the key technical issues for mobile voice interfaces have been resolved or are close to resolution. The mobile terminals do not need to support specialized voice processing beyond a standard voice call. A dedicated telephony server in the infrastructure picks up the voice call. In a speech-in, speech-out application, the telephony server sends the speech to an ASR server, which uses hidden Markov models to implement speaker-independent LVCSR (large vocabulary continuous speech recognition). The results from ASR, usually in the format of ASCII text, are passed on for further linguistic processing.

In existing systems, this linguistic processing is often rudimentary. The input constitutes the user's choice from among a dozen or so options specified in a command grammar. Each recognized input text triggers a system response, which is either predefined text or a text template that contains dynamic information from a database query. The system module that produces this response is a *dialog manager*. The text output combining static and dynamic parts is processed by a TTS engine, which typically employs concatenative synthesis. This approach to speech synthesis searches for the best-fitting speech segments for a computed output target in a large speech database and smoothes the transitions between concatenated segments. Finally, the telephony server plays back the resulting speech output to the user through the voice call.

An interesting alternative architecture for ASR is distributed speech recognition, which performs the first ASR computation steps (feature extraction) at the mobile terminal. This makes it possible to use an optimized, error-robust encoding of the input speech over the air interface of a wireless communication network.

The relative maturity of these technologies and architectures has led to the development of standardized development platforms. The VoiceXML standard has found widespread support in industry.[1] It defines an architecture with modular ASR and TTS, plus an XML-based application language for defining the application-specific dialog manager. VoiceXML's key design objective is to bring voice application development in line with Web development and its simple programming models. This should enable the large base of general Web programmers to port their applications and interfaces to voice without involving hard-to-find specialists for telecom and voice system implementation. Consequently, the range of companies for which developing voice interfaces to their applications is economically feasible should become much larger.

As voice-enabling therefore becomes an issue for many organizations, a new issue for information-systems planning arises, which we should view in the wider KM context. An organization should ask itself the following questions: Is there a significant group of users of in-house IT applications that would benefit from mobile access? Are their typical mobile usage situations such that voice access would be advantageous? If yes, is this a long-term issue, and are there software development resources available that would warrant moving offensively into this area? If all answers are positive, an organization should try to build voice interface competence for its IT application portfolio—inline with the approach that the organization has chosen for other IT-enabled KM activities such as its Web presence, intranet, and extranet.

Multimodality

To avoid repeating previous mistakes made by some human language technology advocates, it is nevertheless important to stress that visual interfaces will also play an important role in mobile data access. Visual interfaces are easy to implement and familiar to users. In addition, for some tasks, they are better suited than voice—for example, to represent spatial information using images.

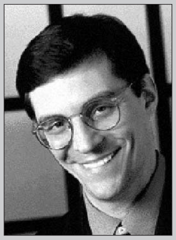
One promising route for research is therefore to develop interfaces that combine visual elements and voice into so-called multimodal interfaces. The two modalities can then be offered as alternatives, so that the user can choose between a visual or a voice mode, depending on the situation. A second mode of combining voice and visual is to have both modalities active concurrently, without explicit coordination. However, only when coordinated, simultaneous use of the voice and visual channels is supported is the potential for a multimodal interface exploited fully. For example, using a terminal that offers visual input through a touchscreen, a user could draw a number of irregular shapes around areas on a city map in a tourism application and say, "Tell me about hotels in these areas." The system would resolve the reference of "these" to the set of selected areas in the present time interval. It might then insert icons for the selected hotels on the map and read through TTS the names and details of the hotels that pay a registration fee to the service provider. While reading the data for a hotel, the corresponding icon could blink on the display.

In contrast to voice-only interfaces, technologies for mobile multimodal interfaces to data services are still at the research stage. The US company AUVO has started field tests of multimodal services on a GPRS-platform, together with Spanish telecom operator AirTel Móvil.[2] Their technology is based on a proprietary coding scheme that must be implemented on the end-user terminal and in the infrastructure to bundle data and packet speech into a common data stream. The protocol working group within the Aurora project at the European Telecommunications Standards Institute has developed a blueprint for multimodal interfaces based on distributed speech recognition.[3] The W3C is preparing a working group on a multimodal dialog markup language comparable to VoiceXML.[4] Most recently, a number of leading companies in the computer industry have set up the SALT consortium to define a set of speech tags that will be supported in HTML and XML documents (www.saltforum.org).

Major players in the field are driving all of these efforts. Each requires substantial dedicated protocol engines in mobile terminals, which a critical mass of equipment manufacturers must provide in a compatible way. It remains to be seen whether any of these proposals will find sufficient support in the industry to make such an approach viable. In our own research at the Telecommunications Research Center, Vienna, we follow a different path based on the assumption that most likely there will not be a successful universal standard for multimodal interfaces.

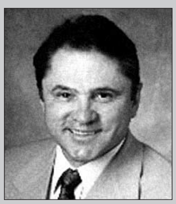
We are working on an architecture and a development model for multimodal interfaces that uses existing technologies and standards so that developing multimodal interfaces for real applications can start immediately. Another requirement that we set is that the development model should be easy to use, especially for general Web programmers. These attributes can help promote a broad base of multimodal interfaces to mobile data applications. Because our institution is a partly industry-financed telecommunications research center, we are also interested in creating foundations for attractive data services for 3G wireless networks, which are often cited as important success factors for this expensive technology.

We have developed a modular architecture that uses HTTP-based visual interface protocols such as HTML or WAP in combination with Java applets and VoiceXML for the voice interface specification.[5] The data streams for the two modalities are transported independently between the mobile terminal and the infrastructure. A dedicated server in the infrastructure, which in turn interfaces to another server that hosts the application logic, performs multimodal integration. Using this approach, we can build simple multimodal interfaces to existing data services. We have already completed a demonstrator that adds a multimodal interface to an existing route-finder application on the Web. To do this, we



Mark Maybury is the executive director of MITRE's Information Technology Division. He has organized multiple international symposia, given tutorials, and published over 50 technical and tutorial articles in the areas of language generation, multimedia presentation, text summarization, and intelligent multimedia information retrieval. He received his M.Phil. in computer speech and language processing and his PhD in artificial intelligence from Cambridge University. He is a member of the Board of Directors of the Object Management Group and is Treasurer/Secretariat for ACM Sigart. Contact him at maybury@mitre.org.

Fabio Ciravegna is a senior research fellow in the Department of Computer Science at the University of Sheffield, UK. His research interests include classical and adaptive information extraction from text, with a particular focus on user-centered methodologies. His work on information extraction has been published at a number of international conferences such as EACL 1999, IJCAI 1999, ECAI 2000, and IJCAI 2001. He coorganized two workshops on the use of Machine Learning for Information Extraction at ECAI 2000 and IJCAI 2001. Contact him at f.ciravegna@dcs.shef.ac.uk.



Dan Moldovan is a professor in the Department of Computer Science at the University of Texas at Dallas. He is also the director of the Human Language Technology Research Center at UTD. His research interests are in natural language processing, particularly in question-answering technology, knowledge acquisition, and knowledge bases. He earned a PhD in electrical engineering and computer science from Columbia University, New York. He founded Language Computer Corporation (www.languagecomputer.com), a company that specializes in question answering, information extraction, and information retrieval research and products. He is a senior member of the IEEE and a member of the ACM and the ACL. Contact him at moldovan@utdallas.edu.

Georg Niklfeld is a senior researcher at the Telecommunications Research Center, Vienna. His interests are applications of human language technology in telecommunications and knowledge management, as well as multimodal systems on small devices. He studied computer science at the TU Vienna, linguistics and philosophy at the University of Vienna, and technology management at the University of Manchester Institute of Science and Technology, UK. Contact him at niklfeld@ftw.at.

did not need access to the existing service but simply attached the new multimodal interface to the CGI interface of the old visual-only service.

While this first demonstrator used a simple type of multimodality, where only one modality is active at any moment, we are currently testing our architecture on a second, more ambitious demonstrator. Here, the user can make inputs that combine events at the visual interface and the voice interface on the basis of time-stamping information. The application is a map-based information and routing service for various types of content including cinemas, pharmacies, drug stores, and gas stations. The visual interface specifies spatial concepts such as map areas or end points for desired routes or selects individual facilities on the map. To express content-specific commands involving variable numbers of objects, either voice or visual can be used, but we expect users to prefer voice for the specification of the commands and a combination of voice and visual for the specification of objects to which a command should apply.

Voice and KM

Due to robustness issues of voice recognition technology, voice interfaces should be used in contexts and applications where they have a chance to work well. Not all applications in IT-based knowledge management appear suitable. For example, voice interfaces to unrestricted search services are problematic because of the huge branching factor in recognition—that is, the number of possible alternatives when recognizing the search terms. Yet, I am convinced that voice-only and combined voice and visual interfaces to applications can be useful beyond "nice to have" in situations that favor them, especially in mobile situations and on small communication terminals. Putting these interfaces in place should be a concern for knowledge-intensive organizations with mobile workers. Whether the resulting new types of information access will also give rise to qualitatively new types of knowledge exchange will be interesting to watch.

Acknowledgments

This research at the Telecommunications Research Center Vienna is supported within the Austrian Competence Center Program Kplus, and by the companies Alcatel, Connect Austria, Kapsch, Mobilkom Austria, and Nokia.

References

1. *Voice eXtensible Markup Language (VoiceXML) version 1.0*, World Wide Web Consortium 2000, www.w3.org/TR/2000/NOTE-voicexml-20000505 (current 23 Oct. 2001).

2. "Auvo Technologies and Vodafone's AirTel Móvil to Trial Multi-Modal Wireless Internet Platform," AUVO, Itasca, III., 2001, www.auvo.com/frmset5.html (current 23 Oct. 2001).

3. "Distributed Speech Recognition in 3GPP," ETSI 2001, Sophia Antipolis, France, www.etsi.org/tbnews/0107_dsr-3g.html (current 23 Oct. 2001).

4. "Next Steps for W3C work on Multimodal Standards," World Wide Web Consortium 2001, www.w3.org/Voice/multimodal-rfp.html (current 23 Oct. 2001).

5. G. Niklfeld, R. Finan, and M. Pucher, "Component-Based Multimodal Dialog Interfaces for Mobile Knowledge Creation," *Proc. EACL/ACL 01 Workshop Human Language Technology & Knowledge Management*, ACM Press, New York, 2001, pp. 121–128.