Human Language Technologies for Knowledge Management: Challenges and Opportunities

Mark Maybury

Information Technology Division
The MITRE Corporation
202 Burlington Road
Bedford, MA 01730, USA
maybury@mitre.org
www.mitre.org/resources/centers/it

Abstract

This paper outlines the central role a range of human language technologies play in the emerging discipline of knowledge management. We articulate several grand challenges, illustrate some early successes, and recommend areas of continued research.

1 Introduction

In the past few years, knowledge management (KM) has received increasing attention from industry, academia, and government. Effective KM is often cited as a key capability for competitive advantage for global enterprises. Human language technology (HLT) plays a central role in KM, as this article aims to elucidate.

2 Knowledge Management Challenges

Knowledge Management is the field of enhancing organizational performance through organizational knowledge sharing, learning, and application of expertise. As an indicator of the importance of knowledge management, many corporations that traditionally measured only the financial aspects of value are beginning to measure human and intellectual value as well.

Knowledge management can be enabled by a range of human language technologies including but not limited to enhanced information retrieval, extraction, summarization, and presentation/generation. Moreover, human language technologies promise to both enhance human's access to information as well as to enhance their interaction with one another such as by increasing their awareness of knowledge artifacts or activities intersecting their interests. Key elements of KM include cataloguing existing knowledge, discovering expertise, and creation of new knowledge, which we briefly discuss in turn.

2.1 Knowledge Mapping

Often times a primary issue for organizations is knowing 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, so much material is created in an organization that its effective organization is daunting. Tools are required that can automatically generate classifications and/or taxonomies of explicit corporate and world knowledge. The success of services such as Yahoo, Northernlight, and Quiver illustrate the value of (and limitations of) current classification-based collections and collaborative filtering approaches.

2.2 Expert and Community Discovery

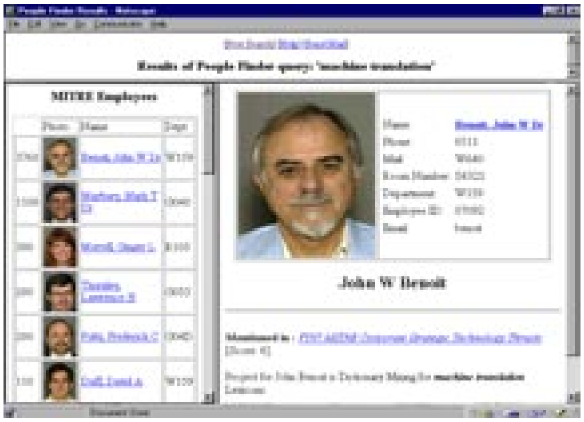
Knowing who to call, who knows a key fact, or who has the know-how or the skill to analyze, diagnose, or recommend in a particular domain is a challenge. Traditional manually-created corporate skills databases are costly, inconsistent across individuals and disciplines, and become rapidly obsolete. Finding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

2.3 Knowledge Discovery

Knowledge discovery is the facilitation of end users or communities of experts directly or collectively in discovering answers to their questions. Accomplishing this can include question answering, discovery of new knowledge through machine learning or data mining, and ultimately new knowledge learning, including ontology induction.

3 Preliminary Results

Human language technologies have already been applied to some key knowledge management areas (Bontcheva 2001). For example, Becerra-Fernandez (2001) reports on NASA's expert finder. In our research we have created systems which automatically extract and correlate information from human created artifacts to create assessments of human expertise (Maybury, D'Amore, House, D., 2000). For example, Figure 1 illustrates the screen MITRE's Expert Finder (also called People Finder) generates after a user types in the keywords "machine translation". Expert Finder processes employee published resumes and documents as well as corporate newsletters that mention individual's names in the context of this topic (using named entity extraction) 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.



Results of People Finder query: 'machine translation'

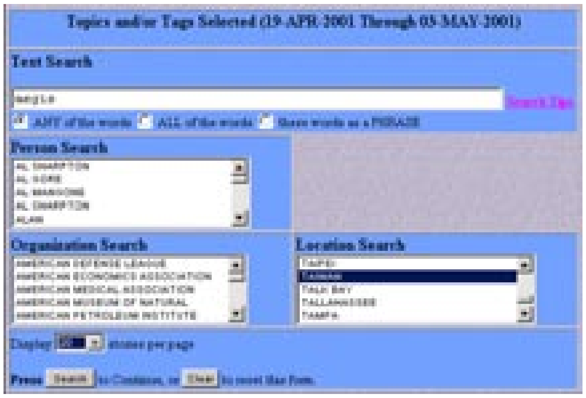
MITRE Employees

John W Benoit

Figure 1. Expert Finder

In spite of low inter-human agreement in determining expertise, an empirical evaluation comparing ten 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% precision and 40% recall when appropriate data was available. This is sufficient performance to find an expert within one phone call, which was the original knowledge management objective.

While we mentioned in Section 2.1 the value of taxonomic search engines, another area of preliminary success is in knowledge discovery. Several research groups are working to create more effective means to access multimedia information sources (Maybury 2000). Figure 2 illustrates MITRE's Broadcast News Nagivator (BNN), the culmination of many years of research and an integration of multiple human language and other technologies (Merlino et al. 1997). BNN applies speech, language, and image processing methods to segment, extract, and summarize broadcast news sources to enable personalized and targeted search of the news. Figure 2 shows the user querying stories from all sources for the last two weeks (19 April to 3 May 2001 at the time of writing) containing the keyword "aegis" and the location "Taiwan". People, organization, and locations menus are dynamically generated from extracted entities.



Topics and/or Tags Selected (19-APR-2001 Through 03-MAY-2001)

Text Search

aegis

Search Tips

● ANY of the words ○ ALL of the words ○ these words as a PHRASE

Person Search

AL SHARPTON
AL GORE
AL MARONE
AL SHARPTON
ALAN

Organization Search

AMERICAN DEFENSE LEAGUE
AMERICAN ECONOMICS ASSOCIATION
AMERICAN MEDICAL ASSOCIATION
AMERICAN MUSEUM OF NATURAL
AMERICAN PETROLEUM INSTITUTE

Location Search

TAIPEI
TAIWAN
TALK BAY
TALLAHASSEE
TAMPA

Display stories per page

Press Search to Continue, or Clear to reset the form.

Figure 2. Broadcast News Navigator

The results of the Figure 2 search are show in Figure 3 which includes 52 stories mentioning "Taiwan" and "Aegis" war ships on multiple programs (e.g., C-SPAN, Fox News at 6pm, 9pm and 10 pm, CNN Headline News, CNN

Morning Headline, CNN World Today, CNN World View, and CNN Moneyline). As shown in Figure 3, BNN presents a quick skim including a keyframe and the top 3 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, 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.

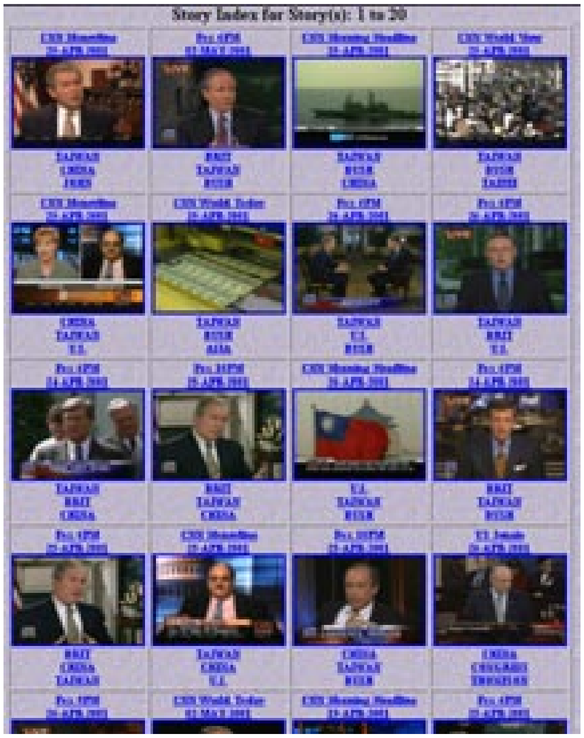


Figure 3. Broadcast News Navigator

4 HLT for KM

Having considered the needs of knowledge management and some preliminary promise of human language technology to provide solutions to these needs, we now more systematically analyze how HLT can contribute to KM. Table 1 outlines a range of functional areas of human language technology that offer potential solutions to some of the elements required to enable knowledge management. We consider the following ten areas in turn: input analysis, information retrieval, information extraction, question answering, translation, dialogue management, user modeling, summarization, presentation generation and awareness/ collaboration.

4.1 Input Analysis

Analysis of 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. Challenges include dealing with imprecise, ambiguous, and/or partial input. Addressing these challenges in multimodal (e.g., text, speech, and gesture) and/or multiplatform (desktop, kiosk, mobile) interfaces provides additional challenges. Input mechanisms that are intuitive as well as user and situation adaptable or automatically adaptive promise to mitigate complexity and increase broad availability of knowledge access.

4.2 Retrieval

The ability to leverage the above advances in input processing (especially query processing) together with advances in content-based access to multimedia artifacts (e.g., text, audio, imagery, video) promises to enhance the richness and breadth of accessible material while at the same time 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 (e.g., chat sessions and speech transcriptions) will offer challenges.

4.3 Extraction

Extraction is the ability to identify and cull out objects and events from multimedia sources (text, audio, video). An example challenge includes extracting entities within media and correlating those across media. For example, this might include extracting names or locations from written/spoken sources and correlating those with associated images. Whereas commercial products exist to extract named entities from text with precision and recall in the ninetieth percentile, domain independent event extractors work at best in the fiftieth percentile and performance degrades further with noisy, corrupted, or idiosyncratic data.

Table 1. Human Language Technology for Knowledge Management

Human Language Technology	Grand Challenges	Benefits to Knowledge Discovery, Access, Exploitation
Input/Query Analysis	Interpretation of imprecise, ambiguous, and/or partial multimodal input. Facilities include spoken query processing, visual query analysis (e.g., sketching), mixed media query (e.g., text and graphics)	Natural (written, spoken, gestural) access to information and knowledge. Decrease in access complexity or user training. Broaden availability of knoweldge to users.
Retrieval	Natural language processing of queries and documents. Content-based retrieval of text, imagery, audio, video.	Enhancements to document retrieval precision and recall. Direct access to media, easing navigational burden of user. Reduction of search time.
Extraction	Segmentation, object and event identification, and extraction from multimedia sources (text, audio, video).	Direct access to information or knowledge elements including specific types that may be user preferred. Reuse of media elements enabling user tailored selection or presentations.
Question Answering	Question analysis, response discovery and generation from heterogeneous sources.	Overcome time, memory or attention limitations required to sift through many returned web pages from a traditional search by providing direct answers to questions.
Translation	Rapid creation of translingual corpora. Effective translingual retrieval, summarization, and translation. Access Verbalization of graphics, Visualization of text.	Cross media/mode information and knowledge access enabling broader access to global information sources using methods such as translingual information retrieval.
Dialogue Management	Mixed initiative natural interaction that deals robustly with context shift, interruptions, feedback, and shift of locus of control	Ability to tailor flow and control of interactions and facilitate interactions. Includes error detection and correction tailored to individual physical, perceptual and cognitive differences. Motivational and engaging life-like agents.
Agent/User Modeling	Unobtrusive learning, representation, and use of characteristics, beliefs, goals and plans of agents (including the user).	Enables tracking of user characteristics, skills and goals in order to enhance interaction as well as discovery of experts by other users or agents.
Summarization	Scaleability, cross-linguality, multimedia summarization.	Increasing speed of reviewing materials. Multimedia summarization, cross-lingual summarization, large multi-document summarization.
Presentation Generation	Automated generation of coordinated speech, natural language, gesture, animation, non-speech audio, generation, possibly delivered via interactive, animated life-like agents (includes challenges of media selection, allocation, coordination and realization)	Mixed media (e.g., text, graphics, video, speech and non-speech audio) and mode (e.g., linguistic, visual, auditory) displays tailored to the user and context. Agents engaging/motivating to younger and/or less experienced users.
Awareness and Collaboration	Topic detection and tracking, place-based asychronous and sychronous collaboration environments.	Enhance awareness of new knowledge as well as other user's interests and expertise and the ability of experts to exchange/integrate knowledge.

4.4 Question Answering

Drawing upon techniques from query processing, retrieval and presentation, an important new class of systems will move us from our current form of search on the web (type in keywords to retrieve documents) to a more direct form of asking questions which are then directly responded to with an extracted answer. Challenges will include source selection, source segmentation, extraction, and semantic integration across heterogeneous sources of unstructured, structured, and semi-structured data.

4.5 Translation

Last year, for the first time, English constituted less than half the material on the web. Some predict that Chinese will be the primary language of the web by 2007. Given that information on the web will increasingly appear in foreign languages and not all users will be fluent in those languages, there will be a need to gist or skim content for relevance assessment and/or provide high quality translation for deeper understanding. New innovative applications include the translation of multilingual conversations (e.g., multilingual chat).

4.6 Dialogue Management

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 the ability to tailor flow and control of interactions and facilitate interactions including error detection and correction tailored to individual physical, perceptual and cognitive differences. Motivational and engaging life-like agents offer promising opportunities for innovation.

4.7 Agent/User Modeling

Computers can construct models of user beliefs, goals and plans as well as models of users' individual and collective skills by processing materials such as documents or user interactions/conversations. While raising important privacy issues, modeling users or groups of users unobtrusively from public materials or conversations can enable a range of important knowledge management capabilities. For example, this might include expertise databases which can be used to enhance organizational awareness and efficiency.

4.8 Summarization

Summarization aims to select content and condense it to present a compact form of the original source or sources. Summaries can be an extraction of or abstraction from original source material as well as informative, indicative or evaluative in purpose. Some challenges include multimedia, multilingual and cross document summarization. Addressing scaleability to large collections and user-tailored or purpose-tailored summaries are active areas of research. Summarization will enable knowledge workers access to larger amounts of material with less required reading time.

4.9 Presentation

Effective presentations require the appropriate selection of content, allocation to media, and fine grained coordination and realization in time and space. Discovery and presentation of knowledge may require mixed media (e.g., text, graphics, video, speech and non-speech audio) and mixed mode (e.g., linguistic, visual, auditory) displays tailored to the user and context. This might include tailoring content and form to the specific physical, perceptual, or cognitive characteristics of the user. It 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 breatdh, or simply increase user satisfaction. A grand challenge is the automated generation of coordinated speech, natural language, gesture, animation, non-speech audio, generation, possibly delivered via interactive, animated life-like agents. Preliminary experiments suggest that, independent of task performance, agents may simply be more engaging/motivating to younger and/or less experienced users (André 2000).

4.10 Awareness and Collaboration

Our gobal web provides unprecedented opportunity for world wide collaboration, both asynchronously and sychronously. Users will need enhanced awareness of both emerging knowledge and of one another's expertise. One such aid is the detection and tracking of topics of interest to facilitate discovery and connection among communities of interest. Another is the creation of expertise profiles based on publically available information (e.g., publications, interviews, public conversations).

5 Summary and Conclusion

Human language technology promises to deliver great value to the challenge of knowledge management. As discussed, many fundamental scientific and technical challenges will need to be addressed to ensure that value accrues. These include:

Heterogeneity – Dealing with the diverse nature of human language artifacts, both in form and in (semantic) content.

Scalability – Addressing the size of corporate collections and global content.

Portability – Creating adaptive methods (e.g., corpus-based machine learning approaches) that enable rapid retargetting 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 and/or ensuring the identity/integrity of source materials.

Privacy – Addressing the legal and social issues of maintaining privacy and user control of a user's model extracted from public user materials or interactions.

Overcoming these HLT challenges will be essential to the advancement of knowledge management.

6 References

André, E., and Rist, T. 2000. Presenting Through Performing: On the Use of Multiple Lifelike Characters in Knowledge-Based Presentation Systems, in: Proc. of the Second International Conference on Intelligent User Interfaces (IUI 2000), pp. 1-8, 2000. Best Paper Award.

Becerra-Fernandez, I. 2001. Searching for Experts with Expertise-Locator Knowledge Management Systems. In ACL '01 Workshop on Human Language Technology and Knowledge Mangement. Toulouse, France.

Maybury, M., D'Amore, R., House, D. Nov-Dec, 2000. Automating the Finding of Experts. *International Journal of Research Technology Management*. 43(6): 12-15.

Maybury, M. (ed.), February 2000. News On Demand. Special Section in *Communications of the ACM*. 43(2): 33-34, 35-79. (www.acm.org/cacm/0200/0200toc.html)

Maybury, M., D'Amore, R. and House, D. forthcoming. Awareness of Organizational Expertise. *Journal of Human Computer Interaction: Special issue on Awareness.*

Merlino, A., Maybury, M., and Morey, D. 1997. Broadcast News Navigation using Story Segments, ACM International Multimedia Conference, Seattle, WA, November 8-14, 381-391.

Morey, D.; Maybury, M. and Thuraisingham, B. editors, 2000. *Advances in Knowledge Management: Classic and Contemporary Works*. Cambridge: MIT Press.

Bontcheva, K., Brewster, C., Ciravegna, F., Cunningham, H. Guthrie, L., Gaizauskas, R. Wilks, Y. 2001. Using HLT for Acquiring, Retrieving and Publishing Knowledge in AKT. In ACL '01 Workshop on Human Language Technology and Knowledge Mangement. Toulouse, France.

7 Acknowledgements

The ideas described in this paper are inspired by the efforts of several human language and knowledge management researchers and groups at The MITRE Corporation including but not limited to 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 have benefit from ideas from Elisabeth André, Thomas Rist and Christoph Thomas and the anonymous reviewers.