Meeting of the MINDS: Future Directions for Human Language Technology

Executive Summary

Donna Harman, Scientist Emeritus, National Institutes of Standards and Technology

Introduction

Motivation:

The Human Language Technology (HLT) area has traditionally included the research disciplines that deal with the processing and understanding of human language. This includes research in machine translation (MT), with a concentration on accurate and fluent translations between languages, and research in information retrieval (IR), where the goal is finding information or answers in text (or other media). A major area of HLT is the natural language processing (NLP) community, who deal with understanding or interpreting text, such as extracting proper names, identifying relationships between terms, or locating co-references. Then there are two research areas that deal mainly with conversion of various inputs to text, such as research in speech, where the input is any type of naturally-occurring speech and the output is usually that speech transcribed to written text (for further processing), or optical character recognition (OCR) that scans printed documents to create text in electronic form.

HLT has had many funded projects over the years, particularly via DARPA and ARDA, plus a number of smaller NSF projects. These projects have usually been focused around some particular application (because of a need to "sell" the projects to upper management), or around some small component part of HLT. But whereas research groups are clearly cumulatively building on their past work, the various funded projects as a group cannot be viewed as a coherent whole. This is for several reasons, including the following:

1) The HLT research community is much fractured in terms of history, educational background, current research, etc. Whereas the MT and NLP communities have a common background and education in computational linguistics (within the artificial intelligence community), the IR community originated in a combination of library science and computer science. The speech community has a background in engineering, in particular acoustical engineering, and OCR also comes from engineering. These diverse backgrounds make for difficult communication between HLT areas, and the areas have tended to develop in isolation. Given the pressures of addressing immediate concerns and issues, there has been a lack of the broader perspective, with no thorough, well-articulated longer-term vision for either HLT or even its more focused subfields (areas).

2) Some effort has been made within many of the funded projects to "join" two components together, e.g., speech and NLP, usually by organizing workshops from two subfields. These have had limited success, and whatever sharing and exchanging of components that has occurred has usually been as a "handoff" rather than as a working together.
3) This fragmentation of vision across the HLT community has led to two major problems. The first is a minimal ability to integrate ideas or components across the subfields, leading to duplication of effort and a restricted understanding of better ways of addressing specific technical issues. It has also led to a disjointed set of corpora and lexical resources. All of the HLT subfields rely on training data to optimize their statistical algorithms, and these expensive investments are often produced only for one subfield, or even worse, for one application in one subfield. There are some very notable exceptions to this, such as WordNet, but even there it would have been useful to have input from the broader HLT community.
4) There is a dichotomy/tension between research that furthers HLT core knowledge versus work that produces something that is obviously needed for a given application. Funders of HLT research must be in for the long haul: the goal of true understanding of language, and therefore more flexible and accurate language processing systems, seems distant today. However there needs to be both a clear direction as to that long haul, and a good mixture of short and long term goals that can be evaluated to show progress.

Workshop organization and reports

The goal of the MINDS (MT, IR, NLP, Data resources, Speech) workshops was to convene a small group of well-respected and broadly-experienced researchers from six areas of HLT to discuss future directions in these areas. These experts were charged with carefully surveying and identifying critical issues for which effective solutions would have the broadest impact in advancing their fields and significant applications dependent on them. In these workshops there was a clear distinction made between identifying future directions versus looking for projects that were fundable. In this way the workshop differed from many past efforts in that the output was not a “white paper” aimed at funding sources.

There were two workshops, both sponsored by Heather McCallum-Bayliss of the Disruptive Technology Office of the Office of the Director of National Intelligence's Office of Science and Technology (ODNI/ADDNI/S&T/DTO). The first took place on November 13-14, 2006 in Chantilly, Virginia. There were 24 participants (see Appendix A), from five areas; both days of the workshop were similar in that most of the work was done in breakouts into the five separate subfields, with several joint sessions held to report progress and discuss issues.

The first workshop was focused by answering the following two questions:

1) Make a list of the 5-10 research discoveries that have led to a major paradigm change in your field. These could be things done within the field, such as some specific parser, or could be something outside, like the development of the Web.

They could be a specific discovery or the strong beginning of some trend, such as the heavier use of statistical methods in NLP

2) Using this list as a guide, create a list of 5-10 research areas that would result in equally important paradigm shifts. If you could propose a 3-5 year program in these areas, how would you design it?

After the first workshop, the groups were asked to start work on a report from each area based on the answers to the above questions. The work was generally split up across the various participants, with the discussion leader playing a major role in getting the reports organized.

The second workshop, held February 25-26, 2007 in Marina del Ray, California, had two goals. The first was to get input from new participants in each area in order to ensure that the final reports had broad coverage. Results from the first workshop were presented and refined, along with incorporating additional thoughts from the incoming participants, as well as a new data-resources group. The morning of the second day was spent in several cross-area discussions and getting input from all groups to the data resources group.

The rest of this executive summary is broken into seven sections. The first section presents answers to question 1, but combined across all of the various HLT areas. It was surprising how similar the issues were and how these issues tended to influence research development in similar ways. The next five sections (one for each area except OCR) cover the answers to question 2. Note that these sections are for the most part literal extractions of the original reports from each area and readers are encouraged to see those reports (on http://www.itl.nist.gov/iaui/894.02/minds.html) for more details. The final section discusses some results from the cross-area discussions; unlike the previous sections, these ideas are only samples of things that could be done rather than an effort to pick research areas leading to paradigm shifts.

History and reasons for past paradigm shifts

A key component of all of the answers to reasons for past paradigm shifts has been the vast improvements in computer hardware (doubling computational power almost every year), with the accompanying availability of very cheap storage. Note that some of the ideas behind current techniques in these HLT areas are old, but there was not enough speed or memory or storage to implement these ideas. For example, the basic algorithms behind the information retrieval systems were started in the 1960’s, but experiments then took days, even working with only 200 document abstracts!! For other areas, there was simply not enough compute power to develop any of the statistical methods that are common today.

The second key component was the public availability, starting in the 1990’s, of online text. This took two important forms. First, simply massive amounts of raw text set in mostly common formats (such as SGML) that allowed for easy processing.

However there soon were corpora (most of which were sponsored by DARPA) specifically built for statistical processing in all the HLT areas:

1) The T146 corpora corpus of isolated spoken words (1980), followed by the DARPA TIMIT corpora of 6300 sentences, for speech recognition,
2) The TIPSTER/TREC test collections, each of 2 gigabytes of text with test questions and relevant document lists (1992) for IR (also used by others as text),
3) The Canadian Hansards collection of parallel French and English text for MT,
4) The Penn Treebank (1993) and WordNet (1995) lexical resources for the NLP.

Along with these valuable corpora, the U.S. government, in particular DARPA, sponsored programs in the 1990s that were highly influential to the various HLT communities. For some of the communities, such as IR and speech, this influence came via common evaluations in retrieval (TREC) and speech (various speech recognition programs), where research groups used common corpora and common metrics to compare results. The NLP community had the MUC evaluation, which not only allowed some of the first statistical work to be evaluated but strongly encouraged groups to move to statistical methods by working in broad domains that could not be tackled using the older rule-based methods. The MT community also had a DARPA evaluation, but particularly important to this community was the development of the BLEU metric which allowed daily training of the MT systems for the first time (the other HLT areas had long had this ability with speech corpora, TREC test collections, and MUC annotated data).

A third key component to paradigm shifts in the HLT communities was the availability of both shared lexicons (such as WordNet) and shared components. The HTK system from the University of Cambridge was (and still is) heavily used by the speech community. The IR community has had several systems available from before the 1990s, such as the SMART system from Cornell and the INQUERY system from the University of Massachusetts; the new Lemur system is now widely distributed, along with the Lucene system. There have been many publicly-available parsers and taggers for the NLP area and these have been critical for the development of the complex systems in this area.

All three of these components played major roles in the paradigm shifts in the past, resulting in the heavy use of language and acoustic modeling that we see in all of the HLT areas today. These types of components will clearly continue to lead to further paradigm shifts, but it is important to critically examine the current environments and current state of the art in HLT in order to identify additional important influences.

Research suggestions to create paradigm shifts

Speech *(extracted from Historical Development and Future Directions in Speech Recognition and Understanding Recognition and Understanding by Janet M. Baker (lead), Li Deng, Sanjeev Khudanpur, Chin-Hui Lee, James Glass, and Nelson Morgan)*

1. Everyday Audio

“Everyday Audio” is a term that represents a large range of speech, speaker, channel, and environmental conditions which people typically encounter and routinely adapt to in responding and recognizing speech signals. Currently automatic recognition systems are challenged, and can deliver significantly degraded performance, when they encounter audio signals that differ sometimes even slightly from the limited conditions under which they were originally developed and “trained.” This 3-5 year research area would focus on creating and developing systems that would be much more robust against variability and shifts in acoustic environments, reverberation, external noise sources, communication channels (e.g. far-field microphones, cellular phones, etc.), speaker characteristics (e.g. speaking style, non-native accents, emotional state, etc.), and language characteristics (e.g. formal/informal styles, dialects, vocabulary, topic domain, etc.).

2. Rapid Portability to Emerging Languages

Today’s state-of-the-art systems deliver top performances by building complex acoustic and language models using a large collection of domain-specific speech and text examples. This set of language resources is often not readily available for rarely-seen languages that could be emerging in the horizon. To prepare for rapid development of such spoken language systems a new paradigm is needed to study speech and acoustic units that are more language-universal than language-specific phones. Three specific research issues need to be addressed, namely: (1) cross-language acoustic modeling of speech and acoustic units for a new target language; (2) cross-lingual lexical modeling of word pronunciations for new language; and (3) cross-lingual language modeling. By exploring correlation between these emerging languages and well-studied languages, cross-language features, such as language clustering and universal acoustic modeling, can be utilized to facilitate rapid adaptation of acoustic and language models.

3. Self-adaptive Language Capabilities

State of the art systems for speech transcription, speaker verification and language identification are all based on statistical models estimated from labeled training data, such as transcribed speech, and from human-supplied knowledge, such as pronunciation dictionaries. Such built-in knowledge often becomes obsolete fairly quickly after a system is deployed in a real world application, and significant and recurring human intervention in the form of retraining is needed to sustain the utility of the system

Like its human counterpart, the system will engage in automatic pattern discovery, active learning and adaptation. Research in this area must address both the learning of new models, as well as the integration of such models into pre-existing knowledge sources. Thus, an important aspect of learning is being able to discern when something has been learned, and how to apply the result.

4. Detection of Rare, Key Events

Current speech recognition systems have difficulty in handling unexpected – and thus often the most information rich – lexical items. This is especially problematic in speech that contains interjections of foreign or out-of-vocabulary words, and in languages for which there is relatively little data with which to build the system’s vocabulary and pronunciation lexicon. Such spoken events are key to tasks such as *spoken term detection* and *information extraction* from speech: accurate detection is therefore of vital importance. A key component of this research will be to develop novel confidence measures and accurate models of uncertainty based on the discrepancy between sensory evidence and *a priori* beliefs. A natural sequel to detection of such events is to transcribe them phonetically when the system is confident that its word hypothesis is unreliable, and devise error-correction schemes.

5. Cognition-derived Speech and Language Systems

A key cognitive characteristic of humans is their ability to learn and adapt to new patterns and stimuli. Although this behavior is very important and relatively well-understood in humans, very little of this knowledge has found its way into automatic speech and language systems. Since it is not possible to predict and collect separate data for any and all types of speech, domains, etc., it is important to enable automatic systems to learn and generalize even from single instances (“episodic learning”) or limited samples of data, so that new or changed signals (e.g. accented speech, noise adaptation, etc.) can be correctly understood. It has been well demonstrated that adaptation in automatic speech systems is very beneficial. . New tools now available for instantaneous brain imaging during speech and language processing may enable us to understand and integrate that knowledge into our automated speech and language systems.

6. Spoken Language Comprehension (Mimicking average language skills at 1st to 3rd grade)

Today’s state-of-the-art systems are designed to transcribe spoken utterances. To achieve a broad level of speech understanding capabilities, it is critical that the community explore building language comprehension systems that can be improved by gradual accumulation of knowledge and language skills. An interesting approach is to compare a system with a child at 1st to 3rd grade for listening comprehension skill.

Data Resources *(extracted from Historical Development and Future Directions in Data Resource Development by Martha Palmer (lead), Stephanie Strassel, Randee Tangi Christiane Fellbaum, and Eduard Hovy*

1. Science of annotation. The success of applying machine learning techniques to annotated training data has had a major impact on the field. This work is limited, however, by the availability of suitably annotated training material. For today's NLP systems, the annotation defines the task, and increasingly rich annotations are the key to more sophisticated systems. Clearly annotation work needs to become much more widely distributed to cope with this need. What precisely should be annotated, however, remains a matter of discussion. Every annotation effort requires that someone design the annotation scheme—its principal classes, their interrelationships, the conditions or circumstances under which the annotations may be added—and without careful thought there is always a danger that the annotation reflects some hasty decisions not based on solid principles. The field needs a concerted effort on creating an explicit description of a step by step process by which useful annotations can be achieved.

2. Robust, extensible annotation infrastructure. Along with a better understanding of a methodology for annotation there should be a set of public domain tools and interfaces that can support, and to a certain degree enforce, "best practice" annotation guidelines. A shift in focus to developing reusable, library-based code with task-specific modules and customizations would greatly improve the overall quality and cohesiveness of annotation systems to the benefit of the entire MINDS community of users. Development of such an industrial strength toolkit will only be possible with **sustained support** for a robust linguistic annotation infrastructure that can marry the "annotation science" developed in (1) with well understood principles of software engineering.

3. Closer integration of emergent technology. There is considerable evidence that the productivity of manual annotation can be speeded up by pre-processing the data with sufficiently accurate automatic taggers. However, current annotation practices frequently fail to take advantage of this approach, possibly because of the difficulty of integrating these systems into new annotation tasks. Even more benefit could be derived from using sophisticated machine learning techniques to aid in the selection of instances to be tagged, in order to maximize their utility and minimize the total annotation effort. Closer ties between annotators and the ASR, NLP, MT, IR and Machine Learning communities are needed for joint efforts to develop techniques to aid data selection and quick access to modular automatic taggers for preprocessing of data. Research is needed to explore issues such as **when** in the annotation pipeline technology can be folded in with the maximal benefit, and **what** levels of accuracy are necessary for the automatic taggers to provide a benefit. The goal should be, in addition to faster and more useful training data, the identification and selection of hard/rare/special data for annotation that supports better use of limited human annotation resources – bigger "bang for the buck." This will stimulate research in machine learning techniques as well as maximizing the impact of limited amounts of annotation, allowing humans to focus on the parts that really require human judgment.

4. Richer annotations. As described above, progress in natural language processing is being led by the definition of increasingly rich levels of representation. Given sufficient, good quality training data, it is likely automatic taggers can be built to replicate whatever representations they were trained on, to a degree relatively close to human performance. This makes the search for increasingly rich levels of representation that can be successfully annotated the highest priority in the field.

5. Language resource kits. There has long been recognition of the need to have basic language processing resources available for a broad spectrum of languages. When a language unexpectedly becomes of vital strategic importance, quickly having access to resources for either the language in question or for a closely related language could be crucial. A complete resource kit would contain:

- **text** - at least 100K words of parallel text (news domain), tagged for nominal entities & co-reference, basic syntactic annotation, basic predicate argument structure, topics + relevance judgments for news articles; part-of-speech taggers, morphological analyzers,

- **speech -** a pronouncing dictionary, a minimum of 100 hours of audio, a minimum of transcription of 10-25 hours, 50% Broadcast News, 50% Broadcast Conversation, possibly interviews.

Ideally these language kits should be made available for 100-200 of the less commonly taught languages.

6. Broad coverage empirically grounded lexical resources. There is widespread agreement that certain new lexical resources are extremely desirable and could be of great benefit. Additionally each application area has its own specially tuned lexicon with information customized to that domain that the other areas do not need and have no interest in providing. Statistical MT systems could be extracting extremely useful bilingual lexicons from aligned corpora, but without phonological information they will be of no use to ASR systems. At the moment the various lexicons are so diverse that different systems cannot ensure that they are referring to the same items. A public domain resource that lists all of the relevant types of information for each lexical unit, thus enabling the ASR, NLP, MT and IR systems to recognize that they are all dealing with the same lexical items is essential.

Natural Language Processing *(extracted from Natural Language Processing by Liz Liddy (lead), Eduard Hovy, Jimmy Lin, John Prager, Dragomir Radev, Lucy Vanderwende, and Ralph Weischedel)*

1. Machine Reading. One of the grandest challenges for Natural Language Processing is for a machine to be able to read text and learn, so that the machine can improve its performance on one or more tasks, e.g., read a user manual and be able to answer complex help questions by a user. Attacking this challenge will focus research on four areas:

a. natural language understanding in an open-ended domain that is expanding through reading the text,
b. going from strings of words to a logical form (richer semantic representation will be required),
c. utilizing multiple levels of annotation, e.g., employ propositional constraints at the same time as employing syntactic constraints or employ word sense disambiguation data in parallel with co-reference annotation; the search space may be smaller and yield more accurate results by applying evidence from more than one level in parallel,
d. designing knowledge representation and reasoning systems with inference methods that are more appropriate to human communication in texts.

2. Socially-Aware Language Understanding: For Natural Language Processing to be able to contribute to the full range of dynamic situations in which language is used, it needs to recognize, interpret, and respond appropriately in all the “contexts” in which language is encountered, not just formal, well-written genres. What is required is a shift in Natural Language Processing to real *in situ* understanding, where the system's language understanding capabilities are human-like in that they include the recognition of communicative goals, of how a conversation proceeds and varies in interesting ways depending on the conversants and the situation, and of how time and place affect interpretation.

3. Annotation Science: Significant improvements and breakthroughs in HLT can be achieved through the creation of large corpora annotated with representations at various levels, including syntax, shallow (word sense) semantics, discourse structure, co-reference, sentiment, etc. Therefore, a new field to be called 'Annotation Science' is proposed. This field would require solid methodology on at least the following seven issues: (1) selecting the phenomena that would most advance the state of the art; (2) designing the appropriate representations as guided by solid theoretical and practical input from the relevant fields of study; (3) selecting the most appropriate corpora and making sure they are representative and balanced; (4) creating simple, effective, and unbiasing annotation interfaces and procedures; (5) selecting and training annotators adequately but not too much; (6) evaluating the results using appropriate and informative measures; and (7) unifying and integrating the results, maintaining them as additional overlays are created, and distributing them with due regard to licensing and other concerns. Regarding scale, complementary approaches need to be explored, including: (1) using non-specialists in creative ways (e.g., social tagging on the web, by turning the drudgery of annotation into fun games, as in ESP); (2) acquiring labeled training data semi-automatically (e.g., using seed examples; applying active learning as a methodology); (3) finding and making use of "found data" (e.g., FAQs can provide Q&A pairs; contemporaneous news stories can provide paraphrases); (4) leveraging Web 2.0 as a vehicle for creative data acquisition efforts.

Information Retrieval *(extracted from Meeting of the MINDS:An Information Retrieval Research Agenda by Jamie Callan (lead), James Allan, Charles L A Clarke, Susan Dumais, David Evans, Mark Sanderson, ChengXiang Zhai)*

1. Heterogeneous Data: While well-edited text may remain the central data type, many people use a richer variety of formats and media, including blogs, instant messaging, text messaging, email, speech, video and images. The volume and complexity of the data generated by these sources precludes any possibility of manual cleaning or organization. While an IR system might filter out material that is outright harmful or adversarial, such as spam or viruses, the remaining material must be retained and made available for searching and browsing. The challenge increases when information arises as a mixture of data types. To cope with this mixture of data, IR systems must seamlessly integrate and correlate information across a variety of media, sources, and formats. The relationships between diverse elements must be apparent to the IR system, and must be exploited to improve information access. Each source and format cannot have its own interactive search interface

2. Heterogeneous Context: As information technologies continue to be used by a more and more diverse user population (students, researchers, analysts, medical professionals) for a wider and wider range of tasks (finding, learning, monitoring, communicating, planning) today’s search technologies and interfaces will need to be extended and improved. While there are many different factors that characterize searchers and tasks, three important classes of contexts seem widely applicable. First, we must do a better job of **understanding the user** who is asking the question and the previous knowledge and skills that they bring to bear on the problem. . Second, we need to better understand and represent the underlying **information domains** (see also heterogeneous data above). Finally, the **larger task** the user is trying to accomplish shapes both the kinds of information that is needed and how it should be presented. There are few tools to help people organize and digest information. Search is not the end goal. It is a tool that can help people accomplish other tasks – certainly a very important part of the process, but only a part. The more we understand the context of the search (the user, the domains of interest and the large tasks), the better we can deliver the right information to the right people using the right means.

3. Availability of Usage Data: In spite of the widespread use of search and text analysis tools, there is surprisingly little public knowledge about how most people use information retrieval tools “everyday”, in part because much of the data that people work with is personal, private, or confidential. Web search engines, large digital libraries, and e-Commerce sites have well-developed methodologies for tracking users and their interactions with information resources, however they do not share such information easily due to competitive and privacy issues. Although there is considerable value to having such basic knowledge in the public domain, it will remain in the private sector until IR has “off-the-shelf” privacy preserving research methodologies. One approach is to develop methods that allow usage data to be shared; this may be a difficult goal to achieve. An alternate approach is to develop standard tools and methodologies for capturing usage data, to make it easier to do this kind of research, and to make it more likely that experiments by different people are roughly comparable even if conducted on different users.

4. Evaluation: IR has been a leader in Computer Science in understanding the importance of evaluation and benchmarking. However the methodologies conceived in the early years of IR and used in the campaigns of today are starting to show their age and new research is required to understand how to overcome the emerging twin challenges of scale and diversity. Potential flaws are starting to emerge as test collections grow beyond tens of millions of documents; solutions are being investigated, but the long term stability of the test collections formed with the new approaches is still unclear. It is increasingly clear that evolution of retrieval is not towards a monolithic solution, but instead to a wide range of solutions tailored for different classes of information and different groups of users or organizations. Each tailored system requires a different mixture of component technologies combined in distinct ways and each solution requires evaluation.

5. IR in Service of HLT Applications: During the last decade *software applications* have emerged as a new class of IR system users. Question answering and information distillation systems are examples of this class of applications. The typical architecture involves a text search engine to efficiently gather “raw” information from a text database, and more sophisticated or specialized processing on the returned documents. In much the same way that e-Commerce systems are built on relational database systems, these applications are built on search engines. Recognizing and providing strong support for search by software applications is important because much recent human language technologies research is now data-driven. NLP, MT, and speech researchers need routine access to large corpora. Currently they must build specialized "one off" solutions to obtain such access; what they need is text search engines that support their information needs.

Machine Translation *(extracted from MINDS Workshops: Machine Translation Working Group Final Report by Alon Lavie (lead), David Yarowsky, Kevin Knight, Chris Callison-Burch, Nizar Habash, Teruko Mitamura)*

The field of Machine Translation (MT) has experienced a major paradigm shift over the course of the last two decades, from labor-intensive manually crafted rule-based systems to a general paradigm that has *computational search* as its core. The most prominent example of this new paradigm is Phrase-based Statistical MT , but this same basic paradigm also underlies most if not all other modern MT approaches, including various syntax-based statistical approaches to MT, Example-based MT approaches, modern transfer-based approaches, and also approaches that combine the output from multiple MT systems (Multi-engine MT). While search-based MT has become the dominant paradigm over the past decade, and research advances have been both significant and impressive, the ultimate goal of fully-automatic broad-coverage high-quality MT for many language pairs remains beyond current state-of-the-art. We believe that current state-of-the-art search-based MT approaches can be characterized by the following main common fundamental weaknesses: (1) **Weak Models:** The models used by current MT approaches are not strong enough to consistently generate correct translations. Consequently, the hypothesis-spaces that are generated by these MT approaches often do not contain correct, or even good possible translations of the input. (2) **Weak Discrimination During Search:** The knowledge resources utilized in today's MT systems are insufficient for effectively discriminating between good translations and bad translations. Consequently, the decoders used in these MT systems are not very effective in identifying and selecting good translations even when these translations are present in the search space. Given the current state-of-the-art, it is our belief that the three most important "grand challenges" for MT research are: **(1) High-Quality MT for many more language pairs; (2) Substantially improved translation quality robustness across domains, genres and language styles;** and **(3) Achieving human-level translation quality and fluency.** The following are five concrete research themes that have the greatest potential, in our opinion, to push the field forward:

1. Effective Sub-sentential MT Models that Generalize: create research scenarios that explicitly encourage MT researchers to focus on the problem of developing MT models at the sub-sentential level that capture correspondences at increasing levels of syntax and semantics. A first critical enabling step in this direction would be to create sentenceparallel corpora annotated with accurate syntactic/semantic structures on one side or both sides. Research scenarios that involve translating a broad mix of genres, domains and text styles will also encourage development of more general models. Additional research scenarios that can encourage such development include an explicit focus on specific sub-problems of MT, such as translation of sentences that exhibit known types of language divergences, which cannot adequately be handled using today's MT approaches.

2. Learning More from Less Data: Research scenarios that require learning "as much as possible" from limited amounts of data that are annotated with higher-levels of representation will create needed incentives for researchers to develop MT models that are far more suitable for realistic scenarios, where only limited amounts of data will be available. The challenge should be how to develop the best performing MT system given such limited amount of training data, and additional NLP tools and resources.

3. MT from English into Other Languages: Research scenarios that involve translation from English into other languages are of extreme importance in that they create challenges that will force MT researchers to deal with language phenomena that have not received adequate attention to date. Adequate treatment of complex morphology is one clear research challenge that translation into languages other than English will require.

4. Multi-Engine Machine Translation: Different approaches using different types of models should continue to be developed, and research scenarios that explicitly encourage such diversity in approach is in the best interest of advancing the MT field. Furthermore, there has recently been a surge in interest in approaches that can synthetically combine different MT engines operating on a common input into a "consensus" translation which surpasses all the individual MT engines in its quality. Teams should be evaluated not only based on the performance of their MT system in isolation, but also for the contribution of their MT system within multi-engine combinations, to encourage research into diverse MT approaches.

5. MT Evaluation: The models used by MT systems today and in the future contain a variety of parameters that need to be tuned for optimal performance. Automatic MT evaluation metrics such as BLEU provide the "target function" for optimizing these parameters for best translation performance. The automatic metrics available today, however, are very crude, and do not correlate well with human judgments of translation quality. As MT systems improve and achieve high levels of translation quality, it becomes ever more important to have evaluation metrics that are sensitive to small differences between translations at the sentence-level, so that minor improvements can still be detected, concrete translation errors can be isolated and identified, and system parameters can be optimized to truly achieve the best translation performance.

Examples of Future Cross-Area Research

The last part of the second MINDS workshop was spent in breakouts with each set of two areas discussing possible cross-area research. A typical question used to start the discussion was to ask each area group to identify technical help that they needed from the other area. Additionally there were discussions of applications that clearly needed input from two areas in order to be successful. What follows are two examples that came out of these discussions. The first example, which is discussed in more detail in the information retrieval report (Appendix E), is that of using information retrieval techniques to help build the language models that are needed by all of the HLT subfields. The second example, taken largely from the Natural Language Processing report (Appendix D), proposes the challenge of generation of minutes from meetings.

IR in Service to HLT Applications *(extracted from Meeting of the MINDS:An Information Retrieval Research Agenda by Jamie Callan (lead), James Allan, Charles L A Clarke, Susan Dumais, David Evans, Mark Sanderson, ChengXiang Zhai), Appendix E*

In cross-area discussions among the MINDS groups, people from other research areas described different types of information needs. For example, a speech recognition system might form a language model of the last few minutes of speech (perhaps it is about basketball), pass it to a search engine as a "find similar documents" query, receive back a set of documents (about basketball), and use them to form a more accurate language model for predicting the speech that will be encountered next.[1] Machine translation and natural language processing researchers described similarly specialized needs, for example, support for recognizing viable translation hypotheses, constructing parallel/comparable corpora, and extending concept hierarchies. The common theme was that their research requires routine access to large text databases. Often there is a specific information need that requires quickly gathering a "small" amount of text – just what full-text search engines are designed to do. Recognizing and providing strong support for search by software applications is important because much recent human language technologies research is now data-driven. NLP, MT, and speech researchers need routine access to large corpora. Currently they must build specialized "one off" solutions to obtain such access; what they need is text search engines that support their information needs.

A Unifying Challenge: Generating Meeting Minutes *(extracted from Natural Language Processing by Liz Liddy (lead), Eduard Hovy, Jimmy Lin, John Prager, Dragomir Radev, Lucy Vanderwende, and Ralph Weischedel), Appendix D*

A unifying grand challenge for all of the Human Language Technologies, and especially for speech processing and for natural language processing is automatic generation of minutes for a meeting. The "minutes" should have three elements:

- A full, accurate transcript of everything stated, including who said what when,
- A searchable index into the audio record of the meeting,
- A set of minutes in the style / genre of that particular group. For example, an informal scientific meeting might include the topics, major important issues / factors raised, conclusions reached, and action items; a group following more traditional rules of order would need a record of motions, who made each, who seconded each, a summary of discussion of each motion, the vote and its outcome

[1] Thanks to Sanjeev Khudanpur for this example.

Attacking this challenge will focus research on six paradigm shifts: from speech recognition of news to everyday, multi-person speech, from transcription to situationally-structured content, from literal meaning to discourse structure and intent recognition, from summaries by extract to transcript synthesis, from one-shot statistically trained systems to self-adapting systems and finally from isolated, independent research in speech and in language understanding to collaboration across the communities to solve the grand challenge. Note also that by including languages other than English, the MT community could be involved, and the IR group would be challenged by the creation of searching abilities across all the rich metadata that would be generated in meeting room minutes.

Appendix A – list of participants

Donna Harman, NIST (organizer/moderator for both workshops)

Information Retrieval (IR)

Jamie Callan, Carnegie-Mellon U.	(discussion leader for both workshops)
James Allan, U. Massachusetts	(workshop #1)
David Evans, Clairvoyance Corp.	(workshop #1)
Mark Sanderson, U. Sheffield, UK	(workshop #1)
ChengXiang Zhai, U. of Illinois	(workshop #1)
Charlie Clarke, U. Waterloo	(workshop #2)
Sue Dumais, Microsoft	(workshop #2)

Natural Language Processing (NLP)

Liz Liddy, Syracuse U.	(discussion leader for both workshops)
Ralph Weischedel, BBN	(both workshops)
Jimmy Lin, U. Maryland	(workshop #1)
John Prager, IBM	(workshop #1)
Lucy Vanderwende, Microsoft	(workshop #1)
Ed Hovy, ISI	(workshop #2)
Drago Radev, U. Michigan	(workshop #2)

Speech

Janet Baker, Saras Institute, MIT	(discussion leader for both workshops)
Sanjeev Khudanpur, John Hopkins U.	(both workshops)
Li Deng, Microsoft	(workshop #1)
Jim Glass (CSAIL, MIT)	(workshop #1)
Nelson Morgan (ICSI)	(workshop #1)
Chin-Hui Lee , Georgia Tech	(workshop #2)

Machine Translation (MT)

Alon Lavie, Carnegie-Mellon U.	(discussion leader for both workshops)
David Yarowsky, John Hopkins U.	(both workshops)
Nizar Habash, Columbia U.	(workshop #1)
Teruko Mitamur, Carnegie-Mellon U.	(workshop #1)
Chris Callison-Burch, U. Edinburgh	(workshop #1)
Kevin Knight, ISI	(workshop #2)
Daniel Marcu, ISI	(workshop #2)

text from images (OCR, only at first workshop)

Henry Baird, Lehigh U.	(discussion leader)
David Lewis, consultant	(workshop #1)
Sargur Srihari, CUNY Buffalo	(workshop #1)
David Doermann, U. Maryland	(workshop #1)

Data resources (only at second workshop)

Martha Palmer, U. Colorado	(discussion leader)
Stephanie Strassel, LDC	(workshop #2)
Randee Tengi, Princeton	(workshop #2