Natural Language Processing

Liz Liddy (lead), Eduard Hovy, Jimmy Lin, John Prager,
Dragomir Radev, Lucy Vanderwende, Ralph Weischedel

This report is one of five reports that were based on the MINDS workshops, led by Donna Harman (NIST) and 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). To find the rest of the reports, and an executive overview, please see http://www.itl.nist.gov/iaui/894.02/minds.html.

Historic Paradigm Shifts or Shift-Enablers

A series of discoveries and developments over the past years has resulted in major shifts in the discipline of Natural Language Processing. Some have been more influential than others, but each is recognizable today as having had major impact, although this might not have been seen at the time. Some have been shift-enablers, rather than actual shifts in methods or approaches, but these have caused as major a change in how the discipline accomplishes its work.

Very early on in the emergence of the field of NLP, there were demonstrations that NLP could develop **operational systems with real levels of linguistic processing, in truly end-to-end (although toy) systems**. These included SHRDLU by Winograd (1971) and LUNAR by Woods (1970). Each could accomplish a specific task — manipulating blocks in the blocks world or answering questions about samples from the moon. They were able to accomplish their limited goals because they included all the levels of language processing in their interactions with humans, including morphological, lexical, syntactic, semantic, discourse and pragmatic. These demonstration systems inspired the new field, but it was to take many years before other systems were to include the more complex levels of processing in real world systems.

Given NLP's lineage, it was not surprising that many of its early theories and methods derived from the field of linguistics. A major shift came in the early 1990s with the move to **a reliance on empirical methodologies** vs. the introspective generalizations that characterized the Chomsky era which held sway in theoretical linguistics. The focus in NLP shifted from what might be possible to do in a language and still have it be grammatically acceptable to what is actually observed to occur in naturally occurring text — that is, performance data. As more and larger corpora became available, empirical methods and evaluation rather than introspection-based methods and evaluation became the norm.

The availability of larger, performance-oriented corpora supported the **use of statistical (machine learning) methods** to learn the transformations that in previous approaches were performed by hand-built rules, eventually providing the empirical proof that statistical processing could accomplish some language analysis tasks at a level comparable to human performance. At the center of this move lay the understanding that much or most of the work to be effected by language processing algorithms is too complex to be captured by rules constructed by human generalization, but rather require machine learning methods. The early statistical Part-Of-Speech tagging algorithms using Hidden Markov Models were shown to achieve performance comparable to humans. A state-of-the-art statistical parser was shown to perform more accurately than a broad-coverage rule-based parser on the test sections of the Penn TreeBank and also on unseen portions of the Brown Corpus (Ringger et al., 2004). Framing questions in the noisy channel model / information theory, with use of Probability Theory, Maximum Entropy, and Mutual Information, produced tangible advances in automatic capabilities.

An enabler of these shifts was the newly available, **extensive electronic resources**, first in the form of sizable corpora, such as the Brown corpus, through the ongoing provision of collections funded by DARPA research programs and collected and distributed by the Linguistic Data Consortium. Later came lexical resources such as WordNet, which provided lexical-semantic knowledge bases, which first enabled use of the semantic level of processing, and the Penn TreeBank, which provided gold standard syntactic resources that led to the development and testing of increasingly rich algorithmic analysis tools.

The increasing availability of realistically-sized resources, coupled with machine learning methods supported a **shift from a focus on closed domains** of the first 30 years of NLP research (from the 60s through the 80s) **to open domains** (e.g., newswire), much of this shift to open domains was brought about originally by DARPA funding (i.e., solving toy problems in narrowly defined domains was not sufficient). The ensuing availability of the broad-ranging textual resources of the web, further enabled this broadening of domains.

Concomitant with these moves towards use of more real world data came the realization that NLP researchers should evaluate their work on a larger scale, and with this came the introduction of **empirically-based, blind evaluations** across systems, as first exemplified in the MUC series of evaluations and conferences, followed by TREC and DUC. These efforts led to the development of metrics such as BLEU and ROUGE that are integral to today's NLP research itself, in part because they can be computed automatically and results fed back into the research.

In parallel with these advances in statistical capabilities, but moving at a slower pace, was the demonstration that **higher levels of human language analysis are amenable to NLP**. The lower levels (morphological, lexical, and syntactic) deal with smaller units of analysis and are considered to be more rule-oriented and therefore more amenable to statistical analysis, while the higher levels (with semantics as a middle level, and discourse and pragmatics as the higher levels) admit of more free choice and variability in usage. That is, these levels permit more variation, with more exceptions, and perhaps fewer regularities. For example, in NLP, Rhetorical Structure Theory, Mann & Thompson, (1988) began to deal with discourse level phenomena, and demonstrated that even these much larger units of analysis (e.g., treatises, instructional guides, etc) were amenable to computational analysis. In information extraction, increasingly complex phenomena, such as subjectivity and opinion are being identified automatically (Wiebe et al., 2003). The most recent machine translation results are demonstrating that syntax-based MT outperforms surface-level word and phrase replacement systems (Charniak et al, 2003; Quirk et al, 2005).

Together these individual developments have resulted in the realization that NLP, by the blending of statistical and symbolic methods, together with lexical resources such as WordNet, and syntactic and semantic resources such as Prop Bank, plus the availability of large scale corpora on which to test and evaluate approaches, is gaining ground on the goal of realistic comprehension and production of human-like language understanding.

Machine Reading

Vision

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. Today, instead of having a machine simply learn by reading, knowledge engineers must work painstakingly to manually encode human knowledge in a Knowledge Representation and Reasoning (KRR) system. Project HALO [1] demonstrated that knowledge engineers could encode the knowledge in an introductory college level chemistry text, and that KRR systems could then answer questions at the Advanced Placement college level for that text, but that manually encoding such knowledge would cost an estimated $10,000 per page[2]. Furthermore, there are relatively few knowledge engineers, while there are vast amounts of human knowledge available in text.

Paradigm shifts: Attacking this challenge will focus research on four paradigm shifts:

1. From limited domains to open-ended domains — From the mid 1970s through the early 1990s, the NLP community had focused on producing a logical form as an interpretation, but always in the context of a 'limited domain', a domain with a pre-specified list of semantic entities and relations among them, e.g., natural language querying against a relational database, making airline reservations, or extracting specific pieces of information (e.g., Persons, Organizations, and Locations) as in the Automatic Content Extraction evaluations. Machine Reading will focus effort on natural language understanding in an open-ended domain that is expanding through reading the text.
2. From strings of words to logical form – From the early 1990s through the present, NLP has focused on operations at the surface level (i.e., on uninterpreted strings of words) rather than on mapping the text into a deeper-level logical form. For example, in Text Summarization, algorithms select sentences/phrases from the original

[1] http://www.projecthalo.com/

[2] N. S. Friedland et al., "Project Halo: Towards a Digital Aristotle," AI Magazine, Volume 25, No. 4, pp. 29 - 47, Winter 2004 Edition. (http://www.projecthalo.com/content/docs/aim_final_submission.pdf)

document(s) and reassemble them in summaries, never capturing the interpretation in a KRR. In Machine Translation, rules are learned that map strings of words in the source language to strings of words in the target language without interpreting the source. To accomplish Machine Reading, richer semantic representation will be required

3. Integrated solutions to language challenges — Well-known challenges in NLP have been tackled in isolation: word sense disambiguation, semantic role labeling, and coreference resolution. Only recently have substantial corpora become available that unify annotation of word sense, semantic role, and coreference: today they are possible through applications such as OntoNotes (Hovy et al., 2005). Algorithms trained on such data can for the first time utilize 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 coreference annotation; the search space may be smaller and yield more accurate results by applying evidence from more than one level in parallel. Yet further levels of annotation are still required in order to accomplish human-like language processing which would require annotation of the pragmatic connotations of language in use for a particular purpose.
4. Toward practical inference — Traditional KRR makes many assumptions that seem invalid in human reading; for example, KRR assumes unambiguous terms and a fully consistent knowledge base. A KRR more appropriate to human communication in texts might prove more practical; fledgling examples of such a new approach have been reported in the Rich Textual Entailment (RTE) evaluations (Dagan et al, 2006).

Challenge	**State of the Art**	**Required**
Full semantic interpretation of NL (incl. ambiguity, negation, metonymy, vagueness, coreference, etc.)	Only small theoretical demos	Able to handle all phenomena in all domains
Robust, practical inference	Narrow and hard to control	Able reliably to integrate knowledge into models and derive new knowledge and answers
Hypothesis management and Machine Reading process control	Pilot systems only	Able to handle complexity of any domain
Adequacy of KRR languages: - shallow KRR - deep KRR (including axioms)	Wide-domain, but few aspects of semantics Only in very small domains	Wide-domain, able to handle most aspects Adequate for domain
Evaluation	Variants of question answering	Generality, not domain-specificity.

Table 1: Technical challenges implicit in attacking Machine Reading

Program Parameters

While the goal of Machine Reading is a system that can read any text in any domain, the same as humans can, an initial program must select:

- A subject area, e.g., biology or software user manuals
- A reading level, e.g., junior high or college level
- The amount and level of prior knowledge built into the knowledge base
- The application and tasks where reading would improve performance, e.g., answering help desk questions regarding software use
- An evaluation paradigm, e.g.,
    - Ask system to answer questions
        - Ask a set of questions before reading a text for the first time
        - System reads the (previously unseen) text
        - Ask the system the same set of questions
        - Measure the delta in question answering capability from before reading to after
    - Instruct system to ask questions regarding text just read
    - Instruct system to summarize what it has read

Impact

The impact will be far reaching. Imagine a personal assistant that reads a user manual and answers your questions, explains how to do something, etc., not just retrieves passages of text that might have an answer for you. Imagine an intelligent tutor that reads a textbook and interacts with and aids the student, based on what it has learned from its reading. Imagine how many truly expert systems there would be, if they could be created without the costs and elapsed time required today.

Socially-Aware Language Understanding

Vision

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. This requires additional levels of interpretation beyond standard semantics, and can be thought of as self-adapting personal language processing, which will incorporate all the sets of features which convey meaning based on linguistic and paralinguistic cues that humans use in their social-communicative interactions, whether in speech or in everyday written communications (e.g., email, text messaging, or instant messaging). Such cues include emphasis marking, and other non-lexical symbols, inflection and energy levels in voice, and stylistically mediated pragmatic effects such as formality and partiality at the lexical level. These phenomena tend to be less formalized and less obviously rule-governed than syntax (addressable by parsing) or semantic roles (addressable by automated role labeling). Even more than well-written genres, interaction-oriented communication is imbued with contextual information — both generic expectations (what general type of situation is this; e.g., discussing a house one is considering buying) and specific knowledge (who is sending me this IM right now; an offeror, a real estate agent, or a friend) as to how this input is to be understood, and perhaps responded to. This would include the ability to recognize and utilize all the appropriate communicative devices such as politeness, skepticism, or sarcasm, as well as to correctly determine the appropriate amount of substantive detail or level of explicitness needed in interacting with a particular individual, or recognize the subtler connotative meaning intended to be understood by the language producer.

Paradigm Shift

Much of NLP to-date has been based on accomplishing the lower levels of language processing, i.e., morphological, lexical, and syntactic, with some degree of semantic level understanding, but only minimal use of the higher levels of language understanding, namely discourse and pragmatic. Currently, when these higher levels of language understanding are incorporated in systems, it is typically only one or a few of that level's phenomena (e.g., just coreference resolution at the discourse level, or pro- and con-opinions at the pragmatic level) that are dealt with. Accomplishing this paradigm shift requires language understanding that incorporates the full range of phenomena at the higher levels of human language processing, and that integrates them effectively into fuller understanding.

The desired paradigm shift would require a system's understanding and production of language that goes beyond literal meaning, that is, from just denotative meaning to connotative meaning. For by staying at the denotative level, systems will not be able to accomplish the true human-level language understanding that is accomplished when two individuals interpret the statements of each other in light of what they have learned as to the thoughts, experience, memories, and knowledge of the other. Two human discussants, each having their own areas of experience and expertise, can still understand the other although their vocabularies and syntax and discourse structure might not match, due to their ability to jointly construct meaning, no matter the level of language at which it is conveyed. This can be prosodics, or a specific lexical choice of a word with a negative connotation vs. an alternative word choice with a neutral connotation.

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.

Challenge	**State of the Art**	**Required**
Understanding the intent behind a statement / question	Subset of intents in restricted domain dialogue systems	Intent recognition capability over underspecified statements
Recognizing emotive dimension, no matter how	Polarity recognition at the lexical level	Detecting all affective dimensions whatever

conveyed linguistically.		linguistic phenomena is used to convey them
Awareness of individual communication partners' knowledge, experience, and personal style	Limited to knowledge engineering approaches	A representation formalism complex enough to capture all pragmatic dimensions

Table 2: Technical challenges implicit in tackling Socially-Aware Language Understanding

Program Parameters

Evaluation calls for a true Turing Test comprised of multiple tasks with multiple partners on multiple topics in multiple situations with multiple communication goals on the part of both partners. Ultimately one could evaluate an NLP-based system's performance as one player in a multi-player social system game, e.g., The SIMS (a simulation of the day-to-day activities of one or more virtual people in a household) or the World of Warcraft, (a massive, multiplayer online role-playing game). Simplified versions of these games would need to be used for the initial testing, but as the systems improved, more complex and realistic versions could be tested. Choice of a gaming environment would ensure capturing the participation of the newest generation of researchers.

Given that today NLP is only now addressing semantics in larger-scale systems, it is infeasible to plan for such 'full' understanding in the near term. Yet limited forms of some of the phenomena are computationally feasible and can have measurable impact. For example, language generation systems that can tailor their output appropriately to the reader and the situation exist and have been evaluated (Hirst et al. 1997). A good example is the Quick!Help system that produces recipes for food handout recipients; where the recipes are tailored for language, cooking skill level, cooking motivation, etc., where it was shown that the recipients use them twice as frequently as compared to generic one-size-fits-all recipes.

To help explore Socially Aware Language Understanding, a funding program needs to identify a set of tasks and task performance goals, target groups of language users, and a set of situations, and then evaluate the effectiveness of meeting increasingly challenging goals under different approaches and styles of computationally mediated language usage.

One necessary requirement of a research program in this challenge area would be a broader inclusion of disciplines that can contribute to an understanding of the issues involved and new theoretical and practical understandings that each of these disciplines could bring to a solution. These are necessary in order to account for the specific effects of, and constraints on, language usage in any one social context in order to understand the nuances of real language in use in a wide range of language situations. Such disciplines would enable identification of key discourse patterns and pragmatic considerations of language users in various modes of social life. Relevant disciplines include Cognitive Linguistics, Sociolinguistics, Stylistics, Discourse Analysis, and Cognitive Psychology.

Impact

While the complete and ultimate achievement of this paradigm-shifting version of Natural Language Processing capability might best be envisioned as conversational agents who could exist in the real world, even incremental advances towards that goal would have substantial impact on many current HLT domains, such as realistic question-answering in Customer Relationship Management, or in truly capturing the fullest meaning in the Meeting Minutes Challenge to be presented later.

Annotation Science

Vision

Several recent papers (Banko & Brill, 2001; Keller & Lapata, 2004; Och, 2005) document the fact that as the amount of training data increases, so do performance scores. A second fact has been observed in almost every Human Language Technology (HLT) application: Performance levels plateau in many applications in which internal processing representations are no 'deeper' than the word level (such as syntax, focus, discourse, etc.). For example, IR results have been 'stuck' at the same levels of Recall and Precision for almost a decade, as has automated speech recognition quality on open-domain speaker-independent speech. Until very recently, neither of these applications have been addressed by methods that use representations other than the surface level (i.e., words). And only in the past years, has syntax-enabled machine translation begun to convincingly outperform surface-level word / phrase-substitution Machine Translation (Charniak et al, 2003; Quirk et al, 2005).

Combining these two facts argues that 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, coreference, sentiment, etc. These corpora can serve as training data for the (semi-) supervised training of new generations of HLT components. The LREC Community in Europe has recognized the need to focus on lexical resources, due to the necessities of dealing with multiple languages, and has provided strong evidence to include annotation across languages as well.[3]

Two principal challenges for creating large amounts of new training data arise:

1. How do we know which phenomena to focus on, develop representations for them, and ensure consistency of the representations?
2. How do we acquire (annotate or otherwise obtain / acquire) the massive amounts of training data necessary to sustain progress?

Paradigm Shift

What is the current state of the art, what needs to be different, and what can be done?

[3] http://www.lrec-conf.org/

Challenge	State of the Art	Required
Selecting phenomena and defining representations	Sometimes ad hoc, sometimes informed by other fields	'Annotation Science'
Acquisition of large scale labeled training data	Manual annotation, Treebanking	Improved methods for acquisition of orders of magnitude more labeled data
Robustness to noise in training data	Varies by domain, some cope better than others	Ability to self-adapt to different conditions
Algorithms for processing massive amounts of data	Giga / terabyte scale	Peta / exabyte scale and beyond

Table 3: Technical Challenges Implicit in Large-scale Annotated Corpora for Learning

Systematizing and better understanding the procedure of creating useful training data deserves much more attention than has so far been the case. Therefore, we propose a new field to be called 'Annotation Science'. 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. Though each of these questions is currently fairly poorly understood by the HLT community, researchers in other fields do address some of these pertinent questions individually (e.g., psychology experimenters for annotation measurements; human factors / HCI researchers for annotation interface design; corpus linguists for balanced corpus creation), and should be enjoined in the new field.

Centrally, annotators need to find the 'sweet spot' among inter-annotator agreement, depth of annotation, and productivity rate. And every annotation effort needs to prove the value of its annotations for HLT applications, which requires close connections with these applications in order to test the incremental value of annotated corpora.

Regarding scale, we would like to have *the effect* of all the training data one would ever want, but this is too hard and expensive to obtain. So 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. In addition, we require algorithms for manipulating massive data structures (e.g., sparse matrices that don't fit into memory), and centralized resource centers that provide services beyond the capabilities of individual research labs. Particularly needed from IR are sophisticated search tools for exploring annotation repositories as they are being built and populated; and then data mining them to investigate corpora characteristics.

Program Parameters

Given the foundational nature of representation, it is hard to design a single specific annotation scheme that will suit all needs. Rather, it makes sense to support the creation of various annotations on some standardized corpora that have a potential for high impact in various applications, and to do so on an ongoing basis as new gaps are brought to light both by developments in HLT applications as well as what is learned from data-mining the annotation repository itself. One can, however, identify a few specific annotation efforts to be pursued that have very general applicability, following the models of the Penn Treebank and WordNet. In addition, research on the development of annotation methodology in and of itself should be fostered and, as discussed in the section on Data Resources, the creation of a repository of annotated corpora, annotation tools, large-scale data processing, and storage algorithms as well.

Impact

While computational resources continue to become increasingly powerful and available, and storage becomes essentially free, the creation of multiple large corpora of training material, suitably structured and annotated, remains as hard and expensive as ever. Yet it is one of the most obvious ways to accelerate progress, because accomplishment of ever-more human-like NLP requires that what is annotated in texts become more sophisticated and incorporate the richer, more complex, and more implicit aspects of language.

Intersections Between NLP and other Areas of MINDS

Vision

Currently, research efforts in the different MINDS communities tend to address similar basic problems independently, and many of them use little or no NLP knowledge or techniques. Synergies are typically rare and researchers from one community (except for a small number of crossover researchers) tend not to follow the other groups' literature. Thus advances in one community are often not picked up in other communities until much later. For example, the development of robust syntactic parsers almost 10 years ago is only gradually beginning to lead to a change in Machine Translation systems.

NLP can not only benefit from the other areas but it can also actively contribute to them. Areas where NLP can use results from other communities include passage retrieval (including from XML documents) from the Information Retrieval (IR) community for providing input to an NLP system, In turn, NLP can contribute to improving IR by allowing better similarity matches (e.g., using syntax-based tree kernels or dependency kernels).

Paradigm Shift

The idea of using NLP representations and techniques to enhance the performance of applications such as IR, information extraction, and MT, is not new and has been tried often, often with little discernible effect. Some researchers draw the conclusion that processing deeper than the surface (word) level is not useful. Two arguments can be made against this claim.

First, one can construe the application task in such a way that deeper levels of processing are not relevant. IR is an example: constraining the input to a few words means that people cannot pose a full NL question, and that IR systems' task ends when they have delivered a set of documents. But most people want answers to specific questions, and no QA system has ever been built that does not include some NLP techniques to augment the IR system at its core, specifically, for parsing the question and candidate answer strings and performing some matching between them. When IR is reconstrued as a special case of QA, the claim that IR does not need NLP techniques no longer holds.

Second, the language research community has in the past focused on precisely those problems for which language processing at deeper levels (semantics, inference, discourse, pragmatics, etc.) can be avoided without too much loss in task performance. This bias may have been prudent at the time, but it does not follow that the other challenges are unimportant. Now, it should be explored if / how NLP can help MT by providing word sense disambiguation, or text generation capabilities as part of a statistical MT pipeline, or entailment-based inference, in question-answering.

Machine Reading is an example challenge that fully requires semantics and inference, and without fully utilizing these capabilities, we will never achieve systems that can educate themselves and become fully useful information partners to people. Similarly, Socially-Aware Language Understanding systems that can correctly interpret and handle interpersonal and other pragmatic cues are necessary for socially sensitive machine translation of conversations, for successful dialogue-based help systems, and for the production of suitably tailored instruction manuals and teaching materials. These important challenges all require NLP representations and processing considerably beyond the current state of the art.

Every HLT community needs to realize that they are working on very similar basic language-based problems and actively join forces with one another towards the solution of problems of common interest.

Program Parameters

A program that optimizes on all HLT areas must focus on the following aspects:

- Joint development of annotated resources to support multiple applications in parallel. For example, a corpus annotated for word sense or theme / rheme should not only consist of written text but also of manually and automatically transcribed speech.

- Joint evaluation. This can be done both intrinsically (e.g., in terms of language model perplexity) or extrinsically (focusing on user satisfaction).
- Joint development of code. A number of toolkits (e.g., SRI-LM for language modeling, the Collins and Charniak parsers, Lemur and Indri for document retrieval) have been actively used by members of multiple HLT communities.
- Development of evaluation pipelines that make it easy for researchers from one of the communities to test their ideas in another community's system without having to learn or build an entire pipeline. Examples include adding word sense disambiguation to spoken language processing or adding semantic role identification to an MT system. It should be easy for the "guest" community to evaluate the influence of their module on the overall "host" pipeline.

Some specific initiatives might include the creation of a Google-scale corpus for research in NLP and IR; the joint formulation of criteria for query representation; the creation of a mixed-source corpus for integrated understanding across languages and media (including backchannels); better identification of semantic similarity, entailment, or paraphrasing; and the automatic creation of minutes or executive summaries of meetings, including question-answering ('*Did the motion pass or fail?*'). The last of these potential unifying programs is provided as an example in the next section.

Impact

As the complexity and sophistication of language technology increases, the core research problems shared by many of the HLT fields become increasingly pressing. If a critical mass of researchers from the various sub-areas were to interact with each other at some length and become aware of each other's problems and their attempts at solution, some of these problems may be solved faster. Cross-pollination may also foster the development of totally new types of systems than any single community would have envisioned alone.

A Unifying Challenge: Generating Meeting Minutes

Vision

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, and
- 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.

Paradigm Shift

Attacking this challenge will focus research on six paradigm shifts (summarized in the table below):

1. From speech recognition of news to everyday, multi-person speech. Most funding in the US has focused on transcription of news, a domain where much of the speech is read from tele-prompters of carefully crafted prose. On the other hand, everyday speech is spontaneous, affected by mood swings (e.g., from calm to agitated or frustrated), and involves interruptions.
2. From transcription to situationally-structured content. While current research, such as the GALE Program, is applying speaker identification to identify who is saying what when, there is much more that is needed. A meeting exhibits an implicit structure which should be captured in a transcript, e.g, an agenda and movement through the agenda. A quality "transcript" should capture at least that pragmatic structure.
3. From literal meaning to discourse structure and intent recognition. Current technology focuses on literal interpretation, e.g., information extraction (ACE), relevance to a stated retrieval need (TREC), or responding to a query without inference (GALE). To provide adequate minutes, the system must handle the various semi-independent threads of conversation, which requires recognizing not only shifts of focus from one agenda item to another, but also knowing when a motion is being made, which portions of the discussion favor a motion and which are arguing against the motion, which conversational turns respond to which earlier ones, etc.
4. From summaries by extract to transcript synthesis. Current summarization systems as seen in DUC select pieces of text from a full document(s) and create a "summary" by creating an ordering on the sentences/phrases selected. For generation of minutes, synthesized summaries, rather than extracts, offer the necessary dimension of identifying motions, key points made in discussion, decisions reached, action items assigned to whom by when, and items tabled for a future meeting.
5. From one-shot statistically trained systems to self-adapting systems. Today's technology is trainable, but both speech and natural language understanding systems are typically trained once, and have no ability to adapt automatically over time. Just as a human learns to adjust to the speech and terminology of a non-native speaker, the technology will be challenged to adapt automatically from examples, e.g., listening to examples of each participant in a series of ongoing meetings, or reading minutes of previous meetings to learn the structure, typical topics/concerns, and style of that group and its minutes, as well as of the individual attendees as described in the Socially-Aware Language Understanding challenge.
6. From isolated, independent research in speech and in language understanding to collaboration across the communities to solve the grand challenge. Though the GALE program has just started to bring the speech-to-text community and machine translation and distillation communities together, this challenge will bring in the speech community, language understanding, and summarization communities.

Challenge	**State of the Art**	**Required**
Everyday, multi-person speech	Broadcast news, e.g., much anchor speech, prepared reports from the field, some interviews	Multi-person, spontaneous speech with interruptions and emotion

Identifying structure and organization	Speaker change detection, speaker clustering (identification), simple thread tracking, and transcription	Recognition of structure, e.g., movement from agenda item to agenda item and introduction of new business items
Capturing connotation and effect	Literal meaning	Discourse structure and intent recognition, e.g., claims and rebuttals
Synthesizing what happens in a meeting	Summaries based on cutting & pasting from the source, e.g., selecting 1st sentence of a news article	Synthesis of major items according to group's style, e.g., action items, motions & results
Improvement through adaptation	One-shot/batch statistically trained systems	Systems that automatically adapt to speakers and domain being discussed

Table 4: Technical Challenges Requiring a Paradigm Shift for Generating Meeting Minutes

Program Parameters

A program must select:

- Data. 'Found' data that already exists or data that can be harvested / collected quickly is crucial. This includes recorded data from meetings in a single room or from teleconferences or videoconferences and should include a sizable corpus of the acoustic signal, transcripts with metadata (e.g., who is speaking, when), and examplar, humanly-produced minutes summarizing each meeting.
- Generality of meetings. Two styles of meetings should be included in a multi-year effort, e.g., a team meeting of scientists, after-action reviews of a unit, or collaboration of a team of analysts. If data is available at the start for one class of meetings, a second style could be added as a later challenge by requiring that participating teams record their own team meetings and generate their own training data.
- An evaluation paradigm. Given acoustic training, transcripts, and example minutes for a given group's meeting, measures of several dimensions, such as:
    - Accuracy in transcribing into text the speech of a meeting
    - Accuracy of minutes (the official summary) content, e.g., recall and precision of the facts included
    - Readability and suitability of various styles of minutes (variations include organization by subtopic, by subtopic thread, or following the conversational flow; producing fluent or headline/bullet-point summaries, etc.)
    - Capture of the intent / tone of meeting, that may be unstated, but a detectable take-away by anyone present in the meeting
    - Increase in accuracy as a result of adaptation after the first meeting
    - Increase in accuracy from adaptation as a result of corrected minutes
    - Increase in appropriateness for various recipients of the meeting minutes

Impact

The impact of systems that produce minutes of meetings may be near ubiquitous, given the number and frequency of meetings and the need to record the decisions and action items from them. Given the growth in distributed meetings enabled by modern telecommunications and the internet, collaboration is only going to increase. Such a suite of technologies can produce collective intelligence and a record of how it grew in a group setting. Furthermore, the collaboration among the various domain's scientists tackling this challenging problem may have unforeseeable technology spin-offs through their working closely together, as opposed to the tradition of the fields operating independently.

References

Banko, M. & Brill, E. (2001). Scaling to Very, Very Large Corpora for Natural Language Disambiguation. In Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics.

Charniak, E., Knight, K., & Yamada, K., (2003). Syntax-based Language Models for Statistical Machine Translation. In Proceedings of MT Summit IX.

Dagan, I., Glickman, O., & Magnini,B. (2006). The PASCAL Recognising Textual Entailment Challenge. Lecture Notes in Computer Science. Vol 3944, Jan 2006, 177-90.

Hirst, G., DiMarco, C., Hovy, E.H. & Parsons, K. (1997). Authoring and Generating Health-Education Documents that are Tailored to the Needs of the Individual Patient. In A. Jameson, C. Paris, & C. Tasso (Eds), Proceedings of Sixth International Conference on User Modeling. Sardinia, Italy. NY: Springer-Verlag (107–118). See *http://lhum.org*.

Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., Weischedel, R., (2006). OntoNotes: The 90% Solution. In Proceedings of HLT-NAACL

Lapata, M. & Keller, F. (2004). The Web as a Baseline: Evaluating the Performance of Unsupervised Web-based Models for a Range of NLP Tasks. In Proceedings of the 2004 HLT/NAACL Conference.

Mann, W.C. & Thompson, S. (1988). Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. Text 8(3). Pp. 243-281.

Och, F.J. (2005). Proceedings of Workshop on Statistical Machine Translation Foundations and Recent Advances. ACL Conference.

Quirk, C., Menezes, A. & Cherry,C. (2005). Dependency Treelet Translation: Syntactically Informed Phrasal SMT. In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics,Ann Arbor, Michigan.

Ringger, E.K., Moore, R.C., Charniak, E., Vanderwende, L. & Suzuki, H. (2004). Using the Penn Treebank to Evaluate Non-Treebank Parsers. In *Proceedings of the 2004 Language Resources and Evaluation Conference (LREC)*, Lisbon, Portugal.

Wiebe, J., Breck, E., Buckley, C., Cardie, C., Davis, P., Fraser, B., Litman, D., Pierce, D., Riloff, E., Wilson, T., Day, D., & Maybury, M. (2003). Recognizing and Organizing Opinions Expressed in the World Press. In Proceedings of 2003 AAAI Spring Symposium on New Directions in Question Answering.

Winograd, T., (1971). Procedures as a Representation for Data in a Computer Program for Understanding Natural Language. MIT-AI-TR-235.

Woods, W. A. (1970). Transition Network Grammars for Natural Language Analysis. Communications of the ACM 13:10.