Historical Development and Future Directions in Data Resource Development

Martha Palmer, Stephanie Strassel, Randee Tangi

Introduction

On February 25,26 2007, the second of two workshops entitled "Meeting of the MINDS: Future Directions for Human Language Technology, sponsored by the U.S. Government's Disruptive Technology Office (DTO), was held in Marina del Rey, California. "MINDS" is an acronym for **M**achine Translation (MT), **I**nformation Retrieval (IR), **N**atural Language Processing (NLP), **D**ata Resources (Data), and **S**peech Understanding (ASR) which were the 5 areas, each addressed by a number of experienced researchers, at this workshop. The goal of these working groups was to identify and discuss especially promising future research directions that could lead to major paradigm shifts and which are yet to be funded. As a continuation of a prior workshop where each group (except Data Resources) had first reviewed major past developments in their respective fields and the circumstances that led to their success, this workshop began by reviewing and refining the suggestions for future research emanating from the first workshop. The different areas then met in pairs to discuss possible cross-fertilizations and collaborations. Each area is responsible for producing a report proposing 5 to 6 "Grand Challenges."

1. Historically Significant Developments in Data Resource Development

The widespread availability of electronic text and the more recent advent of linguistic annotation have revolutionized the fields of ASR, MT and NLP.

Transcribed Speech Beginning with TI46[1] "the availability of common speech corpora for speech training, development, and evaluation, has been critical in creating systems of increasing capabilities. Speech is a highly variable signal, characterized by many parameters, and thus large corpora are critical in modeling it well enough for automated systems to achieve proficiency. Over the years, these corpora have been created, annotated, and distributed to the world-wide community by the National Institute of Science and Technology (NIST), the Linguistic Data Consortium (LDC), and others. The character of the recorded speech has progressed from limited, constrained speech materials to masses of progressively more realistic, spontaneous and “found” speech." [MINDS07 Speech Understanding Report].

TI46 is "a corpus of isolated spoken words which was designed and collected at Texas Instruments (TI) in 1980." There are 16 speakers, both male and female, speaking 46 words. It was followed by TIMIT, The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus. "TIMIT contains a total of 6300 sentences, 10 sentences spoken by each of 630 speakers from 8 major dialect regions of the United States." [2] Equally influential

[1] http://www.ldc.upenn.edu/Catalog/readme_files/ti46.readme.html
[2] http://www.ldc.upenn.edu/Catalog/readme_files/timit.readme.html

have been the "read" Wall Street Journal, the conversational speech Switchboard corpus and Hub4 for Broadcast News.

"The development and adoption of rigorous benchmark evaluations and standards, nurtured by the National Institute of Science and Technology and others have been critical in developing increasingly powerful and capable systems. Many labs and researchers have benefited from the availability of common research tools such as HTK, Sphinx, CMU LM toolkit, SRILM toolkit, etc. Extensive research support combined with workshops, task definitions, and system evaluations sponsored by DARPA and others have been essential to today's system developments." MINDS07 Speech Understanding Report].

Parallel Corpora The rebirth of statistical machine translation in the 1990's, after it was first introduced by Claude Shannon in 1949, is directly attributable to the availability of the Hansards,[3] the official records of the Canadian Parliament. These are kept by law in both French and English, and may be legally reproduced and distributed as long as "it is accurately reproduced and that it does not offend the dignity of the House of Commons or one of its Members." This provided the researchers at IBM with a large parallel French/English corpus of closely aligned, literal translations which is ideal for statistical word alignment techniques. The astonishingly accurate translations their system was able to produce revolutionized the Machine Translation field, and is the basis for the increasingly accurate statistical machine translation of Chinese and Arabic to English which is being currently funded by DARPA-GALE, and which is in daily use in Iraq.

Linguistic Annotation as Training Data The creation of the Penn Treebank (Marcus et al, 1993) and the word sense-annotated SEMCOR (Fellbaum, 1997) have shown how even limited amounts of annotated data can result in major improvements in complex natural language understanding systems. These annotated corpora have led to the training of stochastic natural language processing components which resulted in high-level improvements for parsing and word sense disambiguation (WSD), on the same scale as previously occurred for Part of Speech tagging by the annotation of the Brown corpus and, more recently, the British National Corpus (BNC) (Burnard, 2000). They have also encouraged the development of an increasingly wide variety of corpora with richer and more diverse annotation. These include the ACE annotations (Named Entity tags, nominal entity tags, coreference, semantic relations and events), semantic annotations such as sense tags (Palmer et al, 2004), semantic role labels as in PropBank, NomBank and FrameNet, and pragmatic annotations such as coreference (Poesio and Vieira, 1998, Poesio 2004a), TimeBank (Ferro et al, 2004; Pustejovsky et al, 2005) and the Penn Discourse Treebank.

---

[3] http://www2.parl.gc.ca/housechamberbusiness/ChamberSittings.aspx?View=H&Language=E&Parl=39&Ses=1

Publicly available electronic lexical resources WordNet, is a large lexical database of English, developed under the direction of George A. Miller which has been in widespread use since the 90's. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets). WordNet's structure makes it a useful tool for computational linguistics and natural language processing. Since WordNet is also freely and publicly available for download, there are no financial barriers to its use and it has become a de facto community standard for English vocabulary, providing a common sense inventory for systems doing English language processing. Richer lexical resources such as FrameNet (_) and VerbNet (__) all provide links to WordNet synsets. WordNets have now been created in dozens of other languages, many of them referring to the core English vocabulary from WordNet 1.5; a potential benefit to machine translation. WordNet synsets are organized hierarchically according to hypernyms, or super-types, providing the promise of generalizations from specific words to semantic types. However, this potential has not yet been realized due to the high polysemy of commonly used words. Large scale sense tagging efforts at Princeton, Colorado and ISI should provide sufficient data for successful training of automatic taggers.

Document Collections. "The Text REtrieval Conference (TREC), co-sponsored by the National Institute of Standards and Technology (NIST) and U.S. Department of Defense, was started in 1992 as part of the TIPSTER Text program. Its purpose was to support research within the information retrieval community by providing the infrastructure necessary for large-scale evaluation of text retrieval methodologies." This effort has been highly effective in fostering major advances in the Information Retrieval and Question Answering and has spawned several similar efforts in Europe and Asia.

"The TREC conference series has produced a series of test collections. Each of these collections consists of a set of documents, a set of topics (questions), and a corresponding set of relevance judgments (right answers). Different parts of the collections are available from different places as described on the data page (http://trec.nist.gov/data.html). In brief, the topics and relevance judgements are available at http://trec.nist.gov/data.html, and the documents are available from either the LDC (Tipster Disks 1--3) or NIST (TREC Disks 4--5), information on collections other than English can be found at http://trec.nist.gov/data.html."[4] These collections include: Ad hoc Test Collections, Web Test Collections, Blog Track, Confusion Track, Enterprise Track, Filtering Track, Genomics Track, HARD Track, Interactive Track, Legal Track, Novelty Track, Robust Track, Query Track, Question Answering Track, and the SPAM collection.

Benchmark corpora (to be completed)

2. Future Directions

1) **Science of annotation.** The success of applying machine learning techniques to

[4] http://trec.nist.gov/overview.html

annotated training data has had a major impact on the field. The current depth of NLP research is defined by the depth of linguistic annotation that is available, with supervised machine learning techniques doing an admirable job of automatically producing the same types of annotation. The higher the Inter-Annotator agreement is, and the greater the consistency and coherence of the original annotation, the better the performance of the trained systems. Not surprisingly there is now a resulting demand for more and more annotated data: the same types of annotations for different genres and different languages; newer, richer annotations for the original languages; parallel annotations of parallel corpora; the merging of annotations that were first done independently; and new formatting for pre-existing annotations that makes them easier to merge. For today's NLP systems, the annotation defines the task, and increasingly rich annotations are the key to more sophisticated systems. Clearly the annotation effort needs to become much more widely distributed to cope with this need. Unfortunately, there is evidence that when sites with no experience take on new annotation projects, the result is not always usable as training data, no matter how profound their understanding of the linguistics might be. At the moment successful annotation is more of an art than a science, and there are only three or four sites (in the world) that have the requisite depth of experience and know-how to produce useable annotations is a timely fashion. It is critical that these institutions codify their knowledge in such a way that it can be transferred readily to other sites. The field needs a concerted effort on creating an explicit description of a step by step process by which useful annotations can be achieved. Guidelines and methodologies are needed for community wide answers to the following questions:

- The corpus
    - What constitutes a balanced, representative, and timely corpus?
- The annotation
    - What linguistic phenomenon is the target of the annotation and how can that be represented?
    - How can a manual be created that ensures that the choices the annotators are faced with can be decided rapidly, accurately and consistently?
    - How can the usefulness of the resulting annotated corpora as training data be guaranteed and measured?
- The annotators
    - What qualifications, training and supervision do they need?
    - What are realistic productivity estimates?
    - What tools should they be provided with?
    - What are the principles for good annotation interface design?
- Inter-annotator agreements and disagreements
    - How should agreement be measured (kappas, confusion matrices, etc?) and what are realistic expectations for different types of tasks?
    - How can different sources for disagreements be identified and addressed, such as carelessness, vague or ambiguous instructions, and obscure or ambiguous data? Which ones are susceptible to good software design of annotation interfaces and which ones are not?

Solving these issues will require a substantial research effort which would be greatly facilitated by a regular technical forum for discussion such as an ACL SIG. A measure

of the success of this objective would be substantial progress on providing the public domain infrastructure described in (2). Each different type of annotation requires a stable, language independent methodology based on guidelines and universally accessible tools. The guidelines need to explicate the details of each individual annotation process as well as interactions between the different types of annotation. For instance, the guidelines for the Proposition Bank outline a process that begins with creating a Frame File for each individual verb in the corpus to be annotated. The Frames Files provide invaluable additional direction for the annotation of the individual verbs over and above the general annotation guidelines. The NomBank annotation in turn begins by referencing the verb Frame Files for associated nominalizations whenever possible. The same approach has been used successfully for the Chinese PropBank/NomBank. However, it needs to be tested on at least two or three additional disparate language families before it can be considered a stable part of a comprehensive linguistic annotation process.

The emergence of agreed upon principles and techniques for an "annotation science" is the only way that successful annotation can become widespread and the increasing demand can be met. Even in a rosy distant future where unsupervised techniques that do not need training data will prevail, annotated corpora will still be needed for testing and evaluation purposes.

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 the "best practice." The simplest way to ensure that everyone adheres to best practice guidelines and agreed upon formats is to have everyone using the same tools and APIs. This could also provide a framework for facilitating interaction with automatic taggers as pre-processing components, as described in (3). There are also similar data tracking needs shared by all annotation projects for which a toolkit could provide a general purpose solution. Making such a toolkit available to the international community 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. Coherently layered annotation clearly benefits both technology developers and resource developers. The CRI program at NSF now explicitly acknowledges annotation as an important research area contributing to Computing Research Infrastructure, but those funds are limited. The support of the DOD is needed to turn this into a serious research effort. Some of the key challenges of this program would be:

- Defining data desiderata to support machine learning techniques
- Defining principles for creating modular annotations which can easily be layered with other annotations. For example, clearly defined standoff xml annotations and compatibility between layers would allow for a standard query builder that knows how to access layered annotations.
- Implementing a annotation infrastructure starter kit for new efforts in linguistic resource creation that embodies these desiderate and principles. This would ideally be centrally localized where it can be downloaded, and where people can submit annotated data to be

archived and distributed. Copywrited data can't be shipped back and forth but standoff annotations can be.

- A regular forum that includes funders for more public discussion of data resource priorities such as
  a) New genres (e-mails, blogs, text messages, meetings, …)
  b) New languages for the resource kit
  c) Richer annotations

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 (Penn Treebank, Chinese Treebank). However, current annotation practices frequently fail to take advantage of this approach. 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. For simple classification tasks like Word Sense Disambiguation there are accepted practices such as active learning for doing this, (Chen, et. al., 2006). However, for more complex annotations such as syntactic structure, pinpointing novel or unfamiliar items in the data is an unsolved problem. Fundamental research is needed to develop informed active learning techniques for complex annotation. This will require collaboration between the researchers in the community doing annotation and those developing the machine learning techniques used for system building. Closer ties between annotators and the ASR, NLP, MT, IR and Machine learning communities are needed for joint efforts on developing techniques to aid data selection and quick access to modular automatic taggers for preprocessing of data. This will facilitate faster, easier, more complete integration of emergent technology into the human annotation pipeline, which will in turn improve the quality and availability of annotated data for those communities. The current emphasis on fast, cheap annotation impedes this process, since it does not allow time for experimentation with more sophisticated approaches. 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. A careful study of bootstrapping automatic taggers from small amounts of annotated data to find the minimal amounts of additional annotation required to achieve maximal performance is needed. Recognizing the importance of the improvement of annotation practices as a valuable research area in its own right will allow the time for progress on this front, and will encourage technology developers to hand over state-of-the-art systems for early integration. The goal should be, in addition to faster and most 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 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. We need to annotate smarter, not necessarily harder.

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, automatic taggers can be built to replicate whatever

representations they were trained on. This makes the search for increasingly rich levels of representation that can be successfully annotated the highest priority in the field. This is fundamental, theoretical research requiring expertise in linguistics, annotation and machine learning. Examples of immediate pressing priorities for the following application areas are listed:

- **Machine Translation** - Parallel corpora with parallel annotations, i.e., source documents in multiple languages, translated into multiple languages (all pairings) and sentence aligned; with (complete or partial) cohesive annotations across source & translated docs in each language.

- **Spoken Language Understanding** - Rich corpora transcriptions that include phonetic information, term detection, semantic roles and sense tags, discourse structure, dialect info, etc.

- **Natural Language Processing** – Identification of events and relations between them (causal, temporal, modal, etc.), discourse relations, etc. Also parallel corpora (see above) for research on projecting annotations onto (translated data in) another language given reference annotations in one language.

- **Information Retrieval** – [to be completed]

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. These resources would enable rapid ramp-up of a basic news understanding system for text and for speech, and allow information retrieval over document sets. There is an existing effort at LDC to develop "Language resource kits" for a small set of languages: Urdu, Thai, Hungarian, Bengali, Punjabi, Tamil, and Yoruba. The resources being developed for these languages include:

- monolingual text, parallel text, part-of-speech taggers, morphological analyzers, and Named Entity annotation.

This level of resource development is at best minimal. Speech recognition needs are completely ignored, as are syntactic and semantic structure. A more complete resource kit would also contain:

- **text** - at least 100K words of parallel text (news domain), tagged for nominal entities & coreference, basic syntactic annotation, basic predicate argument structure, topics + relevance judgments for news articles;
- **speech** – a pronouncing dictionary, a minimum of 100 hrs 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**. In the quest for improving the portability of supervised stochastic systems, one under-utilized resource is the lexicon. Many supervised approaches depend heavily on lexical cues, and balk when given data with out-of-vocabulary lexical items. One would think that general purpose lexical resources providing similarity links between items would be helpful in alleviating this impasse. So far that has not been the case. Either the lexical resources are not organized in the most effective way or the stochastic systems are not set up to make use of them, or both. However, there is widespread agreement that certain new resources are extremely desirable and could be of great benefit. For example, in order to address the pressing problem of correctly recognizing references to the same person in different documents which can also be in different languages, the community has been clamoring for a normalized list of references with pointers to instantiations in documents; a cross-document, cross-lingual entity co-ref corpus. From one perspective, this can be viewed as a name lexicon that, in addition to the list of names, includes multilingual spelling variants of each name as well as multiple document instances. It can include nominal entity references as well as named entity references. All of the documents in the set would have all instances of every listed item marked. There are several other examples of "lexicons" that are effectively specific types of lists of items with corresponding pointers to instances in data. A classic example is a bilingual lexicon, which in this case would also include pointers to word usage instances in documents in a parallel corpus for the two languages. Example-based Machine Translation systems contain a vast amount of information which can be seen as another variant of this type of resource. Again, this resource could potentially be of use for improving the portability capabilities of statistical Machine Translation systems. Tying the "word lists" to instances in corpora helps bridge the gap between the classic lexicon and the needs of a stochastic system.

Another drawback of existing lexicons is that 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 be sure 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 could be very useful.

The next step is to create monolingual resources that can provide support for automatically detecting semantic similarity. These resources should also be empirically grounded, with pointers to instances that illustrate the desired paraphrases and inferences. For instance, phrases such as "the rising stock prices" and "the surge in the stock market" can both be linked to a similarity set such as "rise, surge," in the same way that a bilingual lexicon links English/Spanish translations "rising, levantando" and

corresponding aligned sentences. The current surge in sense tagged data is making progress in this area, but much more can be done.

Looking even farther ahead, we will soon need even richer lexical resources (monolingual and bilingual) that describe event hierarchies, and are closely integrated with ontologies associated with rich knowledge bases that can support inferencing. This is the only way we will be able to move forward from our current shallow semantic processing to the deeper levels of comprehension needed for complex question answering and information extraction. Here the empirical grounding might be phrases or parts of sentences that provide the evidence for perceived entailments. These "nuggets" of information correspond closely to the "compositional units" that the MT community would like to be able to focus on.

Bibliography [to be completed]