Some Notes on Performance Evaluation for Natural Language Systems

Jerry R. Hobbs
Artificial Intelligence Center
SRI International
Menlo Park, California

February 24, 2004

The first thing to say is that the universal reaction among researchers at SRI to the question of performance evaluation was an expression of deep reservations about any explicit set of test criteria. Any two systems are probably incomparable since they are aimed at achieving different capabilities. Where they aim at the same capabilities, one may exhibit better performance while the other experiments with methods of greater potential. Enhanced capability in one area may introduce greater problems in another area. Finally, it is often possible to gear a system specifically for the explicit criteria, thereby disguising its real shortcomings. Nevertheless, we have attempted to come up with some guidelines that an evaluator might use. (I've discussed these issues with Paul Martin, Barbara Grosz, Stu Scheiber, Fernando Pereira, David Israel, and others, and have taken many ideas from them, but they can't be held responsible for any mistakes in this attempted synthesis.)

Ultimately, natural language research for the military has to be justified in terms of making information more available to commanders. The processes by which they acquire essential information must be quicker, more economical, or more reliable. A natural language system is ultimately part of an organization's "knowledge amplification" capability. It is necessary to find tasks where a natural language capability can be shown off to its best advantage. There are two principal candidates: tasks involving a user's nonroutine interactions with a computer system (e.g., with a data base management system), where the freedom of expression allowed by natural language makes it unnecessary to learn and figure out how to express the problem in a special-purpose language; and tasks now involving human processing of texts already in natural language, where the processing is routine, highly repetitious, and well-understood and where the computer can capitalize on its greater speed and reliability (e.g., message processing). There are two ways of demonstrating the utility of such capabilities to commanders, first by means of demo systems, second by placing usable systems in the field. However, demo systems often dazzle observers for the wrong reasons, and field systems employing emergent technology often fail to take hold for reasons unrelated to their ultimate promise or to the progress they represent in technology development. Finer-grained evaluation procedures are therefore required. These procedures, however, can only be convincing to people already convinced of the utility of a natural language capability.

Just as administrators cannot evaluate new airplanes themselves, they cannot evaluate natural language systems themselves. They must depend on the evaluation of experts. Their confidence is enhanced to the extent that the procedures are systematized. One can imagine then having a panel of established experts in computational linguistics evaluate various experimental systems and research efforts. They should be able to present a succinct summary of the results of the evaluation that, although it does not go into the full complexities of the problem, at least does not tell lies about the state of the art and correctly reflects their assessment of the progress that is being made.

It ought to be possible to make up a set of guidelines for these evaluations. Whenever any of us tries out someone's system, we have our ways of banging on it. What we would like is some systematization of this banging. The rest of these remarks constitute a first pass at such guidelines. It is largely a list of problem areas. It should be viewed as a kind of "cheat list" for evaluators. It is important that evaluation remain broken into categories. One should not attempt to determine some kind of "composite" score for a system, and one should definitely not use these criteria for guiding research. The real breakthroughs in research come from viewing problems in new ways.

The first distinction that should be made is between the competence of a system and its performance. An evaluation of competence would say things like "It handles relative clauses and interprets compound nominals." An evaluation of its performance would say things like "It answers 62% of the queries real users ask it." These two aspects of evaluation should be kept separate. A system that represents significant progress in competence may be a disaster in performance for some trivial reason, for example, simply because it lacks a large lexicon. If one were to make up a collection of test sentences, there should be two sets: a set of sentences motivated by theoretical, linguistic considerations, and a set of sentences collected from typical users. The former would be used to test the competence of specific components of the system. The latter would be used to evaluate performance, i.e., the overall functionality of the system in some environment. Since performance evaluations are highly dependent on the task environment, I will concentrate on competence evaluations for the rest of these remarks.

It is not appropriate to have a simple "checklist" of problems that we can tick off when we've solved them. All the problems in this field are open-ended. We don't solve them; we only make progress on them. It would be more appropriate to have a two-dimensional chart. Along one dimension would be the problem areas. Along the other would be an indication of the level of sophistication attempted by the system. Where a component of a system lies along this dimension could be a matter of subjective judgment, assigning it to one of four or five categories. More likely, the levels of sophistication would correspond to some abstract characterization of the techniques we know. Ultimately, of course, what is important is not some measure of a component's sophistication, but the algorithms that are used. These should be made explicit. (This is perhaps a place to say that claims of "trade secret" are generally nonsense. What AI companies are selling is not their knowledge but their sweat.) Moreover, an attempt should be made to describe them in a manner that is as nonidiosyncratic as possible, although it often requires a person with some distance to do the rational reconstruction.

Several caveats are in order. First, two systems with components attempting different levels of sophistication are essentially incomparable. A simple syntactic approach to pronoun resolution, for example, might work on 90% of the test cases, whereas an inferential approach only worked on 40%. Yet the syntactic approach would have reached its inherent limit, while the inferential approach had the potential of surpassing that limit. Secondly, a more sophisticated approach is not necessarily better. In many cases the problem is to find intermediate solutions that combine the potential coverage of the sophisticated solutions with the efficiency of the simple solutions. Thirdly, it is okay for a system to be at the simple end of the scale in some category if another part of the system is going to pick up the slack. For example, in the TACITUS system the semantic representations of sentences lie at the trivial end of the scale on the assumption that pragmatic processing will compensate, and similarly, many systems built at Yale sacrifice sophistication in syntax on the assumption that sematics and pragmatics will make up the difference. Finally, I've written of statistics on a set of test sentences for expositional purposes, but it is not really desirable to make that the mode of evaluation; it is too coarse-grained and militates against adding difficult cases to the test criteria.

The following is a first cut at a list of problem areas:

Syntax:

Parsing technique: What parsing algorithm is used? (This area is well-understood enough that this can be stated in standard terminology.) How does the parsing interact with other processes?

Grammar formalism: To what extent is the grammar formalism a variant of well-understood formalisms? For new formalisms, how does it enhance modularity, i.e., limit propagation of side-effects from small changes in the grammar. Does it make the grammar easier to write and easier to understand?

Subcategorization: What range of complements can verbs take? How are subcategorization constraints implemented?

Determiners: Which determiner structures are recognized and which are excluded? E.g. "Many fewer than all the expected five hundred people arrived."

Sentential and infinitival noun phrases: E.g., "That John exists is a fact." "To think is to exist." "Seeing is believing."

Auxilliaries: What compound tenses and modals are handled? E.g. "He could not have been seeing her."

Verb phrase deletion: E.g. "John is leaving and Mary is too."

Comparatives: E.g., "John was very much happier than Mary was sad." "John is more than three inches taller than Bill."

Wh-movement, including questions, relative clauses, and headless relatives: Pied piping: "a book the author of which ..." Unbounded dependencies: "Which banks do they appear to be willing to consider allowing us to remove gold from?" Interaction with conjunction.

Extraposition: Are the constituents attached correctly?

Adverbials: What range, e.g., of subordinate clauses or measure or time phrases, are recognized? What positions are they allowed in? What means are used to capture regularities in spite of the positional freedom?

Conjunction: Is a general approach attempted, or are only a few of the most common cases handled? Is gapping handled? Are conjunctions handled as part of the grammar or as part of the parser? What is done about the structural ambiguities introduced by conjunctions? What about the ambiguities of “the red and black cars”?

Syntactic ellipsis: What methods are used to recover the full sentence?

Syntactic ambiguity: Are different parses generated for syntactically ambiguous sentences (implicitly or explicitly)? Is some (syntactic) attempt made to choose among them or to rank them? This may differ for different phenomena; for example, DIALOGIC generates separate parses for prepositional phrase ambiguities, but not for very compound nominal ambiguities.

Ill-formed input: Is there an attempt to characterize ill-formed input systematically? What happens when the parser fails? Are partial results given? Is some attempt made to make sense out of the partial results?

Semantics:

Case relations: How are they represented? Are they recovered systematically? E.g., “John broke the window.” vs. “The hammer broke the window.” vs. “The window broke.”

Tense and aspect: What kind of information is extracted from the fact that a sentence is in, e.g., the perfect or progressive tense?

Other encodings of temporal information: E.g., clock and calendar, temporal prepositions, indexicals. Temporal information encoded in other lexical items, as in “The patient awoke with a fever.”

Spatial prepositions and other encodings of spatial information: How much information is extracted from them? What approach is taken toward their inherent vagueness and context-dependence?

Quantifiers: Not just “all” and “some” but also “most”, “three”, “few”, etc.

Collective entities and substances: E.g., “Three men lifted the piano.” “All Eastern coal contains some sulphur.”

Abstractions: E.g., “red” in “Red is my favorite color.” “Copper is an element.”

Comparatives: Among the relevant entities in “John is taller than Bill.” are John, Bill, John’s height, Bill’s height, and a tall-ness scale. How are all of these pieced together in the logical form of the sentence?

Attributives: E.g. representation of the reference group. Relation to quantitative descriptions of same information.

Events, actions, and processes: Are they represented explicitly or only implicitly?

Adverbials: Are they properties of events and conditions? Higher operators? Predicate modifiers?

Modalities: What range of modalities are handled and how are they represented?

Propositional attitudes, like "believe": How are they represented? How do they interact with disjunction and negation? What approach is taken to the classical problems of intensional operators, the de re-de dicto distinction and the problem of identity? Does the approach coordinate with an approach to reasoning within propositional attitudes?

Questions and imperatives: How are they represented?

Local Pragmatics:

What speech acts are handled? E.g., only questions, only declaratives. Is "I want to know the length of the Kennedy." treated as an assertion or a question? If as a question, what method is used, the trivial translation, "I want to know P." ==¿ "P?", or a more sophisticated approach involving reasoning about WANTs?

Referring expressions: There are two orthogonal classifications of referring expressions: the syntactic-lexical distinctions among proper names, definite noun phrases, indefinite noun phrases, pronouns, one-anaphora, and omitted arguments; and the functional distinctions among referential, non-specific, generic, and attributive uses. For any given system, we must ask not just whether it handles definite noun phrases, but also whether it just handles the referential case or makes some attempt at the others. There is a range of sophistication for possible solutions to, e.g., the pronoun resolution problem. One can use purely syntactic criteria; one can use selectional information; one can attempt to do more general inferencing to match the properties given by the pronoun's environment; one can try to determine and employ attentional factors like focus; or pronoun resolution can "fall out" of recognizing the sentence's place in a text or task structure.

Compound nominals: What is the implicit relation being encoded? Again there is a range of sophistication, from encoding fixed phrases in the lexicon, to choosing among a small set of pre-established relations on the basis of selectional information, to using more general inferential processing to find the most plausible relation.

Interpretation of vague predicates, like "have", "do", "of", "in" and the possessive: Again, what specific relation is intended?

Metaphor: Again there is a range of sophistication, from having separate, precoded word senses, to attempting to determine the interpretation by inferential processes. Note that there is a trade-off between having precoded metaphorical word senses and the lexical ambiguity problem.

Metonymy, indirect reference, or coercion: What is the intended referent and the implicit relation between it and the apparent referent? Again, one can identify a range of sophistication in approaches. The embedding operators can be given separate word senses. There can be a small set of possible coercion functions distinguished by selectional information. Or more general inferential processing can be attempted.

Lexical Ambiguity: What sense should be chosen for lexically ambiguous words? Again a range, from selectional criteria to more general inferencing.

Syntactic Ambiguity: Which parse should be chosen? Heuristics based on syntax vs. selectional criteria vs. more general inferencing.

Pragmatic Ellipsis: For sentence fragments embedded in discourse, how is their interpretation to be determined?

Coherence: How is the structure of the text recognized?

Task Pragmatics:

Question-Answering: I assume Bonnie Webber can give a more elaborate list of issues than I can. E.g., response-planning.

Task-oriented dialogues, message processing: How are the elements of the text matched with the model of the domain?

To summarize, competence evaluation could be done in terms of a matrix of the sort shown in Figure 1 (the numbers on the chart represent how some random system might perform, although let me reiterate a distaste for statistics of this sort):

For each problem area there could be a set of test sentences that exercised that particular capability and not the others. E.g., for relative clauses, sentences like "This is the book John believed Bill bought."

Each system would have to be able to provide a "probe" for each problem area, some way of determining whether the system succeeds or fails on the test sentence. For most syntactic analyzers, for example, the parse tree would serve as a probe. In a system which did not produce parse trees, some other evidence would have to be produced that, e.g., the book in the above sentence is recognized as the object of the buying event.

The method the system uses would then need to be classified as to its level of sophistication – simple, intermediate, or sophisticated. This is a subjective judgment, but one that an expert in computational linguistics can make. What is being judged here is the POTENTIAL of the method, regardless of its current performance.

The system could then be tested and given a score in each problem area in the column corresponding to its method's level of sophistication. This gives a rough overall characterization of the system's capabilities.

Problem \ Level of Area \ Technology	None	Simple	Intermediate	Sophisticated
Syntax:				
Subcategorization				90%
Relative Clauses				77%
Conjunction			46%	
...				
Semantics:				
Case Relations			92%	
Tense & Aspect		51%		
...				
Pragmatics:				
Pronoun Resolution		85%		
Metonymy				12%
Metaphor	X			
...				
Response Planning				
...				
System Robustness				

Comparison:

This evaluation can then be used in two ways and not used in a third:

1. A system can be evaluated at two different times and the evaluations compared. Progress has occurred if the numbers go up or if the methods used have become more sophisticated.

2. The evaluation would give a reasonable statement of the system's capabilities in some absolute sense. We could see how far it is from the "great system in the sky" that is sophisticated and achieves 100% in every area.

The comparison that cannot be made is one between two different systems. Every system will attempt to achieve a different subset of capabilities by means of methods of different levels of sophistication. For example, if System A has sophisticated, effective syntax and no pragmatics, and System B has mediocre syntax and mediocre, experimental pragmatics, there is no way we can say that one is better than the other.