Workshop on the Evaluation of Natural Language Processing Systems

Wayne Hotel
Wayne, Pennsylvania

December 7 - 9, 1988

Martha Palmer
Unisys Paoli Research Center

Tim Finin
Unisys Paoli Research Center

Sharon M. Walter
Rome Air Development Center

1. Introduction

In the past few years, the computational linguistics research community has begun to wrestle with the problem of how to evaluate its progress in developing natural language processing systems. With the exception of natural language interfaces there are few working systems in existence, and they tend to focus on very different tasks and equally different techniques. There has been little agreement in the field about training sets and test sets, or about clearly defined subsets of problems that constitute standards for different levels of performance. Even those groups that have attempted a measure of self-evaluation have often been reduced to discussing a system's performance in isolation - comparing its current performance to its previous performance rather than to another system. As this technology begins to move slowly into the marketplace, the lack of useful evaluation techniques is becoming more and more painfully obvious.

In order to make progress in the difficult area of natural language evaluation, a *Workshop on the Evaluation of Natural Language Processing Systems* was held in December of 1988 at the Wayne Hotel in Wayne, Pennsylvania. There were two basic premises for this workshop:

- It should be possible to discuss system evaluation in general without having to state whether the purpose of the system is ``question-answering" or ``text processing." Evaluating a system requires the definition of an application task in terms of input/output pairs which are equally applicable to question-answering, text processing, or generation.

- There are two basic types of evaluation, *black box evaluation* which measures system performance on a given task in terms of well-defined input/output pairs, and *glass box evaluation* which examines the internal workings of the system. For example, glass box performance evaluation for a system that is supposed to perform semantic and pragmatic analysis should include an examination of predicate-argument relations, referents, and temporal and causal relations. Since there are many different stages of development that a natural language system passes through before it is in a state where black box evaluation is even possible (see Figure 1), glass box evaluation plays an especially important role in guiding the development at early stages.

With these premises in mind, the workshop was structured around the following three sessions:

1. Defining the notions of ``glass box evaluation" and ``black box evaluation" and exploring their utility.

2. Defining criteria for ``glass box evaluation."

3. Defining criteria for ``black box evaluation."

Calls for participation were sent by electronic mail to several national and international mailing lists and posted on numerous internet newsgroups and resulted in a large response. A program committee consisting of Beth Sundheim (NOSC), Ed Hovy (ISI), Tim Finin (Unisys Paoli Research Center), Lynn Bates (BBN), Martha Palmer (Unisys Paoli Research Center), Mitch Marcus (CIS, University of Pennsylvania) was put together to plan the workshop and invite participants. Those respondents interested in participating in the workshop were asked to describe their interest in the topic, describe any relevant work done in the area, and provide an abstract on evaluation topics they would want to present. A total of fifty people were invited to participate.

It was hoped that the workshop would shed light on the following questions:

1. What are valid measures of ``black box" performance?

2. What linguistic theories are relevant to developing test suites?

3. How can we characterize efficiency?

4. What is a reasonable expectation for robustness?

5. What would constitute valid training sets and test sets?

6. How does all of this relate to measuring progress in the field?

2 Background

Before looking at the distinctions between ``black box" and ``glass box" evaluation, it is necessary to examine the development of a natural language system a little more closely. There are several different phases, and different types of evaluation are required at each phase. The various phases are summarized in Figure 1.

NLP System Development Steps

1. Picking the application.
2. Characterizing the necessary phenomena.
3. Selecting relevant theories, if available.
4. Developing and testing algorithms that implement these theories.
5. Implementing the first pass at the system.
6. Characterizing new phenomena that appear, especially those having to do with interactions.
7. Fine-tuning algorithms to improve efficiency, and also replacing algorithms as the characterization of the phenomena changes.
8. Second pass at implementation.
9. Third pass at an implementation in which a focus is placed on issues of extensibility.
10. Fourth and final pass at the implementation in which the system moves into a production environment. This stage pays special attention to issues of robustness.

Figure 1: There are a number of different stages in the development of a natural language processing system. Different kinds of evaluations are required and/or possible at the different stages.

Speaking very roughly, the development of a natural language processing system is usually sparked by the needs of the particular application driving it, whether it be question-answering, text processing or machine translation. What has happened in the past is that, in examining the requirements of such an application, it has quickly become apparent that certain phenomena, such as pronoun reference, are essential to the successful handling of that application. It has also quickly become apparent that for many of these phenomena, especially semantic and pragmatic ones, past linguistic analysis has very little to offer in the way of categorization. Even where it might offer a fairly rigorous account of the phenomenon, as in the case of syntax, it has very little to say about useful algorithms for efficiently producing syntactic analyses and even less to say about interaction between different types of phenomena. So, almost before beginning implementation, a great deal of effort in the computational linguistics community must of necessity be devoted to tasks which can rightly be seen as belonging to theoretical linguistics. The *Discourse Canon* that Bonnie Webber prepared for the Mohonk Darpa Workshop [7] is an excellent example of the type of ground work that must be done prior to serious attempts at implementation, and must be continued throughout and subsequent to said implementation. The field needs many more such ``canons" for other semantic and pragmatic phenomenon.

Algorithm development is equally important, and can also be carried out independently of, or in parallel with, an implementation. We have several different algorithms for syntactic parsing, and ways of comparing them (and ways of proving that they are all equivalent), but very few algorithms for semantics and pragmatics.

Implementing an algorithm and using it for an application is a separate stage of development. Progress cannot, however, be measured in terms of accurate output until a system that uses particular algorithms to handle particular phenomenon has been implemented. In addition to methods for measuring the performance of entire systems, we also need ways of measuring progress in characterizing phenomenon and developing algorithms which will contribute to system development.

Once a system is up and running, the accuracy of its output can then be measured. The different pieces of the output can be associated with the phenomena that have to be handled in order to produce each piece. For example, consider a phrase from the *trouble failure report* domain [1]:

``Replaced interlock switch with new one."

In order to accurately fill in the slot in the database field associated with the new_part_installed(_) relation, the ``one" anaphora has to be correctly resolved, requiring a complex interaction between semantic and pragmatic analysis.

It is possible to have two systems that produce the same output, but do it very differently. This is where such issues as efficiency, extensibility, maintainability and robustness come in. A more efficient implementation, for example, may be able to support a larger domain. With a more general implementation, it should be easier to extend the scope of the system's domain or to port the system to an entirely new domain. A system with a more convenient or robust interface will be easier to use and, one would suppose, be used more often.

2.1 Black box evaluation

Black box evaluation is primarily focused on ``what a system does." Ideally, it should be possible to measure performance based on well-defined Input/Output pairs. If accurate output is produced with respect to particular input, than the system is performing correctly. In practice, this is more difficult than it appears. There is no consensus on how to evaluate the correctness of semantic representations (see below), so output has be in terms of some specific application task such as *data base answering or template fill.* This immediately introduces an astonishing amount of variation among systems. However, black box evaluation is an important goal for natural language systems and progress is being made [6]. In addition to the accuracy of the output systems could also be evaluated in terms of their user-friendliness, modularity, portability and maintainability. How easy are they to use, how well do they plug into other components, can they be ported and maintained by someone who is not a system expert? In general, it should be possible to perform black box evaluation without knowing anything about the inner workings of the system - the system can be seen as a black box, and can be evaluated by system users.

2.2 Glass box evaluation

In contrast, glass box evaluation attempts to look inside the system and find ways of measuring how well it does something, rather than simply whether or not it does it. To this end, glass box evaluation should include an examination of relevant linguistic theories and how faithfully they are implemented. It should measure the system's coverage of a particular linguistic phenomena or set of phenomena and the data structures used to represent them. And it should be concerned with the efficiency of the algorithms being used. Many of these tests could only be performed by a system builder. They are especially useful in attempting to measure progress when a system is under development. In addition, one way of looking inside a system is to look at the performance of one or more components. A black box evaluation of a particular component's performance could be seen as a form of glass box evaluation.

A glass box evaluation involves performing evaluations on the parts of the system with respect to some decomposition. The parts obtained depend on the nature of the decomposition. Two common ways of decomposing an NL system are: into *functional modules* each of which performs a specified task, and into *analysis phases* during which different functions can be performed.

3. Workshop Format

The workshop began with a set of presentations that discussed evaluation methods from related fields: Speech Processing, Machine Translation, and Information Retrieval. This was followed by a panel of reports on evaluations of natural language systems. After lunch the workshop heard a presentation on black box evaluation, and then broke up into working groups for the rest of the afternoon. After reporting on the results of the working groups, there was a short break before the banquet which included a demonstration of the Dragon speech recognition system. The second day began with the presentation of another method of black box evaluation, and then moved on to introduce the topic of glass box evaluation. A panel discussed necessary characteristics for corpora to serve as training sets and test sets for black box evaluation of systems and components of systems. The workshop then broke up into working groups again. The final session was devoted to reports of working groups and summarization of results. These activities are all described in more detail below.

3.1 Reports from related fields

The opening presentation at the Workshop on Evaluation of Natural Language Processing Systems was made by Dave Pallett of the National Institute of Standards and Technology (formerly the National Bureau of Standards). Mr. Pallett's presentation demonstrated depth of technical knowledge, including ``lessons-learned" from the historical development of speech recognition benchmark tests. This presentation, and Mr. Pallet's subsequent active participation, set a high scientific tone for the workshop.

Experiences in evaluation from the NLP-related technical fields of Speech Processing, Information Retrieval, and Machine Translation were next presented, and critiqued, by knowledgeable researchers in each of the respective areas. The issues common to each presentation, with respect to the specific technical area, were: the criteria for evaluation (what should be measured), the measurement procedure (how to get quantitative measures), the problems associated with the evaluation methods, and the similarities/differences between evaluation in their field and that of NLP.

3.2 Panel on evaluation attempts

The next session of the workshop was made up of presentations by people who either had particular suggestions for evaluation, or who had experiences in ``evaluating" systems to relate. Lyn Bates (BBN) suggested using the BBN corpus of inputs for a Navy database as a standard corpus for evaluation. The corpus is a resource of BBN, Inc but is in the public domain and available without fee.

John Nerbonne of Hewlett-Packard Labs discussed his ideas for a test suite of inputs for evaluating database query systems and the associated criteria for evaluation. A necessity for formulating the suite of inputs would be, according to Nerbonne's plan, a description of the range of phenomena to be captured in the suite. This idea was expressed numerous times by various people during the workshop and is the goal of the RADC/COES project ``Benchmark Investigation/Identification". The objective of this project is to develop a "descriptive terminology" with which software components for processing Natural Language (NL) can be described in a standard way. The Descriptive Terminology will be a collection of non-subjective terms with associated definitions and procedures for applying them appropriately. Terminology is to be developed for two areas of NL Interface concern: processing issues, those issues relating to human-computer interaction, and issues relating to interface-to-target application integration procedures, those issues addressing the process of applying an NL interface to domain software.

Debbie Dahl of Unisys presented her experiences in applying the DARPA-sponsored, Prolog-based PUNDIT system to an Information Retrieval task on Air Force RAINFORM messages. The main items to rate in Information Retrieval are ``recall", the percentage of relevant messages retrieved and ``precision", the percent of retrieved messages that are determined to be relevant. Dr. Dahl included a critique of their evaluation in her presentation.

Anatole Gershman is with Cognitive Systems, Inc. Cognitive Systems engineers Natural Language Interfaces to specific user needs, basing their systems on Roger Schank's Conceptual Dependency theory. The Cognitive Systems experience was the most concrete, grounded-in-experience, and objectively evaluated example of ``engineering" NL interfaces. System users develop interface specifications: specific questions to be asked and their variations, the interface vocabulary, user goals to be satisfied, and a target interaction scenario. In the system evaluation, users are given task assignments to complete using the interface and success is measured by the successful completion of the task.

Dick Kitteridge, of Odessey Research Associates, described a task to assess the feasibility of using commercial machine translation systems for English-to-French translation of patents. Three candidate systems, based on the same technology, were compared in the effort.

3.3 Black box evaluation working groups

The first set of working groups met on the afternoon of December 8th. Beth Sundheim of NOSC introduced the topic by giving a possible task for black box evaluation of natural language processing message understanding systems [6]. This was a very specific proposal suggesting a training set of 100 messages from a specific domain, and two separate test sets, one consisting of twenty messages and another of ten. The performance was to be evaluated with respect to a frame-filling task. Working groups were divided according to type of system application: Message Understanding, Text Understanding, Data Base Question-Answering, Spoken Language Understanding, and Generation. The Message Understanding Group was to discuss and refine Beth's evaluation proposal. The other groups were to make as much progress as they could on defining similar evaluation tasks for their application areas. After meeting for several hours, the groups gathered together to exchange results.

3.3.1 Results

There was general agreement among the workshop participants that useful black box evaluations can be done for the message understanding and data base question-answering task domains. It was also agreed that more general systems aimed at text understanding and dialogue understanding were not good candidates for black box evaluation due to the nascent stage of their development, although individual components from such systems might benefit from evaluation. Workshop participants were pleasantly surprised by the results of the generation group which came up with a fairly concrete plan for comparing performance of generation systems based on the message understanding proposal. A perennial problem with all of these proposals, with the exception of the message understanding proposal is the lack of funding. Conferences and workshops need to be organized, systems need to be ported to the same domain so that they can be compared, etc., but there is very little financial support for evaluation of systems that could make these things possible.

Message Understanding. Chair: Ralph Grishman

The working group on message understanding formed a consensus for the structure of the evaluation task to be performed at the second Message Understanding Conference (MUCK-II). A domain of 135 messages was selected. 100 messages were to be distributed as a training set. A set of twenty messages were reserved as a preliminary test set, with a further five messages reserved as a final test set. At least ten messages were put aside for a possible additional test at a third Message Understanding Conference. The measurements of performance would be based on a common frame filling task, in which the frame summarized key information in the message. For further information on details of how the test was set up, see [6].

Data Base Question Answering. Chair: Harry Tennant

The data base group was one of the most reticent in terms of making concrete proposals for evaluation. This is undoubtedly because they were the only group that had actually had experience with system evaluation. They had seen how difficult it is to design black box evaluation tasks that measure just what they are supposed to, and which cannot be misled. A standard problem with these tasks is that someone can develop a system that is aimed solely at performing well on the task, and which pays only lip service to the original problem. The data base group committed themselves to an overall design of an evaluation that would include a standard corpus with a database, along with sets of questions and answers that could be shared among all participants. In the past the availability of such material has proven a major stumbling block to comparative evaluations of question-answering systems. It is possible that the official airline guide database, to which some in the spoken language community are committing may also prove of benefit to the question answering community. This group was quite definite about the benefits to be accrued from a conference devoted to natural language processing for database applications, similar to the Message Understanding Conference (MUCK) organized by Beth Sundheim, but pointed out that there was currently no funding available to organize such a conference.

Text Understanding. Chair: Lynette Hirschman

For the purposes of the workshop, text understanding was defined to be distinct from message understanding. It is expected to consist of "well-formed," i.e., not telegraphic, sentences and is multi-paragraph. It was generally agreed that thorough analysis of text of this length was beyond the state of the art. It was suggested that it might be possible to do a partial analysis that could be used for information retrieval purposes. Dave Lewis agreed to be a focal point for discussion of applying text analysis techniques to information retrieval.

Dialogue Understanding. Chair: Mitch Marcus

Dialogue understanding was defined as being naturally occurring dialogue, either written or spoken. With the exception of very limited domain applications, such as portions of the dialogue between an air traffic controller and a pilot, this was also considered to be beyond the state of the art. This group poured an astonishing amount of creativity into attempting to design a black box evaluation task but so many portions of the task had to be contrived that it was clearly more trouble than it was worth.

Generation. Chair: Dave McDonald

This group may have started out being the most skeptical about the possibility of reasonable evaluation but ended on a surprisingly positive note. They suggested that generation systems first be ported to the same domain as the message understanding systems, the Opreps domain. They could then be given the task of generating messages based on the filled-in frames that are the output of the message understanding systems. It is immediately obvious that the filled-in frames leave out a lot of the information about style and other important presentation attributes that are normally fed into a generator, so it would be possible to generate several different textual versions of the same messages. This is also a disadvantage however, in making it difficult to determine the ``correct" generation.

3.4 Introduction to glass box evaluation

The second day began with a presentation by Fred Jelinek (IBM) in which he advocated an approach for black box evaluations of parsers. This suggestion depended on gathering large amounts of data that was marked with syntactic parts of speech and bracketed in such a way that it could be automatically compared to the output of a syntactic parser. This type of data collection using ``naive grammarians" (briefly trained people off of the street) has already been initiated by the University of Lancaster in a joint project with IBM. Jelinek also suggested that evaluating the performance of the parsers should involve the use of probabilistic evaluation criteria.

This was followed by an overview of the goals of glass box evaluation by Martha Palmer. Black box evaluation tests a system building effort and has little implication for future systems. Glass box evaluation, however, attempts to test the state of a technology and should indicate whether the particular methods employed show promise for the future.

The next item was a panel chaired by Jerry Hobbs with Mitch Marcus and Ralph Weischedel as participants. The panel discussed various aspects of glass box evaluation. In particular, black box evaluation of components, such as the need for training sets and test sets and the need for public support were discussed. During this panel, Mitch Marcus outlined a proposal the University of Pennsylvania was submitting to Darpa to collect data to be used for testing parsers automatically, borrowing the name TREEBANK from the Lancaster project. This data bank has subsequently been funded and includes both written and spoken language. Even at this point in the workshop it could be seen that the group was moving towards a consensus on the evaluation of parsers, but that there were still major difficulties with respect to other components. Semantics was mentioned as a principal stumbling block. The task of a natural language system is the analysis of sentences into their ``meanings." Since there is no consensus in the field on how to represent this meaning, it makes it especially difficult to evaluate the output of such systems.

After the panel Bonnie Webber outlined the task on glass box evaluation for the afternoon working groups. The groups were divided according to class of phenomena: syntax, semantics, discourse, and knowledge representation. There was also a group that discussed overall system issues. In her presentation Bonnie drew distinctions between methodology (features and behaviors) and metrics (measures made on those features or behaviors.) She also distinguished evaluating a theory from evaluating the algorithm that attempts to implement it. She cited two examples. The first contrasted *Tree Adjunction Grammar* (TAG) as a linguistic theory [4] with the several algorithms have been used to implement TAG parsers: Extended CKY parser, Extended Earley parser, Two-pass extended Earley parser based on lexicalized TAGs, and a DCG parser using lexicalized TAGs. The second example described *Centering* as a theory for resolving anaphoric pronouns [5,3] as opposed to attempts to use a centering approach to resolving pronouns in an implementation [2].

The groups were tasked with identifying, within their area, the range of items (components, algorithms, theories, data structures, interactions, etc.) that are suitable for evaluation and for evaluation metrics. If possible they were to suggest methodologies and metrics for those items. They were also supposed to indicate how the other items could be made ready for evaluation.

3.5 G'ass box working group results

Syntax. Chair: Richard Kittredge

This group focused mainly on Mitch Marcus's proposal and added several refinements. By the time the results were presented a consensus had been reached that this would be a valuable direction for the community, and that it was indeed feasible. The proposal as it is being implemented consists of a large amount of data, both written language and spoken language which will be divided into training sets and test sets. Similarly with Jelinek's TREEBANK, it would involve marking the data with a theory-neutral syntactic structure. One difference is that linguistics graduate students would do the annotations instead of the ``naive grammarians" used in the Lancaster work. It was agreed that the annotation could probably include lexical class labels, bracketing, predicate argument relationships and possibly reconstruction of control relationships, wh-gaps, and conjunction scope. Eventually it would be desirable to include coreference anaphora, prepositional phrase attachment and comparatives although it is not clear how to ensure consistent annotation. The parsers would be delivered to the test site with the ability to map their output into the form of the corpus annotation for automatic testing. The test results would be returned to parser developers with overall scores as well as scores broken out by case, i.e., percentage of prepositional phrase bracketings that are correct.

Semantics. Chair: Christine Montgomery

This was clearly one of the most controversial areas, with very little agreement among the group members as to even lists of phenomena. Predicate argument relations, temporal relations and modifiers (including prepositional phrase attachment) were perhaps the only phenomena agreed upon as semantic, with no agreement as to instructions for annotation or methodologies for evaluation. There are several different major camps in this area: a conceptual dependency camp that shows little interest in relating semantic deep structures to surface syntactic phenomena, a lexical semantic camp that wishes to rely solely on dictionary input for building semantics, and a logic camp that attempts to reduce everything to Montague semantics, to mention just a few. There are a few people who are trying to bridge some of the gaps, but at the moment there are as many different styles of semantic representation as there are researchers in the field. This means that the only possible form of comparative evaluation must be task related. Good performance on such a task might be due to all sorts of factors besides the quality of the semantic representations, so it is not really an adequate discriminator. At the recent Darpa Spoken Language Workshop in Philadelphia, Martha Palmer proposed three necessary but not sufficient steps for moving towards more consensus in this crucial area:

1. agreement on characterization of phenomena
2. agreement on mappings from one style of semantic representation to another
3. agreement on content of representations for a common domain

An obvious choice for a common domain would be the Opreps domain recently used for the Message Understanding Conference. There are several state of the art systems that are performing the same task for the same domain using quite different semantic representations. It would be quite useful to take four of these systems, say NYU, SRI, Unisys and GE systems, and compare a selected subset of their semantic representations in-depth. It should be possible to define a mapping from one style of semantic representation to another and pinpoint the various strengths and weaknesses of the different approaches.

Pragmatics and Discourse. Chair: Candy Sidner

This group attempted to list all of the capabilities that might be demonstrated by an interface claiming to have pragmatic understanding and discourse capabilities, and then determine definitive measures of those capabilities. Toward the end of the session Dr. Sidner asked if group members felt that the work was of benefit: if we remained, locked in the room and paid our salaries, was the task worthwhile. There was unanimous agreement that it was.

The group's basic premise was that they would need a large corpus annotated with discourse phenomena. This would allow them to evaluate the effect of individual components upon the system as a whole and upon other components such as syntax and semantics. It would also allow an individual subcomponent's behavior to be observed. They listed the discourse phenomena shown in Figure 2, and marked the ones about which they could develop instructions for an annotator. The others might take a bit more thought. It was agreed that the topics for a subsequent meeting would include experimenting with text annotations and designing training sets and test sets.

o turn-taking
- o referring expressions, including anaphora, ``do so", respectively

o multi-sentence text
o sensitivity to user's goals and plans
o model of user's beliefs, goals, intentions, etc.
o use of Gricean maxims
o use of speech acts
o interpretation and use of temporal and causal relationships

- o *part/whole, member/set relationships*

o vague predicate specification

- o determination of implicit arguments in predicate-argument relationships

o metaphor and analogy
o schema matching
o varying depth of processing on the basis of certain criteria
o focus of attention and saliency of entities

- o ellipsis

o style and social attitudes
o deixis

Figure 2: A list of discourse phenomena to use in annotating a discourse corpus. A dash (-) indicaties that precise annotation instructions could be given.

Knowledge Representation Frameworks. Chair: Tim Finin

This group looked at approaches for evaluating *knowledge representation and reasoning* (KR&R) systems in support of natural language processing applications. They began by pointing out that KR&R services provided fall into two classes: (1) providing a *meaning* *representation language* (MRL), (2) providing inferential services in support of syntactic, semantic and pragmatic processing. The group noted that the MRL class should probably be broadened to include languages for representing dialogues, lexical items, etc. In addition, the group laid out a spectrum of activities which are included in a KR&R, shown in Figure 3.

- **theory** - Is there an underlying theory which gives meaning to the KR&R system? What is known about the expressiveness of the language and the computational complexity of its reasoning?
- **languages** - How does the KR&R system as a practical language for expressing knowledge? How easy or difficult is it to define certain concepts or relations or to specify computations?
- **systems** - KR&R systems are more than just an implementation of an underlying theory. They require good development environments: knowledge acquisition tools, debugging tools, interface technology, integration aids, etc. How extensive and good is this environment?
- **basic models** - A KR&R system often comes with some basic, domain-independent modules or models, such as temporal reasoning, spatial reasoning, naive physics, etc. Are such models available and, if they are, how extensive and detailed are they?

Figure 3: There are several dimensions along which a knowledge representation and reasoning system might be evaluated.

The group suggested three evaluation methodologies. The first was aimed at evaluating a KR&R system's suitability as a meaning representation language. One way to evaluate a potential MRL is to have a standard set of natural language expressions to try to express in the MRL. This provides an evaluator with some idea of the expressiveness and conciseness of the KR&R system as an MLR.

A second evaluation methodology follows the ``Consumer's Reports" paradigm and involves developing a checklist of features. An extensive list of KR&R features could be developed for each of the dimensions given in Figure 3. Scoring how well KR&R systems handle each of these features provides a way to compare different systems.

The final evaluation technique is to hold a MUCK-like workshop aimed at evaluating the performance of the NLP system's underlying KR&R system. The group outlined a proposal for organizing a workshop to do an evaluation of the KR&R aspects of a natural language processing system based on the MUCK Workshop models.

3.6 Final session

The ``next steps" toward development of NLP evaluation capabilities was the topic of a final workshop discussion session. Mitch Marcus of the University of Pennsylvania offered to be the focal point and coordinator for an informal bulletin board for the development of a standardized database for NLP evaluators. As mentioned above, a number of workshops focusing on evaluation of NLP systems for various tasks (Message Understanding, Generation, Data Base Question-Answering, and Knowledge Representation Frameworks), were proposed for the future. Candy Sidner made note that RADC's Benchmark Investigation/Identification effort was of particular relevance to the workshop proceedings and that the activities and results of that effort may be of wide interest as they unfold.

4. Workshop Conclusions

Several concrete results came out of the workshop. In particular, a consensus was reached on the black box evaluation task for the second Message Understanding Conference, and a consensus was also reached on the desirability of a common corpus of annotated language, both written and spoken, that could be used for training and testing purposes. Since the workshop, the Message Understanding Conference has been held with interesting and useful results and the TREEBANK project at the University of Pennsylvania has received funding and has begun. This should eventually lead to more formalized testing and comparisons of parsers. Evaluation is becoming a more prevalent topic at NL workshops, such at the one to be held at RADC in September of 1989, and the Darpa Spoken Language Community is working hard to construct a general evaluation procedure for the various contractors. However, most of the other specific workshops suggested, such as Data Base Question-Answering, Generation, Knowledge Representation and Pragmatics and Discourse do not have any funding sources available. The most difficult problems remain unresolved. We still do not know how to effectively measure improved performance during the crucial development phase, apart from peer review. We have little agreement on semantic representations and there are still large classes of phenomena that have yet to be characterized in a scholarly fashion. However, a first step has been made and with sufficient focus and, of course, sufficient funding, the next steps will follow.

References

[1] Catherine N. Ball. Analyzing explicitly-structured discourse in a limited domain: trouble and failure reports. In *Proceedings of the DARPA Speech and Natural Language Workshop*, Philadelphia, PA, February 1989.

[2] Susan E. Brennan, Marilyn W. Friedman, and Carl J. Pollard. A centering approach to pronouns. In *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics*, pages 155-162, Stanford, CA, 1987.

[3] Barbara Grosz, Aravind Joshi, and Scott Weinstein. Providing a unified account of definite noun phrases in discourse. In *Proc. 21st Annual Meeting*, pages 44-50, Assoc. for Computational Ling., Cambridge MA, June 1983.

[4] A.K. Joshi and S. Weinstein. Control of inference: centering. In *7th !nternational Conference on Artificial Intelligence*, pages 385-387, Int'l Joint Conf. on Artificial Intelligence, Vancouver, Canada, August 1981.

[5] Aravind K. Joshi and Leon S. Levy. Phrase structure trees bear more fruit than you would have thought. In *Proceedings of the 18th Annual Meeting of the Association for Computational Linguistics,* University of Pennsylvania, Philadelphia, PA, June 1980.

[6] B. M. Sundheim. Plans for a task-oriented evaluation of natural language understanding systems. In *Proceedings of the DARPA Speech and Natural Language Workshop*, pages 197-202, Morgan Kaufman Publishers, Inc., Philadelphia, PA, 1989.

[7] Bonnie Lynn Webber. Discourse canon. in preparation. Presented at the Mohonk Darpa Workshop, May, 1988.