# **NEAL-MONTGOMERY NLP SYSTEM EVALUATION METHODOLOGY**

#### Sharon M. Walter

Rome Laboratory RL/C3CA Griffiss AFB, NY 13441-5700 walter@aivax.rl.af.mil

# ABSTRACT

On what basis are the input processing capabilities of Natural Language software judged? That is, what are the capabilities to be described and measured, and what are the standards against which we measure them? Rome Laboratory is currently supporting an effort to develop a concise terminology for describing the linguistic processing capabilities of Natural Language Systems, and a uniform methodology for appropriately applying the terminology. This methodology is meant to produce quantitative, objective profiles of NL system capabilities without requiring system adaptation to a new test domain or text corpus. The effort proposes to develop a repeatable procedure that produces consistent results for independent evaluators.

## **1. INTRODUCTION**

An appreciable drawback to current corpus-based (eg., [BBN; 1988], [Flickinger, et al; 1987], [Hendrix, et al; 1976], [Malhotra; 1975]) and task-based (eg., ["Proceedings"; 1991]) methodologies for evaluating Natural Language Processing Systems is the requirement for transportation of the system to a test domain. The expense and time consumption are sizable and, as the port may be minimal or incomplete, the evaluation may be based on a demonstration of less than the full potential of the system. Further, current evaluation methodologies do not fully elucidate NLP system capabilities for possible future applications.

Under contract to Rome Laboratory, Dr. Jeannette Neal (Calspan Corporation) and Dr. Christine Montgomery (Language Systems Incorporated) are in the final months of developing an NLP system evaluation methodology that produces descriptive, objective profiles of system linguistic capabilities without a requirement for system adaptation to a new domain. The evaluation methodology is meant to produce consistent results for varied human users.

### **1.1. Evaluation Methodology Description**

Within the Neal-Montgomery NLP System Evaluation Methodology each identified linguistic (lexical, syntactic, semantic, or discourse) feature is first carefully defined and explained in order to establish a standard delimitation of the feature. Illustrative language patterns and sample sentences then guide the human evaluator to the formulation of an input that tests the feature on the NLP system within the system's native domain.

Based on clear and specific evaluation criteria for test item inputs, NLP system responses are scored as follows:

S: The system successfully met the stated criteria and demonstrated understanding with respect to the feature under test.

C: The system responded in a way that was correct (that is, correctly answered the question posed), but the criteria were not met.

P: The system responded in a way that was only partially correct.

F: The system responded in a way that was incorrect, failing to meet the criteria.

N: The system was unable to accept the input or form a response (for example, the system vocabulary lacks appropriate words to complete a test input).

Each linguistic feature is tested by more than one methodology item to make sure that results are not based on spurious responses, and each item examines only one asyet-untested capability, or one as-yet-untested combination of capabilities. Test inputs that are dependent on capabilities previously shown to be unsuccessful are avoided. Scores are then aggregated into percentages for hierarchically-structured classes of linguistic capabilities which produce descriptive profiles of NLP systems. The profiles can be viewed at varying levels of granularity. Figure 1 shows a sample system profile from the top level of the hierarchy.

Note that the scoring nomenclature (above) has been refined and expanded since project experiments produced the profiles and results presented in this paper. In Figures 1 and 2, "Unable to Compose Input" is equivalent to an 'N' in the newer nomenclature. A score of "Indeterminate" earlier meant the human evaluator could not determine if the NLP

	System XYZ		Successes		Failures		Unable to Compose Input		Indeterminate		Average Time Per Item
		#	%	#	%	#	%	#	%		
I.	Basic Sentences	18	81.82	2	9.09	0	0.00	2	9.09	0:50	0:02:16
II.	Simple Verb Phrases	5.5	78.57	1.5	21.43	0	0.00	0	0.00	0:36	0:05:09
Ш.	Noun Phrases	63	56.25	34	30.36	6	5.36	9	8.03	4:31	0:02:25
IV.	Adverbials	3.5	70.00	1.5	30.00	0	0.00	0	0.00	0:23	0:04:36
V.	Verbs and Verb Phrases	12.5	65.79	3.5	18.42	0	0.00	3	15.79	1:40	0:05:16
VI.	Quantifiers	36	45.00	39	48.75	1	1.25	4	5.00	3:04	0:02:18
VII.	Comparatives	25	39.06	38	59.38	1	1.56	0	0.00	3:00	0:02:49
νШ.	Connectives	28.5	83.82	5.5	16.18	0	0.00	0	0.00	2:10	0:03:49
IX.	Embedded Sentences	2	40.00	3	60.00	0	0.00	0	0.00	0:20	0:04:00
<b>X</b> .	Reference	6	50.00	5	41.67	1	8.33	0	0.00	1:18	0:06:30
XI.	Ellipsis	5	29.41	10	58.82	2	11.76	0	0.00	1:05	0:03:49
XII.	Semantics of Events	14.5	37.18	17.5	44.87	2	5.13	5	12.82	2:17	0:03:31

Figure 1: A Top Level Evaluation Profile of an NLP System

system correctly processed the test input. The new system of scores will be applied for the final project selfassessment activities.

The columns at the far right of Figure 1 display the total time (in hours and minutes) the user required to complete that section of the evaluation, and the average time per item (hours:minutes:seconds) for the section.

Figure 2 displays part of the evaluation to the methodology's most detailed level of granularity.

### 2. PROJECT SELF-ASSESSMENT

In March and September of 1991 rigorous project assessments provided valuable feedback into the design of the Neal-Montgomery NLP System Evaluation Methodology. For each assessment, three people applied the methodology to each of three NLP systems, for a total of eighteen applications. Assessment personnel, knowledgeable with respect to interface technology but not trained linguists, were distinct from the methodology development team.

The consistency of system profiles resulting from these applications, the examination of test inputs composed during the assessments, records of oral commentary by evaluators, and responses to a post-evaluation questionnaire have been used as measures of the accuracy of methodology results. For the September assessment phase, Figure 3 shows, for each section of the methodology, the percentage of items for which the assessment team gave the same score to each system. For example: the data points for the adverbial section indicate that all three people gave the same assessment of System 2's skills for adverbials (they agreed in every instance), they agreed 60% of the time on System 1's adverbial skills, and they agreed only 20% of the time for System 3's adverbial skills. The inconsistency of scores in this section has prompted the development team to refine the methodology's adverbial section.

NLP systems used for assessments to date have included three NL database query systems and two MUC-3 systems.

Focusing on reliability rather than feedback into methodology design, four people will apply the Neal-Montgomery NLP Evaluation Methodology to each of two systems for the third (and final) project self-assessment in April 1992.

### **3. TOWARD THE FUTURE**

Evaluation "standards" are not developed and adopted without a period of review, rumination, and tweaking by the relevant user community. It is our hope therefore, in distributing the Neal-Montgomery NLP System Evaluation Methodology to the technical community, to stir interest that may lead to the eventual consideration of the methodology as the basis for a standard evaluation tool for NLP system capabilities.

The Neal-Montgomery NLP System Evaluation Methodology is due for completion and delivery to Rome Laboratory in May of 1992. It will become immediately available at that time to all interested parties. Requests should be made to the author of this paper. Reviewer comment, critique, and suggestions for the methodology are invited.

System XYZ		Successes		Failures		Unable to Compose Input		terminate	Total Time	Average Time Per Item
	#	%	#	%	#	%	#	%		
									0.50	0.00.16
I. Basic Sentences	18	81.82	2	9.09	0	0.00	2	9.09	0:50	0:02:16
1 Declarative Sentences	0	0.00	0	0.00	0	0.00		100.0		
2 Imperative Sentences	1	100.00	0	0.00	0	0.00	0	0.00		
3 Interrogative Sentences	17	85.00	2	10.00	<u>0</u>	0.00	1	5.00		
3.1 What-questions	5	100.00	0	0.00	0	0.00	0	0.00		
3.1.1 What as Pronoun	2	100.00	0	0.00	0	0.00	0	0.00		
a) with BE			0		0		0			
b) with DO			0		0		0			
3.1.2 What as Determiner	3	100.00	0	0.00	0	0.00	0	0.00		
a) with verb	1	I	0		0		0			
b) with BE	1		0		0		0			
c) with DO	1		0		0		0			
3.2 Who-questions	2	100.00	0	0.00	0	0.00	0	0.00		
a) with verb	1		0		0		0			
b) with DO	1		0		0		0			
3.3 Where-questions	0.5	25.00	0.5	25.00	0	0.00	1	50.00		
a) with BE	0.5		0.5		0		0			
b) with DO	0		0		0		1			
3.4 When-questions	1.5	75.00	0.5	25.00	0	0.00	0	0.00		
a) with BE	0.5		0.5		0		0			
b) with DO	1		0		0		0			
3.5 Which-questions	3.5	87.50	0.5	12.50	0	0.00	0	0.00		
a) with BE np	1		0		0		0			
b) with verb	1		0		0		0			
c) with BE adj	1		0		0		0			
d) with DO	0.5		0.5		0		0			
3.6 How-questions	1.5	75.00	0.5	25.00	0	0.00	0	0.00		
3.6.1 How [Adj][BE-Verb][NP]?	1	100.00	0	0.00	0	0.00	0	0.00		
3.6.2 How [Adv][BE-Verb][NP]?	0.5	50.00	0.5	50.00	0	0.00	0	0.00		
3.7 Yes/No questions	3	100.00	0	0.00	0	0.00	0	0.00		
a) with BE np	1		0		0		0			
b) with BE adj	1		0		0		0			
c) with DO	1		0		0		0			

Figure 2: Detailed Evaluation Profile for 'Basic Sentences'

My sincere thanks to Jeannette Neal of the Calspan Corporation and to Beth Sundheim for their valuable critique on early versions of this paper.

# REFERENCES

1. BBN Systems and Technologies Corporation, Draft Corpus for Testing NL Data Base Query Interfaces, NL Evaluation Workshop, Wayne, PA, December 1988.

- Flickinger, D., Nerbonne, J., Sag, I., and Wasow T. "Toward Evaluation of Natural Language Processing Systems", Hewlett-Packard Laboratories Technical Report, 1987.
- 3. Hendrix, G.G., Sacerdoti, E.D. and Slocum, J. "Developing a Natural Language Interface to Complex Data", Artificial Intelligence Center Technical Report, SRI International, 1976.

- 4. Malhotra, A., "Design Criteria for a Knowledge-Based Language System for Management: An Experimental Analysis", MIT/LCS/TR-146, 1975.
- 5. Neal, J. G., Feit, E.L., and Montgomery, C.A., "An Application-Independent Approach to Natural Language Evaluation", submitted to ACL-92.
- 6. Neal, J.G. and Walter, S.M. (ed.) "Natural Language Processing Systems Evaluation Workshop", Rome Laboratory Technical Report, 1991.
- Read, W., Quilici, A., Reeves, J., Dyer, M., and Baker, E.; "Evaluating Natural Language Systems: A Sourcebook Approach", Coling-88.
- 8. "Proceedings of the Third Message Understanding Conference", Morgan Kaufmann Publishers, 1991.



Figure 3: Percentage of Agreement Among Evaluators for Each System