[Human Language Technology, Plainsboro, 1994]

Evaluation in the ARPA Machine Translation Program: 1993 Methodology

John S. White, Theresa A. O'Connell

PRC Inc.
McLean, VA 22102

ABSTRACT

In the second year of evaluations of the ARPA HLT Machine Translation (MT) Initiative, methodologies developed and tested in 1992 were applied to the 1993 MT test runs. The current methodology optimizes the inherently subjective judgments on translation accuracy and quality by channeling the judgments of non-translators into many data points which reflect both the comparison of the performance of the research MT systems with production MT systems and against the performance of novice translators. This paper discusses the three evaluation methods used in the 1993 evaluation, the results of the evaluations, and preliminary characterizations of the Winter 1994 evaluation, now underway. The efforts under discussion focus on measuring the progress of core MT technology and increasing the sensitivity and portability of MT evaluation methodology.

1. INTRODUCTION.

Evaluation of Machine Translation (MT) has proven to be a particularly difficult challenge over the course of its history. As has been noted elsewhere (White et al., 1993), assessment of how well an expression in one language is conveyed in another is loaded with subjective judgments, even when the expressions are translated by professional translators. Among these judgments are the extent to which the information was conveyed accurately, and the extent to which the information conveyed was fluently expressed in the target language. The inherent subjectivity has been noted, and attempts have been made in MT evaluation to use such judgments to best qualitative advantage (e.g., van Slype 1979). The means of capturing judgments into quantifiably useful comparisons among systems have led to legitimate constraints on the range of the evaluation, such as to the scope of the intended end-use (Church and Hovy 1991), or to the effectiveness of the linguistic model (Jordan et al. 1992, Nomura 1992, Gambäck et al. 1991).

The ARPA MT Initiative encompasses radically different approaches, potential end-uses, and languages. Consequently, the evaluation methodologies developed for it must capture quantifiable judgments from subjectivity, while being relatively unconstrained otherwise. This paper presents the 1993 methodologies, and the results of the 1993 MT evaluation. We further discuss the preliminary status of an evaluation now underway that greatly increases the participation of the entire MT community, while refining the sensitivity and portability of the evaluation techniques.

2. MT EVALUATION IN THE ARPA MT INITIATIVE

The mission of the ARPA MT initiative is "to make revolutionary advances in machine translation technology" (Doddington, personal communication). The focus of the investigation is the "core MT technology." This focus tends, ultimately, away from the tools of MT and toward the (fully automatic) central engines. It is well understood that practical MT will always use tools by which humans interact with the algorithms in the translation process. However, the ARPA aim is to concentrate on fully-automatic (FA) output in order to assess the viability of radical new approaches.

The May-August 1993 evaluation was the second in the continuing series, along with dry runs and pre-tests of particular evaluation methods. In 1992, evaluation methods were built on human testing models. One method employed the same criteria used in the U.S. government to determine the competence of human translators. The other method was an "SAT"-type evaluation for determining the comprehensibility of English texts translated manually into the test source languages and then back into English. The methods have been replaced by methods which maintain familiarity in terms of human testing, but which are both more sensitive and more portable to other settings and systems. The Fluency, Adequacy, and Comprehension evaluations developed for the 1993 evaluation are described below; system outputs from 1992 were subjected to 1993 methods, which determined their enhanced sensitivity (White et al., op. cit.).

The 1993 evaluation included output from the three research systems, five production systems, and translations from novice translators. Professional translators produced reference translations, by which outputs were compared in the Adequacy evaluation, and which were used as controls in the Comprehension evaluation.

The research systems were:

- CANDIDE (IBM Research: French - English(FE)), produced both FA and human-assisted (HA) outputs.

Candide uses a statistics-based, language modeling MT technique.

- PANGLOSS (Carnegie Mellon, New Mexico State, University of Southern California: Spanish - English (SE)), produced three output types: fully automatic pre-processing, interactive pre-processing, and post-edited (PE). Both pre-processing operations are mapped into one version (XP) for evaluation purposes, though the difference in performance between the operational types was measured. The Pangloss system uses both knowledge-based and linguistic techniques.
- LINGSTAT (Dragon Systems: Japanese - English (JE)), performed in human-assisted mode. Lingstat is a hybrid MT system, combining statistical and linguistic techniques.

To provide comparison against state of the art FAMT, production systems ran in fully automatic mode. These systems are in current commercial use and developed over a wide range of subject areas. SPANAM, from the PAN AMERICAN HEALTH ORGANIZATION (PAHO) produced SE. SYSTRAN, a commercial system, produced FE. Three unidentified systems based in Japan each contributed JE. Their outputs were made available to the test and evaluation by Professor Makoto Nagao at Kyoto University.

Manual translations (MA) were provided by novice, usually student, translators at each of the research sites. These persons also developed the human-assisted outputs, controlled for pre-/post-test bias. Finally, expert manual translation of the same material into English was performed as a reference set as noted above.

SYSTEM TESTS

The first phase of the ARPA MT Evaluation was the System Test. The research and production sites each received a set of 22 French, Japanese or Spanish source texts for translation into English. Each set comprised eight general news stories and 14 articles on financial mergers and acquisitions, retrieved from commercial databases. The lexical domain was extended in 1993 to include general news texts to determine whether the training and development of the systems was generalizable to other subject domains. French and Spanish texts ranged between 300 and 500 words; Japanese articles between 600 and 1,000 characters.

EVALUATION COMPONENTS

The evaluators were eleven highly verbal native speakers of American English. Evaluation books were assembled according to a matrix based on a Latin square, designed to guarantee that each passage was evaluated once and that no evaluator saw more than one translation version of a passage. Because of technical problems, two of the Kyoto system outputs were evaluated in a subsequent evaluation that reproduced as closely as possible the construct of the preceding evaluation. The 1993 series tested the systems with source-only text, measuring the results with a suite of three different evaluations.

All participants evaluated first for fluency, then adequacy and finally for comprehensibility. Fluency and an adequacy components contained the same 22 texts. The comprehension component included a subset of nine to twelve of these texts. The Comprehension Evaluation was presented to evaluators last, in order to avoid biasing the performance of the fluency and adequacy over the passages that appeared in the comprehension set.

Fluency Evaluation

The Fluency Evaluation assessed intuitive native speaker senses about the well-formedness of the English output on a sentence by sentence basis. Evaluators assigned a score from one to five with five denoting a perfectly formed English sentence.

Adequacy Evaluation

The Adequacy Evaluation measured the extent to which meaning present in expert translations is present in the FAMT, HAMT, PE and MA versions. In order to avoid bias toward any natural language processing approach, passages were broken down into linguistic components corresponding to grammatical units of varying depths, generally confined to clause level constituents between 5 and 20 words in length. Average word count within a unit was 11 for SE and FE, 12 for JE. The average number of fragments for a passage varied: 33 for FE, 41 for JE, 31 for SE. The evaluators viewed parallel texts, an expert translation broken into brackets on the left and the version to be evaluated presented in paragraph form on the right. They were instructed to ascertain the meaning present in each bracketed fragment and rate the degree to which it was present in the right column on a scale of one to five. IF the meaning was absent or almost incomprehensible, the score was one; if it was completely represented the score was five.

Comprehension Evaluation

The Comprehension Evaluation measured the amount of information that is correctly conveyed, i.e. the degree to which a reader can find integral information in the passage version. This evaluation was in the format of a standardized comprehension test. Questions were developed based on the expert versions and then applied to all translation versions. Evaluators were instructed to base their answers only on information present in the translation. The Comprehension Evaluation is probably the most portable evaluation, as it is a common test format for literate English speakers.

RESULTS OF THE 1993 EVALUATION

The evaluations resulted in a total of over 12,500 decision points. These are in turn represented on two axes: the time ratio (x-axis) and normalized quality (y-axis). Both axes represent results as scores on a 0-1 scale. The time ratio is the ratio of the time taken to produce a system translation compared to the time taken for the novice MA translation. Thus, the novice MA translations all appear at time value 1. Since time taken to translate is not recorded for the FAMT systems, all of these are set at time 0. The quality (that is, fluency, adequacy, or comprehension) axis is the raw score, divided by the scoring scale (5 for fluency/adequacy, 6 for comprehension), in turn divided by the number of decision points (sentences for fluency, fragments for adequacy, or questions for comprehension) in the total passage set for that language pair.

Common characteristics can be observed in all of the evaluation measurements taken in 1993. First, it is evident that all of the HAMT systems performed better in time than the corresponding MA systems. This is a change from 1992, where one system took more time to operate than it took the same persons to translate manually. Each PE system also performed better in adequacy, and very slightly better in fluency, than the MA translations. While a reasonable and desirable result, this outcome was not necessarily expected at a relatively early stage in the development of the research systems. Another general observation is that PE versions scored better in quality than non-post-edited (i.e., raw FAMT or interactively pre-processed) versions. This too is an expected and desirable result. The benchmark FAMT for French and Spanish (SPANAM and SYSTRAN, respectively) scored better in quality than the non-post-edited research systems, except in fluency, where CANDIDE scored .040 higher than SYSTRAN's .540.

It was expected that comprehension scores would rise with the amount of human intervention. This proved true for FE. At .896, CANDIDE HAMT scored highest for FE comprehension; SYSTRAN (.813) scored above CANDIDE FAMT (.729). PANGLOSS SE scores also demonstrated this trend: FA at .583, HA at .750 and PE at .833, however, the HA and PE are unexpectedly below SPANAM (.854). LINGSTAT HA .771 also scored higher than the JE FAMT: KYOTO A (.479) KYOTO B (.5625) and KYOTO C (.563).

COMPARISON BETWEEN 1992 AND 1993 SYSTEM PERFORMANCE

Figures 1 and 2 show comparisons and trends between 1992 and 1993 for the elements of data and evaluation that are comparable. These include the fluency and adequacy measures for all of the 1993 test output and that portion of the 1992 data that was based on source-only text. The Comprehension Evaluation was not compared, since the 1992 data involved back-translations, and the numbers of questions per passage was different, thus creating the potential for uncontrolled bias in the comparison.

In 1993 all systems improved both in time and in fluency / adequacy over 1992. The PANGLOSS system shows the most apparent improvement in time, from 1.403 in 1992 to. 691 in 1993. LINGSTAT also shows a considerable improvement from .721 to .395. All ARPA research systems showed improvement in fluency and adequacy over 1992 scores. CANDIDE FAMT scores increase from .511 to .580 in fluency and .575 to .670 in adequacy. PANGLOSS PE improved from .679 to .712 for fluency and rose from .748 to .801 in adequacy. LINGSTAT improved from .790 in 1992 to .859 in fluency and went from .671 to .707 in adequacy.

It should also be noted that the benchmark systems used in both 1992 and 1993 (SYSTRAN French and SPANAM) showed improved fluency/adequacy scores as well. For fluency, SYSTRAN improved from .466 to .540; for adequacy, SYSTRAN went from .686 to .743. SPANAM went from .557 to .634 for fluency and from .674 to .790 for adequacy. It was verified that these are reflections of system improvements.

1993 demonstrated a significant increase in sensitivity of the evaluation methodology. Sensitivity is gauged by computing an F ratio, i.e., the correlation between independent values. A high F ratio indicates that the range of values is wide; the wider the range of values the more sensitive the method is. For the Fluency Evaluation, the F ratio rose from 3.158 in 1992 to 12.084 in 1993. In the Adequacy Evaluation, the F ratio rose from 2.753 to 6.696.

1994 EVALUATION IN PROGRESS

The 1994 Evaluation presently underway focuses on core FAMT technology. Its scope has been broadened to increase sensitivity and portability. In keeping with the ARPA MT Initiative goal to foster development of FAMT, input will move away from HAMT and include a larger proportion of FAMT. To better measure the expanded lexical capabilities of the systems under development, half of the test passages will be general news articles. The Winter 1994 evaluation alone will generate 25,000 data points to manage human subjectivity. This increase in data points has been accomplished by successfully porting the methodology to evaluation of 14 production systems in addition to the three ARPA research systems. To maximize the randomness of passage assignment in the evaluation matrix, the Latin square has been replaced with a matrix ordered by a random number generator. The methodology has been simplified to optimize the elicitation of intuitive judgments. For example, the fluency component which formerly measured only well-formedness has been modified to recognize the influence of contextual meaning.

The broadened scope of the 1994 Evaluation offers benefits for the evaluation of the core technology for the profoundly different systems of the ARPA MT Initiative. It also contributes to the advancement of the MT community as a whole through providing a consistent portable suite of evaluation methodologies.

REFERENCES

1. Church, Kenneth, and Eduard Hovy. 1991. "Good Applications for Crummy Machine Translation." In Jeannette G. Neal and Sharon M. Walter (eds.) *Proceedings of the 1991 Natural Language Processing Systems Evaluation Workshop.* Rome Laboratory Final Technical Report RL-TR-91-362.
2. Gambäck, Björn, Hiyan Alshawi, David Carter, and Manny Rayner. 1991. "Measuring Compositionality in Transfer-Based Machine Translation Systems." in Neal and Walter (eds.).
3. Jordan, Pamela W., Bonnie J. Dorr, John W. Benoit. 1993. "A First-Pass Approach for Evaluating Machine Translation Systems:" to appear in *Machine Translation.*
4. Nomura, Hirosato. 1992. *JEIDA Methodology and Criteria on Machine Translation Evaluation.* Japan Electronic Industry Development Association.
5. van Slype, Georges. 1979. "Critical Study of Methods for Evaluating the Quality of Machine Translation." Final Report to the Commission of the European Communities Directorate General Scientific and Technical Information and Information Management.
6. White, J.S., T.A. O'Connell, L. M. Carlson. 1993. "Evaluation of Machine Translation". *(Proceedings of the 1993 Human Language Technologies Conference.* Morgan Kaufmann.

FLUENCY - '93 v. '92 Output - 1993 ARPA MT Evaluation

	CA93.FA	CA93.HA	SY93.FR	MA93.FR	SP93.SP	PA93.XP	PA93.PE	MA93.SP	KY93.JP	LI93.HA	MA93.JP	5/93 Mean	STD DEV 5/9	VAR 5/93	AVG STD DEV
(Pass.Scores/5)/	0.580	0.838	0.540	0.833	0.634	0.370	0.712	0.709	0.401	0.859	0.849	0.666	0.169	0.029	0.023
Num of Fragments															
Var of 22 pts	0.017	0.008	0.010	0.010	0.012	0.011	0.011	0.012	0.018	0.005	0.008	0.011		FRATIO	
std dev of 22 pts	0.129	0.092	0.101	0.100	0.111	0.120	0.120	0.111	0.135	0.073	0.089	0.107		12.084	
Norm Time	0.000	0.625	0.000	1.000	0.000	0.084	0.691	1.000	0.000	0.395	1.000	0.436	0.420		
Var of 22 times	0.000	0.027	0.000	0.000	0.000	0.018	0.078	0.000	0.000	0.021	0.000	0.013			
std dev of 22 time	0.000	0.165	0.000	0.000	0.000	0.133	0.279	0.000	0.000	0.144	0.000	0.066			

	CA92.FA	CA92.HA	SY92.FR	MA92.FR	SP92.SP	PA92.XP	PA92.PE	MA92.SP	SY92.JP	LI92.HA	M92.JP	pretest mean	TD DEV prete	VAR pretest	AVG STD DEV
(Pass.Scores/5)/	0.511	0.819	0.466		0.557		0.679		0.239	0.790		0.580	0.187	0.026	0.055
Num of Fragments															
Var of 6 pts	0.035	0.015	0.023		0.019		0.023		0.002	0.022		0.020		F RATIO	
std dev of 6 pts	0.187	0.123	0.152		0.136		0.151		0.039	0.148		0.134		3.158	
Norm Time	0.000	0.688	0.000	1.000	0.000		1.403	1.000	0.000	0.721	1.000	0.581	0.512		
Var of 6 times	0.000	0.010	0.000	0.000	0.000		0.762	0.000	0.000	0.025	0.000	0.080			
std dev of 6 times	0.000	0.099	0.000	0.000	0.000		0.873	0.000	0.000	0.157	0.000	0.113			

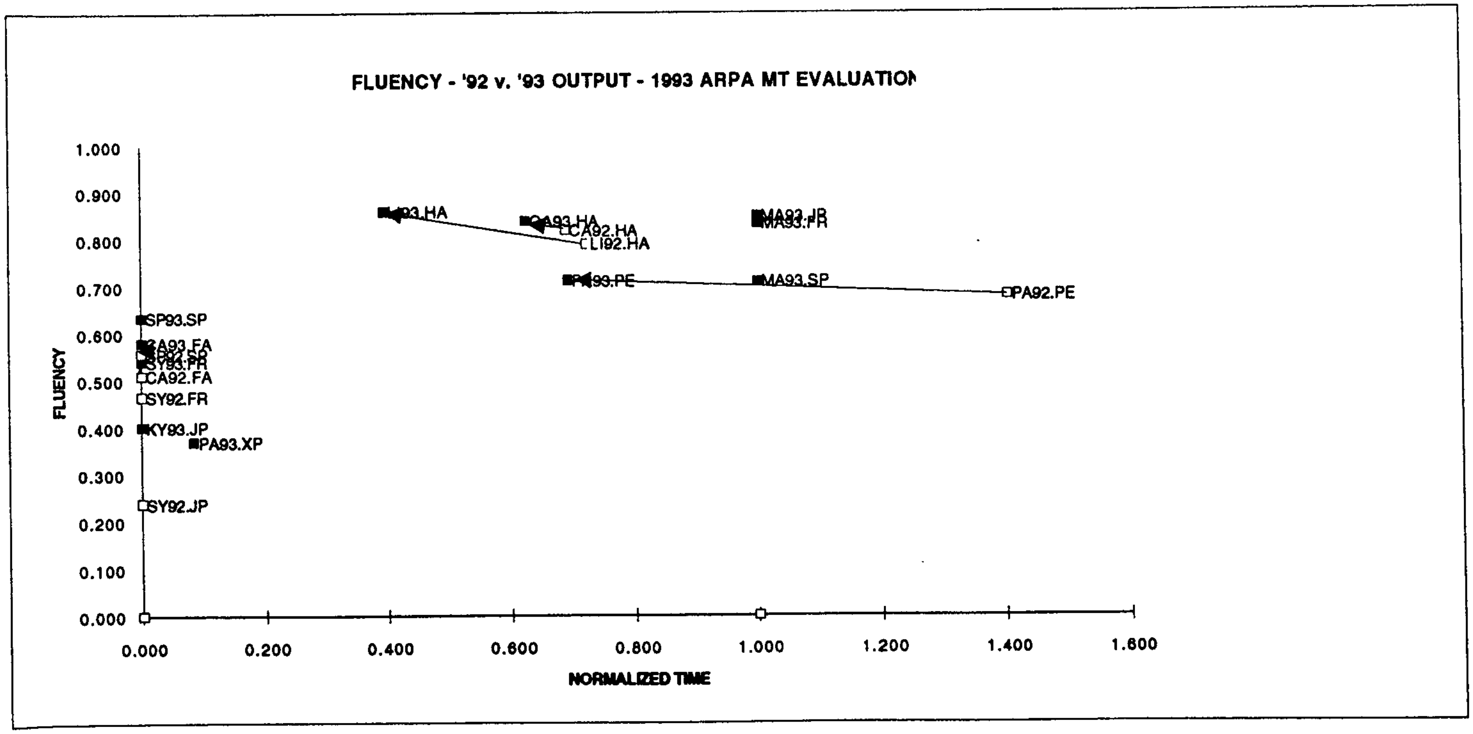

Figure 1.

ADEQUACY - '92 v. '93 OUTPUT - 1993 ARPA MT EVALUATION

	CA93.FA	CA93.HA	SY93.FR	MA93.FR	SP93.SP	PA93.XP	PA93.PE	MA93.SP	KY93.JP	LI93.HA	MA93.JP	5/93 Mean	STD DEV 5/9	VAR 5/93	AVG STD DEV
(Pass.Scores/5)/ Num of Fragments	0.670	0.850	0.743	0.840	0.790	0.548	0.801	0.766	0.363	0.707	0.651	0.703	0.137	0.019	0.024
Var of 22 pts	0.010	0.012	0.019	0.007	0.007	0.020	0.010	0.011	0.015	0.013	0.021	0.013		FRATIO	
std dev of 22 pts	0.102	0.109	0.136	0.083	0.086	0.140	0.098	0.103	0.123	0.114	0.144	0.113		6.696	
Norm Time	0.000	0.625	0.000	1.000	0.000	0.084	0.691	1.000	0.000	0.395	1.000	0.436	0.420		
Var of 22 times	0.000	0.027	0.000	0.000	0.000	0.018	0.078	0.000	0.000	0.021	0.000	0.013			
std dev of 22 time:	0.000	0.165	0.000	0.000	0.000	0.133	0.279	0.000	0.000	0.144	0.000	0.066			

	CA92.FA	CA92.HA	SY92.FR	MA92.FR	SP92.SP	PA92.XP	PA92.PE	MA92.SP	SY92.JP	LI92.HA	MA92.JP	pretest mean	TD DEV prete	VAR pretest	AVG STD DEV
(Pass.Scores/5)/ Num of Fragments	0.575	0.837	0.686		0.674		0.748		0.260	0.671		0.636	0.170	0.006	0.027
Var of 6 pts	0.008	0.009	0.001		0.006		0.001		0.001	0.009		0.005		F RATIO	
std dev of 6 pts	0.091	0.094	0.039		0.078		0.034		0.038	0.096		0.067		2.753	
Norm Time	0.000	0.688	0.000	1.000	0.000		1.403	1.000	0.000	0.721	1.000	0.581	0.512		
Var of 6 times	0.000	0.010	0.000	0.000	0.000		0.762	0.000	0.000	0.025	0.000	0.080			
std dev of 6 times	0.000	0.099	0.000	0.000	0.000		0.873	0.000	0.000	0.157	0.000	0.113			

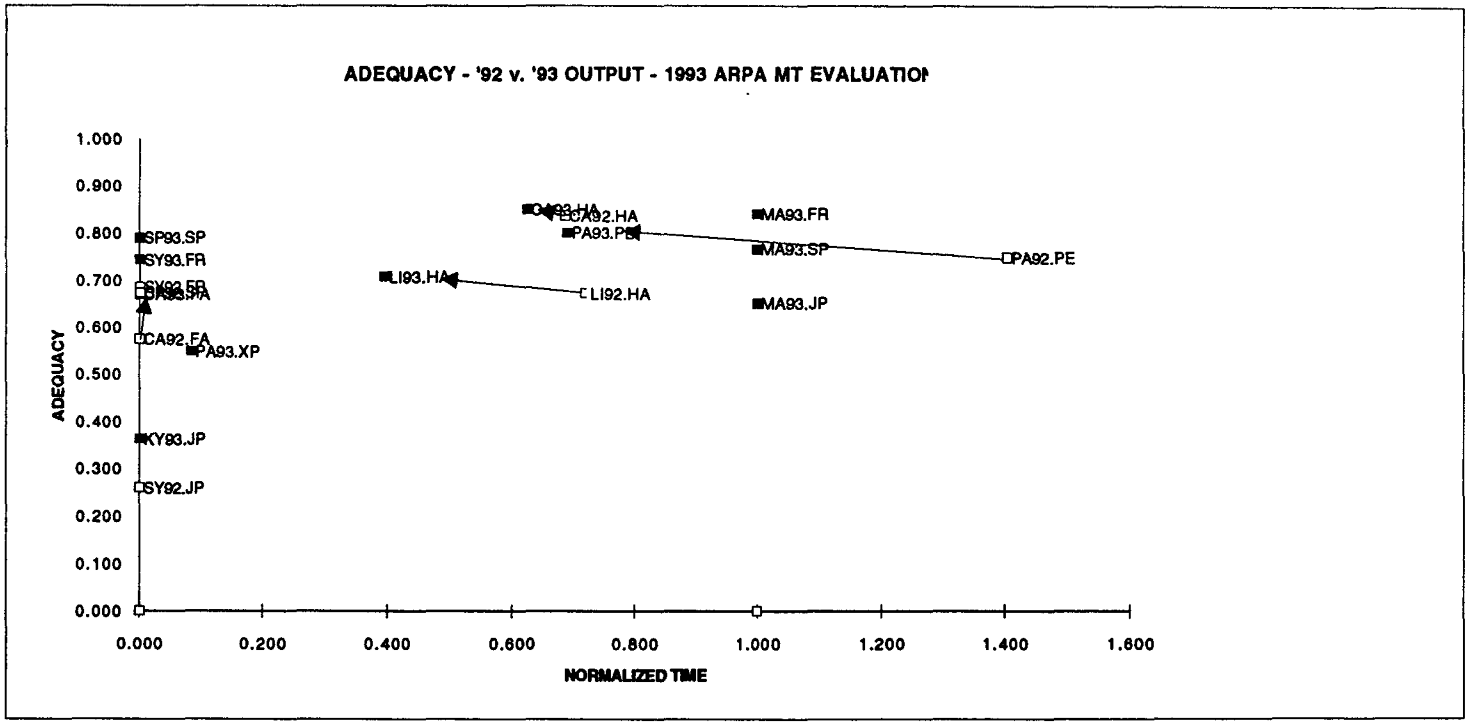

Figure 2.