The ARPA MT Evaluation Methodologies: Evolution, Lessons, and Future Approaches

John S. White
Theresa O'Connell
Francis O'Mara
PRC Inc.
1500 Planning Research Drive
McLean, VA 22102
tel: 703-556-1899
email: {white_john, oconnell_teri, omara_francis}@po.gis.prc.com

Abstract

The ARPA MT Evaluation methodology effort is intended to provide a basis for measuring and thereby facilitating the progress of MT systems of the ARPA-sponsored research program. The evaluation methodologies have the further goal of being useful for identifying the context of that progress among developed, production MT systems in use today. Since 1991, the evaluations have evolved as we have discovered more about what properties are valuable to measure, what properties are not, and what elements of the tests/evaluations can be adjusted to enhance significance of the results while still remaining relatively portable. This paper describes this evolutionary process, along with measurements of the most recent MT evaluation (January 1994) and the current evaluation process now underway.

1 Background

The ARPA MT Initiative is part of the Human Language Technologies Program of the Advanced Research Projects Agency Software and Intelligent Systems Technology Office. The mission of the ARPA MT effort is to make revolutionary advances in MT technology (White *et al.* 1993, 1994a, 1994b). As with all the initiatives under the Human Language Technologies Program, evaluation plays a fundamental role in achieving the success of this mission.

Since the inception of the evaluation process in 1991, there have been certain challenges to meaningful evaluation. Principally, judgments as to the correctness of a translation are highly subjective, even among expert human translations and translators. Thus evaluation must exploit intuitive judgments while constraining subjectivity in ways that minimize idiosyncratic sources of variance in the measurement. In addition, the ARPA situation is probably unique in that the three research systems employ radically different theoretical approaches to their core translation engines, originated envisioning quite different end-user applications, and have, up to now, translated different

languages. Thus the ability to compare the progress of approaches is difficult. In this paper we explore the evaluation methods intended to address the mission and associated challenges, while at the same time enabling an optimum level of portability of the methodology to a wider range of MT systems.

1.1 The research systems

There are three research projects under the ARPA MT Initiative:

- CANDIDE from IBM Thomas Watson Research Laboratory (Brown *et al.* 1993). CANDIDE uses a statistics-based, language modeling MT technique. It translates French to English(FE), with a Spanish-English (SE) system planned. Its originally envisioned application use is as a batch-oriented, non-interactive, fully automatic translation engine. Any user tools associated with CANDIDE were to be considered peripheral and outside the operation of the engine itself.
- PANGLOSS from the Carnegie Mellon University, Center for Machine Translation; New Mexico State University, Computing Research Laboratory; and University of Southern California, Information Sciences Institute (Frederking *et al.* 1993). The PANGLOSS system uses both knowledge-based and linguistic techniques, integrating new approaches to lexical acquisition and target generation. It translates Spanish into English and has a Japanese-English (JE) system for the August evaluation. Its originally envisioned application use involved user interaction in an integrated fashion, specifically as a means of assisting and teaching the translation engine for lexical and structural disambiguation, unknown patterns and lexicon, etc.
- LINGSTAT from Dragon Systems(Yamron *et al.* 1994). LINGSTAT uses a combination of statistical and linguistic techniques. It translates Japanese to English and also has a Spanish-to-English system . Its originally envisioned application use was in a desktop, monolingual environment, allowing a user to write a translation from cues provided by the translation engine.

1.2 The original evaluation concepts

In 1991, representatives of PRC, the U.S. Government, and the MT research projects developed a variety of evaluation methods and goals, originally intended to focus on the anticipated strengths of particular theoretical and end-use approaches. It was believed at the time, for example, that PANGLOSS would be more accurate at the outset than others (being human-assisted), and thus its evaluation path might track its progression to more complete automation. At the same time, an evaluation of fundamentally automatic CANDIDE might track its progression to higher accuracy. To address these concerns, multiple evaluation methods were adopted to cover the apparent strengths of each system type while affording some ability to measure the

progress of each in isolation and compared with each other. At the time of the first evaluation in 1992, two evaluations measured the comprehensibility and quality of both automatic and human-assisted outputs.

2 Evolution of the Evaluation

As a result of the first evaluation, the initiative learned a number of lessons about the feasibility/desirability of certain measures, and modified subsequent evaluations accordingly (first verifying the new methods against outputs used in previous evaluations). The methods used in 1992 were either modified dramatically or abandoned.

2.1 Direct Comparability

The 1992 comprehension evaluation was intended to derive direct comparisons of the tested systems, even though they translate different languages. To accommodate this, English newspaper articles about financial mergers and acquisitions were professionally translated into the respective source languages, and then submitted to the MT systems and control processes for translation back into English. The MT outputs, the controls, and the original English were then presented to monolingual native speakers of English in the form of an "SAT"-like comprehension test (cf. Church and Hovy 1991), in which they answered multiple choice questions about the content of the articles.

Since the original source of all the articles was English, it was believed at the time that we could compare the comprehension results of systems that translated a foreign language version of these back into English. However, it is evident that any human manipulation, even professional translation, has too great a potential of modifying the content of a text, and thus that there is no way to tell whether a particular result reflects the performance of a system or the competence of the original translation from English. Consequently we abandoned that aspect of the comprehension evaluation. However, we have preserved this evaluation component as a valuable measure of the informativeness preserved by an MT system as it translates an original foreign language text into English.

2.2 Quality Panel

The 1992 evaluation used measurement tools for judging the quality of translation, that is, its lexical, grammatical, semantic, and stylistic accuracy and fluency. This process involved subjecting the outputs and controls to a panel of professional, native-English-speaking translators of the relevant languages. These "quality panels" used a metric modeled on a standard US Government

metric for grading these outputs as if they were the work of a human translator, in order to determine the proficiency level the "translator".

The natural appeal of the quality panel is that the metric used to make the evaluation is externally motivated, i.e., was developed for a more general purpose of grading translators, and not specifically for the purpose of judging MT outputs. This prevents certain biases from being inadvertently introduced into the metric. However, we found that the quality panel concept was difficult to deploy logistically. (It is very hard to get a sufficient number of pair-specific translation experts committed for a week or more of such effort and most difficult for such a panel to come to a consensus.) This compromised the ultimate goal of portability of evaluation. Moreover, it was not possible to maintain the exact structure of the metric: the nature and proliferation of MT errors necessitated alterations of the grading limits in the original method, thus introducing the potential for the very bias that its use was intended to avoid. Consequently the quality panel evaluation was abandoned.

2.3 Adequacy and Fluency Measures

The three subsequent MT evaluations have used two evaluation methods that cover the relevant measurements of the quality panel without involving the large number of expert evaluators. In an adequacy evaluation, literate, monolingual English speakers make judgments determining the degree to which the information in a professional translation can be found in an MT (or control) output of the same text. The information units are "fragments", usually less than a sentence in length, delimited by syntactic constituent and containing sufficient information to permit the location of the same information in the MT output. These fragmentations are intended to avoid biasing results in favor of linguistic-compositional approaches (which may do relatively better on longer, clause-level strings) or statistical approaches (which may do better on shorter strings not associated with syntactic constituency).

In a fluency measure, the same evaluators are asked to determine, on a sentence-by-sentence basis, whether the translation reads like good English (without reference to the "correct" translation, and thus without knowing the accuracy of the content). Their task is to determine whether each sentence is well-formed and fluent in context.

The adequacy and fluency evaluations, along with the modified comprehension (or "informativeness") evaluation, have become the standard set of methodologies for the ARPA MT evaluation. A variety of other issues have been raised as a result of the three evaluations completed in 1994. Several of these have to do with human factors issues (minimizing fatigue and other bias sources for the evaluators), but a more fundamental issue concerns what types of outputs actually provide results that measure what progress the different MT approaches are making.

2.4 Human-Assisted Measurements

As described above, the initial positions at the beginning of the program held that there was a need to measure a machine translation approach in terms of the productivity it potentially affords to a user of machine translation. A human-assisted MT (HAMT) process operated by a novice translator should be faster, if not better than, a manual translation produced by the same person. For the first three evaluations, then, the research systems submitted human-assisted output as well as fully-automatic output (in most cases). The HAMT processes included post-editing of automatic MT, query-based interaction with a user during the translation, or actual composition of the translation by the user with supports from the MT system. These outputs were submitted to the same evaluations as the fully-automatic (FAMT) outputs, along with manual translations produced by the same novice translators who operated the HAMT systems. In principle, the results should reflect the degree to which HAMT was faster (and possibly better) than manual, and whether that difference was greater than it was in preceding evaluations.

The measurement of HAMT appeared valuable, especially since no MT system currently under development is likely to be used without a significant human-assisted component. However, it, like the "back translation" problem in the original comprehension measure, introduced effects unrelated to the performance of a particular translation approach:

- The certification of a person as a "novice" translator allowed too much variability in translation skill level, and in capabilities unrelated to translation. If the translator is better than novice level, the time improvement artificially decreases; time improvement artificially increases where the translator has a facile command of the specific HAMT tools and workstation environment Thus the HAMT measurements, in reality, reflected the expertise of the translator and the sophistication of the human-computer interface. They did not measure the core technology.
- The HAMT systems generally performed better in fluency, adequacy, and comprehension than the FAMT outputs, as expected. However, there was no apparent means of extracting the contribution of the peripheral tools. Indeed what constitutes peripheral tools, versus designed human interaction as part of the essential translation process, is certainly unclear in some of the cases.
- Given that all production MT systems in use today employ some human-assisted tools, the evaluation was faced with the difficulty of determining whether the production systems should have also been asked to submit HAMT as well as FAMT outputs. This prospect served to magnify the issue of inability to control the effect of the non-core translation technology from the core technology.

Because of all these reasons, the evaluation effort ultimately took the position that advancing the core technology of machine translation could best be served by eliminating, as much as possible, everything that could mask the function of the core translation approach — the engine that renders a string in one language into a string in another. Thus, we came to believe that these core technologies should be evaluated from FAMT outputs, avoiding the extraneous effects of HAMT tools. Given this position, the current evaluation does not use HAMT output.

3 Involvement of Production Systems

As alluded to above, the participation of non-research, or "production" systems (i.e., either commercial or institutional MT systems which are used on a regular basis) has been a part of the evaluation process from the beginning. In the January 1994 evaluation, the invitation to production systems was broadened significantly, and 13 production systems participated. The value of this participation cannot be understated; understanding the current position of the research systems among the production systems preserves a grounding in the state of the art and practice of MT technology. Beyond this, their participation helps the evaluation methodologies improve as well, particularly in the direction of enhanced portability.

4 The January 1994 ARPA MT Evaluation

Each ARPA MT Evaluation has two parts: a test and an evaluation. The test consists of translations performed by MT systems. The evaluation comprises a collection of human judgments on the quality of those translations and analyses of those human judgments. The first of two evaluations scheduled for 1994 started in January.

4.1 January 1994 Test Process

The January 1994 test differed from its predecessors principally in the proportion of financial to general news texts in the test sets and in the increased number of production systems participating.

There were 20 passages from each source language. As in 1993, their length remained at 300 to 500 words for French and Spanish and an average of 800 characters for Japanese. The proportion of financial M&A texts was reduced to 50% of the test set; the remaining ten texts in each set were general news articles. This reduction continued a trend of moving away from the financial domain as the research systems mature and become able to process a wider variety of lexical items.

The MT systems and human translators produced a combined output of 500 translations: ten FE versions, nine SE versions and six JE versions. Each of the three ARPA research systems produced FAMT: CANDIDE (FE), LINGSTAT (JE), and PANGLOSS (SE). In addition, there were 13 production systems: GLOBALINK, POWER TRANSLATOR (FE, SE); LINGUISTIC PRODUCTS, PC TRANSLATOR (SE); MICROTAC, FRENCH/SPANISH ASSISTANT (FE, SE); NEC, PIVOT (JE); PAHO, SPANAM (SE); SIETEC, METAL (FE); SOCATRA, XLT (FE); SYSTRAN TRANSLATION SYSTEMS, INC. (FE, JE, SE); and WINGER A/S, WINGER (FE). The larger number of production systems provided a wider base of comparison against the performance of the research systems.

Once again, the research sites also supplied level two manual and human-assisted translations of the test sets. At CANDIDE and LINGSTAT, two level two translators each performed manual translation of ten sequential texts and produced HAMT of the other half of the test set. At PANGLOSS, these tasks were divided among four translators.

Systems were required to freeze development upon commencing the test. In the production of FAMT, all fully-automatic functions were permitted; human intervention and lexical development were expressly prohibited.

4.2 January 1994 Evaluation Process

All translations were randomly distributed into a matrix which governed the assembly of 30 evaluation books. In prior evaluations, matrices had been ordered according to a Latin square. The purpose of automatic random distribution was to minimize context effects whereby marks for a translation would be influenced by the difficulty level of the text immediately preceding it in the evaluation book. Within the matrix, each translation appeared one time and no book contained more than one translation of any source passage. The proportion of half general to half financial texts was preserved in each evaluation book. Every book contained translations from all three source languages.

In January 1994 there were three components to the evaluation suite. They appeared in all evaluation books in the following sequence: comprehension, fluency and adequacy. The same matrix was used for each component so that each evaluator saw the same texts in the same order in each component.

Professional expert translations were used as reference versions in the adequacy component and were evaluated in the comprehension component.

Thirty evaluators each completed one evaluation book. Within this book evaluators judged the same 16 passages for comprehension, fluency and adequacy. In the comprehension section, they evaluated an additional two expert translations. Evaluators made an average of 769 judgments during the course of the one day evaluation.

Evaluators received both written and spoken instructions. The goal was to instruct each evaluator to the same level of understanding. Written instructions included examples in degrading order. In the fluency and adequacy sections an unidentified practice text preceded the evaluation texts. The purpose of the practice texts was to increase consistency of judgment by providing an opportunity to learn the fluency and adequacy tasks.

Evaluators were highly verbal persons ranging in age from high school seniors to persons in their sixties. To minimize fatigue, evaluators were allowed to take breaks whenever they desired. Prior evaluations had shown that the incidence of omission errors was highest among persons in their teens and early twenties. Therefore, evaluators in this age group were required to take a ten minute break in the middle of each task. This planned break virtually eliminated omission errors .

4.3 January 1994 Results

Figures 1, 2, and 3 contain the FAMT results by each of the three measures.

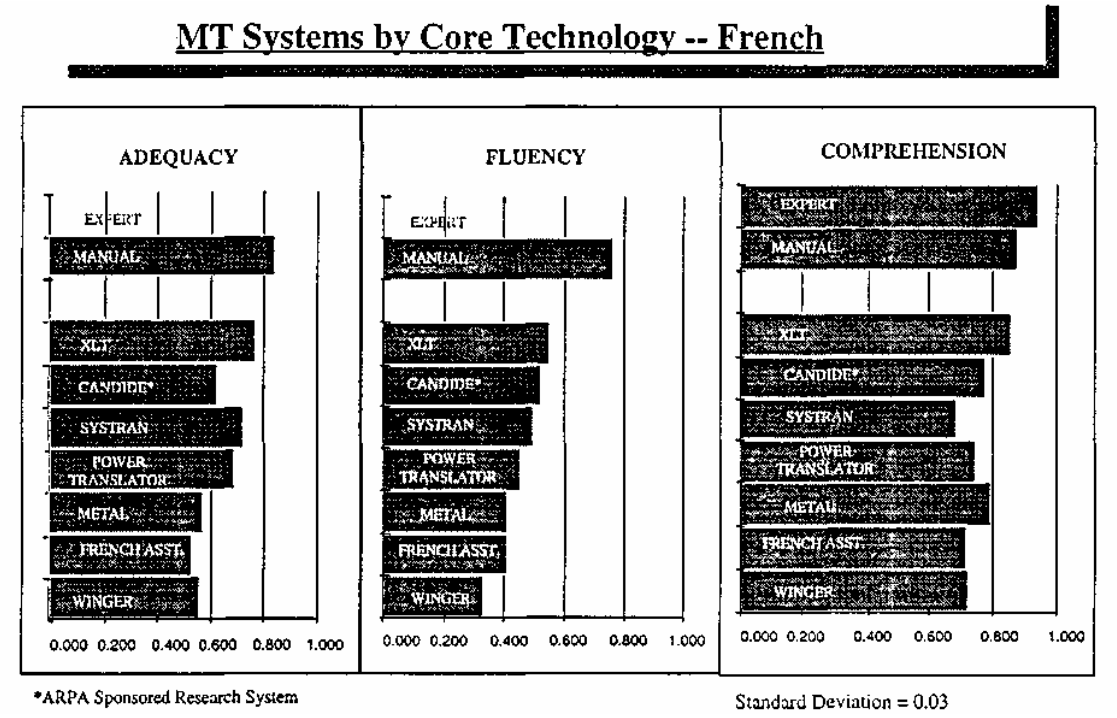

Figure 1. French FAMT results

In total, 23,060 data points were tallied: 3,000 in comprehension; 6,744 in fluency; and 13,316 in adequacy.

4.3.1 January 1994 FAMT

For FAMT, passage scores for all three components were plotted between 0 and 1 according to the following formulas (see Figures *1,2,3):*

Comprehension(P) = #Correct/6
Fluency(P) = _((Judgment point - l)/(5-l))/#Sentences in passage
Adequacy(P) = _((Judgment point - l)/(5-l))/#Fragments in passage

For passage scores, the mean and standard deviation over 20 passages were calculated. The standard deviation was also calculated for each system's scores. The F-Ratio, used as a measure of sensitivity, was calculated as the variance of the system means over the mean of system variances.

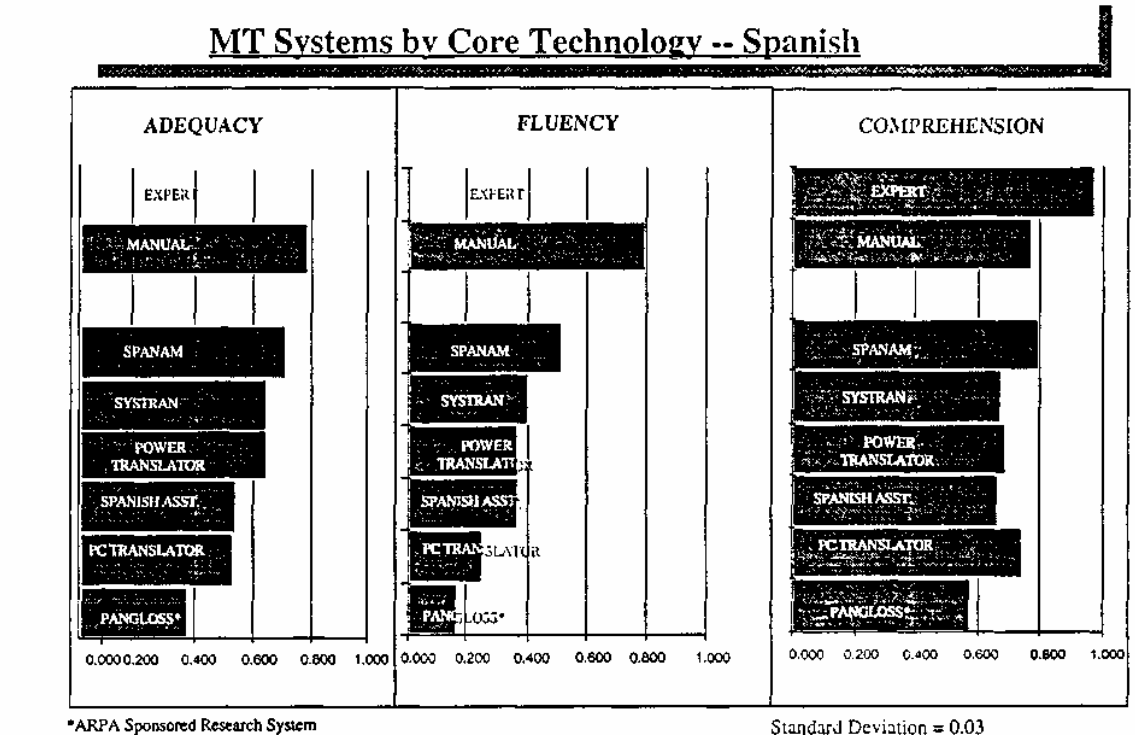

Figure 2. Spanish FAMT result

FE comprehension scores ranged from .851 for XLT to .684 for SYSTRAN. CANDIDE scored .781. FE fluency scores ranged from .554 for XLT to.339 for WINGER. CANDIDE scored .524. FE adequacy scores ranged from .786 for XLT to .548 for French Assistant. CANDIDE scored .638.

JE comprehension scores ranged from .509 for LINGSTAT to.386 for SYSTRAN. JE fluency scores ranged from .211 for NEC to .114for LINGSTAT. JE

adequacy scores ranged from .223 for NEC and SYSTRAN to .147 for LINGSTAT.

SE comprehension scores ranged from .798 for SPANAM to .570 for PANGLOSS. SE fluency scores ranged from .520 for SPANAM to.176 for PANGLOSS. SE adequacy scores ranged from .719 for SPANAM to .377 for PANGLOSS.

Research system scores were compared to 1993 scores. An apparent PANGLOSS decline was an artifact of the way tests were constructed in 1993. The production systems which participated in 1993 and 1994, namely SYSTRAN French and SPANAM Spanish, also improved in three evaluation measures. Both SYSTRAN and SPANAM had undergone significant system enhancements in the 1993-94 interim, and the improvements reflected in the January evaluation.

The F-Ratio across all research system outputs for comprehension rose from .375 to .756 demonstrating increased sensitivity in this component in January 1994 over 1993. The total F-Ratio for fluency dropped from 12.584 to 7.466. The total F-Ratio for adequacy evaluations rose from 3.664 to .4.892. The pooled standard deviation for all results was .03.

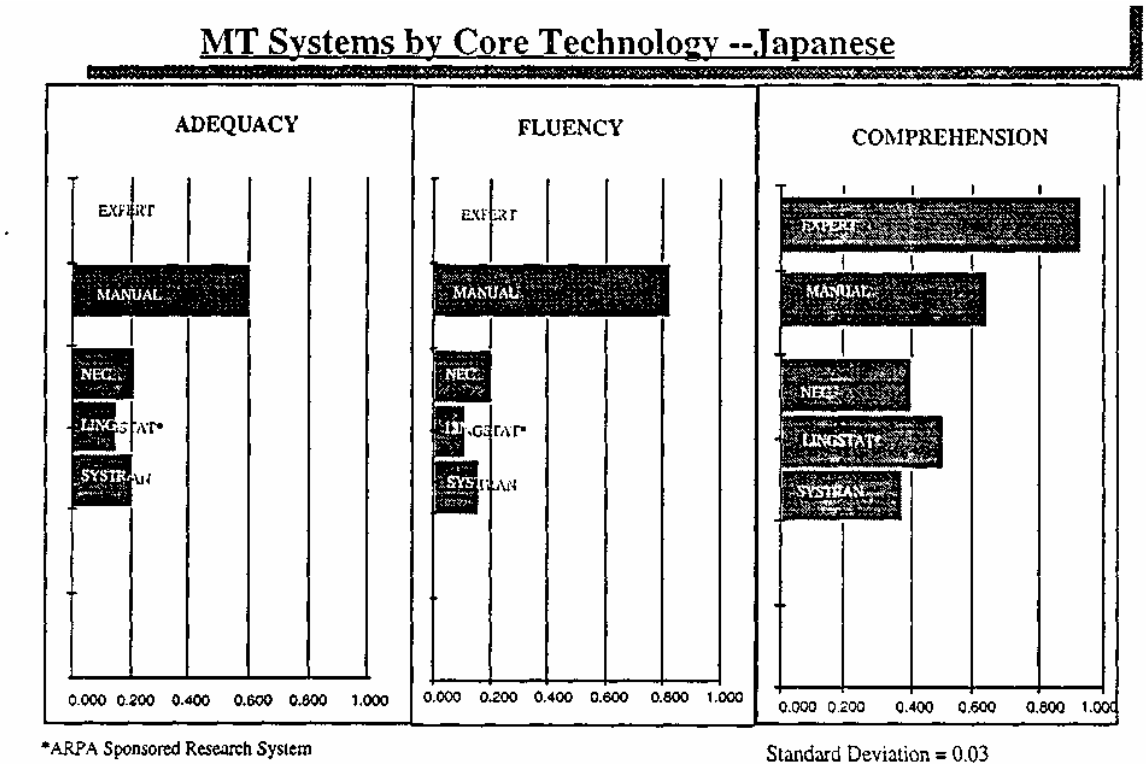

Figure 3. Japanese FAMT results

Overall, the results shown in Figs. 1-3 illustrate that the variety of approaches represented in the groups have merit, particularly when matured through end-use experience. In addition, systems that translate multiple language pairs seem to be ordered similarly across the languages, which may reflect a consistency of development approach.

4.4 January 1994 Analyses of Variance, Findings

To maximize the sensitivity of the analysis to possible system differences, it was important to minimize the error (noise) variance that was due to the effect of four methodological factors: evaluator differences, text differences; domain differences; and text order in the evaluation books. We performed analyses of variance (ANOVAs) to determine if these existed in the data thereby clouding the distinctions between systems. To determine the extent to which this was so, the ANOVAs estimated the nature and magnitude of effects of the four methodological factors. The ANOVAs identified significant effects of all four factors with the magnitude of the effects varying somewhat across subgroups of evaluators and domains.

These factors are being controlled in the design features of the subsequent evaluation. While the ANOVAs served to verify the significance of the January results, they also showed us ways to enhance the sensitivity of the system differences so that more subtle differences and trends could be identified.

5 The August 1994 ARPA MT Evaluation

Like its predecessors, the August 1994 ARPA MT Evaluation measured the progress of the ARPA research systems by contrasting present performance to past performance and to the performance of production systems. Its scope expanded to increase the validity of its results. Its design evolved to reflect lessons learned from prior evaluations.

5.1 August Test

The August test differed from its predecessors in: output from the research sites and expert translators; the exclusion of HAMT and level two manual translations; and the number and domain of test texts.

Systems invited to participate in the August evaluation included top performers among January 1994 production sites. Each system translated the same source languages into English as they had in the January 1994 Evaluation. In August, all three research systems produced FAMT. Two of the research systems added a second source language. For the first time, LINGSTAT produced SE FAMT and PANGLOSS produced JE FAMT.

In contrast to all earlier evaluations, for reasons discussed above in Section 2.4, the August Evaluation did not use HAMT or level two manual translations. Two sets of level five expert translations were procured. Half of each expert translation set served as reference versions in the adequacy Evaluation; the other expert translations were evaluated in each of the three components. This was the first time expert translations were evaluated for fluency and adequacy.

To minimize the effect of passage variances, the number of source texts increased to 100 from each language. These were all general news stories; while it was probable that some financial texts would appear in the test set, there was no directed retrieval of M&A texts.

5.2 August Evaluation Process

There were three components to the August evaluation suite, informativeness, fluency and adequacy. In contrast to earlier evaluations where the components appeared in the same order in all evaluation books, the sequence of the components varied from book to book.

Matrix design was improved to minimize context and halo effects. Distribution of system output was pre-coordinated rather than random. Each book contained output from each system. However, to reduce context effects, output from every system preceded output from every other system at least one time within the set of matrices. In prior evaluation books, evaluators saw the same text three times, once in each section. In the current evaluation, evaluators saw different texts in each component to prevent halo effects, whereby an evaluator would remember that a text had earned a certain range of marks in an earlier component and demonstrate bias toward assigning similar marks when the translation appeared again. As in prior matrices, only one translation of any source text appeared in any book.

The random order of evaluation components in the books will provide insight on whether the placement of a component in the book affects evaluator performance. Such effects could be caused by fatigue. All evaluators took planned breaks to minimize omission errors.

One practice text will precede the evaluation texts in all three components. While the primary goal of these texts is to train the evaluators, they were selected to also serve two other purposes. To better monitor progress of the research systems and to contrast performance of current evaluators to performance of past evaluators, a set of 100 practice texts was assembled from translations used in January 1994. These texts included output from all three January 1994 research systems and some of the production systems. Their January 1994 scores ranged from high to low. In 1995, each text in this practice set will be re-evaluated for informativeness, fluency and adequacy.

The number of judgments tallied for current output from each system increased by a factor of five over January 1994. For each system, the number of

judgments for informativeness increased from 120 to 600; for fluency from approximately 347 to 1735, and for adequacy from 605 to 3025.

The evolutionary modifications that have led to the shape of the August 1994 evaluation should provide a more sensitive measure of results, while being optimally comparable to the previous results in order to show progress.

6 Conclusion

The ARPA MT Test and Evaluation process bears an integral responsibility to the overall mission of developing and promoting revolutionary advances in machine translation technology. The series of evaluations to date have demonstrated not only the positive progress of the research systems, but also the place of this progress in the context of the progress of the industry as a whole. At the same time, the experiences to date have refined the evaluation methodologies to be more focused on the core technologies, more accurate in controlling for human factor sources of variance, and more sensitive in measuring the results of the ARPA MT tests.

References

Brown, P.F.,S.A. Della Pietra, V.J. Della Pietra, and R.L. Mercer. 1993. "The Mathematics of Statistical Machine Translation: Parameter Estimation". Computational Linguistics vol. 19:2. pp. 263-312

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.

Frederking, R., D. Grannes, P.Cousseau and S. Nirenburg. 1993. "An MAT Tool and Its Effectiveness". Human Language Technology. March 1993, pp. 196-201.

White, J.S., and T.A. O'Connell. 1994a. "Evaluation in the ARPA MT Program: 1993 Methodology". Proceedings of the ARPA Workshop on Human Language Technology Workshop. Plainsboro NJ: March 1994.

White, J.S. and T.A. O'Connell. 1994b. "Evaluation Methodologies in the ARPA Machine Translation Initiative". Proceedings of the Symposium on Advanced Information Processing and Analysis Steering Group. Tysons Corner, VA: March 1994. p.91.

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.

Yamron, J., J. Cant, A. Demedts, T. Dietzel, Y. Ito. 1994. 'The Automatic Component of the LINGSTAT Machine-Aided Translation System". Proceedings of the ARPA Workshop on Human Language Technology Workshop. Plainsboro NJ: March 1994.