EVALUATION OF MACHINE TRANSLATION

John S. White, Theresa A. O'Connell

PRC Inc.
McLean, VA 22102

and

Lynn M. Carlson

DoD

ABSTRACT

This paper reports results of the 1992 Evaluation of machine translation (MT) systems in the DARPA MT initiative and results of a Pre-test to the 1993 Evaluation. The DARPA initiative is unique in that the evaluated systems differ radically in languages translated, theoretical approach to system design, and intended end-user application. In the 1992 suite, a Comprehension Test compared the accuracy and interpretability of system and control outputs; a Quality Panel for each language pair judged the fidelity of translations from each source version. The 1993 suite evaluated adequacy and fluency and investigated three scoring methods.

1. INTRODUCTION

Despite the long history of machine translation projects, and the well-known effects that evaluations such as the ALPAC Report (Pierce et al., 1966) have had on that history, optimal MT evaluation methodologies remain elusive. This is perhaps due in part to the subjectivity inherent in judging the quality of any translation output (human or machine). The difficulty also lies in the heterogeneity of MT language pairs, computational approaches, and intended end-use.

The DARPA machine translation initiative is faced with all of these issues in evaluation, and so requires a suite of evaluation methodologies which minimize subjectivity and transcend the heterogeneity problems. At the same time, the initiative seeks to formulate this suite in such a way that it is economical to administer and portable to other MT development initiatives. This paper describes an evaluation of three research MT systems along with benchmark human and external MT outputs. Two sets of evaluations were performed, one using a relatively complex suite of methodologies, and the other using a simpler set on the same data. The test procedure is described, along with a comparison of the results of the different methodologies.

The authors would like to express their gratitude to Michael Naber for his assistance in compiling, expressing and interpreting data.

2. SYSTEMS

In a test conducted in July, 1992, three DARPA-sponsored research systems were evaluated in comparison with each other, with external MT systems, and with human-only translations. Each system translated 12 common Master Passages and six unique Original Passages, retrieved from commercial databases in the domain of business mergers and acquisitions. Master Passages were Wall Street Journal articles, translated into French, Spanish and Japanese for cross-comparison among the MT systems and languages. Original Passages were retrieved in French, Spanish, and Japanese, for translation into English.

The 1992 Evaluation tested three research MT systems:

CANDIDE (IBM, French - English) uses a statistical language modeling technique based on speech recognition algorithms (see Brown et al., 1990). It employs alignments generated between French strings and English strings by training on a very large corpus of Canadian parliamentary proceedings represented in parallel French and English. The CANDIDE system was tested in both Fully Automatic (FAMT) and Human-assisted (HAMT) modes.

PANGLOSS (Carnegie Mellon University, New Mexico State University, and University of Southern California) uses lexical, syntactic, semantic, and knowledge-based techniques for analysis and generation (Nirenburg, et al. 1991). The Spanish-English system is essentially an "interlingua" type. Pangloss operates in human-assisted mode, with system-initiated interactions with the user for disambiguation during the MT process.

LINGSTAT (Dragon Systems Inc.) is a computer-aided translation environment in which a knowledgeable non-expert can compose English translations of Japanese by using a variety of contextual cues with word parsing and character interpretation aids (Bamberg 1992).

Three organizations external to the DARPA initiative provided benchmark output. These systems ran all the test input that was submitted to the research systems. While these systems are not all at the same state of commercial robustness, they nevertheless provided external perspective on the state of FAMT outside the DARPA initiative.

The Pan American Health Organization provided output from the SPANAM Spanish-English system, a production system used daily by the organization.

SYSTRAN Translation Systems Inc. provided output from a French - English production system and a Spanish - English pilot prototype.

The Foreign Broadcast Information Service provided output from a Japanese-English SYSTRAN system. Though it is used operationally, SYSTRAN Japanese-English is not trained for the test domain.

3. MT EVALUATION METHODOLOGIES

The 1992 Evaluation introduced two methods to meet the challenge of developing a black-box evaluation that would minimize judgment subjectivity while allowing a measure of comparison among three disparate systems. A Comprehension Test measured the adequacy or intelligibility of translated outputs, while a Quality Panel was established to measure translation fidelity.

The 1992 Evaluation provided meaningful measures of performance and progress of the research systems, while providing quantitative measures of comparability of diverse systems. By these measures, the methodologies served their purpose. However, developing and evaluating materials was difficult and labor-intensive, involving special personnel categories.

In order to assess whether alternative metrics could provide comparable or better evaluation results at reduced costs, a Pre-test to the 1993 Evaluation was conducted. The Pre-test was also divided into two parts: an evaluation of adequacy according to a methodology suggested by Tom Crystal of DARPA; and an evaluation of fluency. The new methodologies were applied to the 1992 MT test output to compare translations of a small number of Original Passages by the DARPA and benchmark systems against human-alone translations produced by human translators. These persons were nonprofessional level 2 translators as defined by the Interagency Language Roundtable and adopted government-wide by the Office of Personnel Management in 1985.

In the second suite, three numerical scoring scales were investigated: yes/no, 1-3 and 1-5. Two determinations arise from the comparison: whether the new methodology is in fact better in terms of cost, sensitivity (how accurately the variation between systems is represented) and portability, and which scoring variant of the evaluation is the best by the same terms.

The methodologies used in the 1992 Evaluation and 1993 Pre-test are described briefly below.

3.1. Comprehension Test Methodology

In the 1992 Evaluation, a set of Master Passage versions formed the basis of a multiple-choice Comprehension Test, similar to the comprehension section of the verbal Scholastic Aptitude Test (SAT). These versions consisted of the "master passages" originally in English, professionally translated into the test source languages, and translated back into English by the systems, benchmarks and human translators.

Twelve test takers unfamiliar with the source languages answered the same multiple choice questions over different translation versions of the passages. They each read the same 12 passages, but rendered variously into the 12 outputs represented in the test (CANDIDE FAMT, CANDIDE HAMT, PANGLOSS HAMT, LINGSTAT HAMT, SPANAM FAMT, SYSTRAN FAMT for all three language pairs, human-only for all three pairs, and the Master Passages themselves.) The passages were ordered so that no person saw any passage, nor any output version twice.

3.2. Quality Panel Methodology

In the second part of the 1992 Evaluation, for each source language, a Quality Panel of three professional translators assigned numerical scores rating the fidelity of translated versions of six Original and six Master Passages against sources or back-translations. Within a given version of a passage, sentences were judged for syntactic, lexical, stylistic and orthographic errors.

3.3. Pre-test Adequacy Methodology

As part of the 1993 Pre-test, nine monolinguals judged the extent to which the semantic content of six baseline texts from each source language was present in translations produced for the 1992 Evaluation by the test systems and the benchmark systems. The 1992 Evaluation level 2 translations were used as baselines. In the 18 baselines, scorable units were bracketed fragments that corresponded to a variety of grammatical constituents. Each monolingual saw 16 machine or human-assisted translations. Each evaluator saw two passages from each system. The passages were ordered so that no person saw the same passage twice.

3.4. Pre-test Fluency Methodology

In Part Two of the Pre-test, the nine monolinguals evaluated the fluency (well-formedness) of each sentence in the same distribution of the same 16 versions that they had seen in Part One. In Part Two, these sentences appeared in paragraph form, without brackets.

4. RESULTS

In both the 1992 Evaluation and the 1993 Pre-test, the quality of output and time taken to produce that output were compared across:

- human-alone translations
- output from benchmark MT systems
- output from the research systems in FAMT and/or HAMT modes.

The results of the Comprehension Test (in which all systems used what were originally the same passages) are similar to the results of the Quality Panel, with some minor exceptions (see White, 1992). Thus for the purpose of the discussion that follows, we compare the results of the second, adequacy-fluency suite against the comparable subset of the Quality Panel test from the first suite.

The Pre-test evaluation results are arrayed in a manner that emphasizes both the adequacy or fluency of the human-assisted and machine translations and the human effort involved to produce translations, expressed in (normalized) time. For each part of the Pre-test, scores were tabulated, entered into a spreadsheet table according to scoring method and relevant unit, and represented in two dimensional arrays. The relevant unit for Part 1 is the adequacy score for each fragment in each version evaluated. For Part 2, the relevant unit is the score for fluency of each sentence in each version evaluated.

Performance for each of the systems scored was computed by averaging the fragment (or sentence) score over all fragments (or sentences), passages, and test subjects. The method for normalizing these average scores was to divide them by the maximum score per fragment (or sentence); for example, 5 for the 1-5 tests. Thus, a perfect averaged normalized system score is 1, regardless of the test.

Three evaluators each saw two passages per system; thus there was a total of six normalized average scores per system. The mean for each system is based on the six scores for that system. The eight system means were used to calculate the global variance. The F-ratio was calculated by dividing the global variance, i.e. the variance of the mean per system, by the local variance, i.e. the mean variance of each system. The F-ratio is used as a measure of sensitivity.

The Quality Panel scores were arrayed in a like manner. The quality score per passage was divided by the number of sentences in that passage. The six Original Passages were each evaluated by 3 translators producing a total of 18 scores per system. Adding the 18 scores per system together and dividing by 18 produced the mean of the normalized quality score per system. The means, variances and F-ratios were calculated as described above for adequacy and fluency.

4.1. Quality Panel Evaluation Results

Figure 1 is a representation of the Quality Panel evaluation, from the first evaluation suite, using the comparable subset of the 1992 data (i.e., the original passages). The quality scores range from .570 for Candide HAMT to .100 for Systran Japanese FAMT. The scores for time in HAMT mode, represented as the ratio of HAMT time to Human-Only translation time, range from .689 for Candide HAMT to 1.499 for Pangloss Spanish HAMT. The normalized time for FAMT systems is set at 0.

4.2. Adequacy Test Results

Figure 2 represents the results of the adequacy evaluation from the second suite. Using the 1-5 variation of the evaluation, the adequacy (vertical axis) scores range from .863 for Candide HAMT to .250 for Systran Japanese FAMT. The time axis reflects the same ratio as is indicated in Figure 1.

4.3. Fluency Test Results

Figure 3 represents the results of the fluency evaluation from the second suite. Using the 1-5 variant, fluency scores range from .853 for Candide HAMT to .214 for Systran Japanese FAMT. The time axis reflects the same ratio as is indicated in Figure 1.

5. COMPARISON OF METHODOLOGIES

The measures of adequacy and fluency used in the second suite are equated with the measure of quality used by the 1992 Evaluation Quality Panel. The methodologies were compared on the bases of sensitivity, efficiency, and expenditures of human time and effort involved in constructing, administering and performing the evaluation.

Cursory comparison of MT system performance in the three results shown in Figures 1 through 3 shows similarity in behavior. All three methodologies demonstrate higher adequacy, fluency and quality scores for

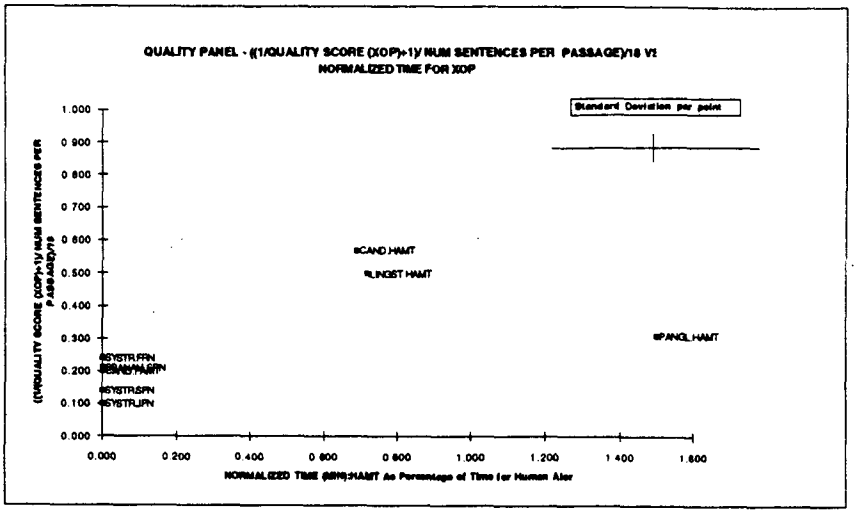

Figure 1: Quality Panel Results

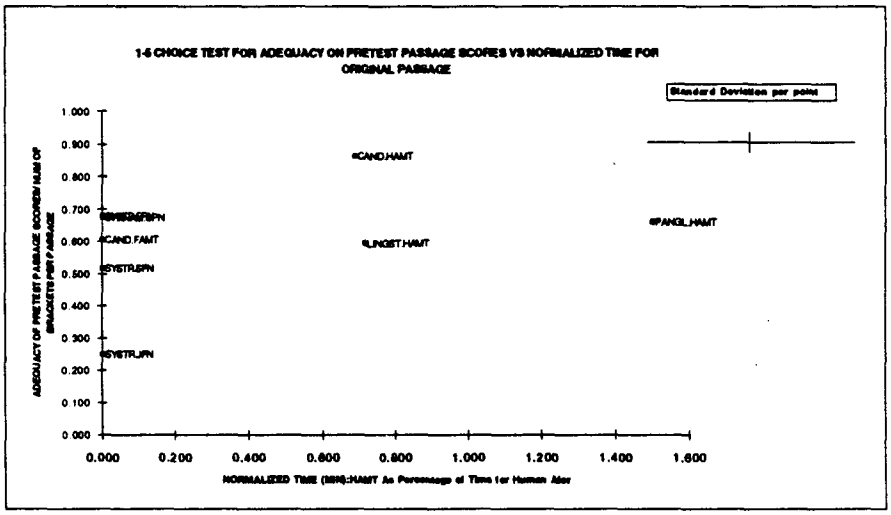

Figure 2: Adequacy Evaluation Results

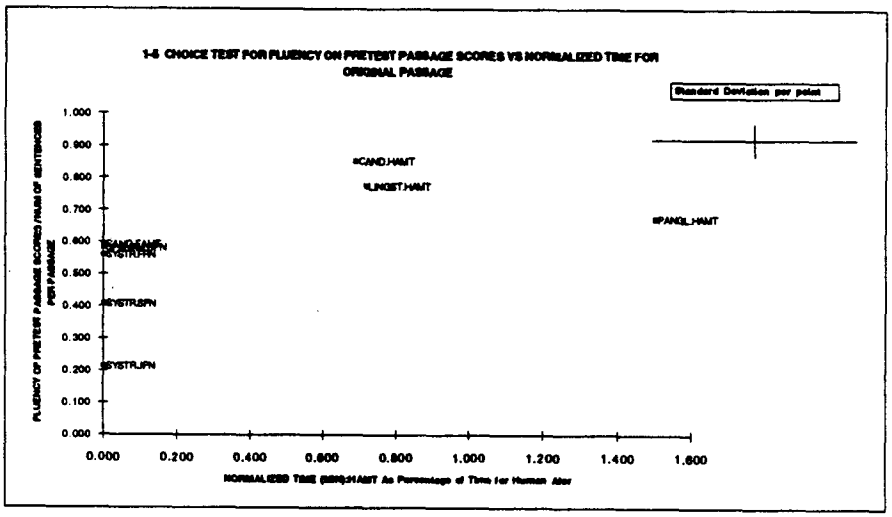

Figure 3: Fluency Evaluation Results

HAMT than FAMT. Candide HAMT receives the highest scores for adequacy, fluency and quality; Systran Japanese FAMT receives the lowest. Bounds are consistent, but occasionally Lingstat and Pangloss trade places on the y axis as do SpanAm FAMT and Systran French FAMT.

Given a similarity in performance, the comparison of evaluation suite 1 to evaluation suite 2 should depend upon the sensitivity of the measurements, as well as the facility of implementation of the evaluation.

To determine sensitivity, an F-ratio calculation was performed. For the suite 1 (Quality Panel) and suite 2, as well as for the variants that were performed on the suite 2 set (yes/no, 1-3, 1-5). The F-ratio statistic indicates that the second suite is indeed more sensitive than the suite 1 tests. (The Quality Panel test shows an F-ratio of 2.153.) The 1-3 and 1-5 versions both have certain sensitivity advantages: the 1-3 scale is central for adequacy (1.329.), but proves most sensitive for fluency (3.583). The 1-5 scale is by far the most sensitive for adequacy (4.136) and central for fluency (3.301). The 1-5 test for adequacy appears to be the most sensitive methodology overall.

The suite 2 methodologies require less time/effort than the Quality Panel. For all three scoring variants used in the second suite, less time was required of evaluators than Quality Panelists. The overall average time per passage for the Quality Panel was 26 minutes per passage, while average times for the Pre- tests were 11 minutes per passage for the 1-5 variant of adequacy and four minutes per passage for the 1-5 variant of fluency.

The level of expertise required of evaluators is reduced in the second suite; monolinguals perform the Pre-test evaluation, whereas Quality Panelists must be native speakers of English who are expert in French, Japanese or Spanish. The second suite eliminates a considerable amount of time and effort involved in preparation of texts in French, Spanish and Japanese for the test booklets.

6. NEED FOR ADDITIONAL TESTING

Human effort, expertise, and test sensitivity seem to indicate that the suite 2 evaluations are preferred over the suite 1 sets. However, the variance within a particular system result remains quite high. The standard deviations (represented in the figures as standard deviation of pooled variance) are large, due perhaps to the sample size, but also due to the fact that the baseline English used in this Suite 2 Pre-test evaluation were produced by level 2 translators, and not by professional translators. Accordingly, we intend to re-apply the evaluation of the 1992 output, using professional translations of the texts as the adequacy baseline. Results will again be compared with the results of the 1992 Quality Panel. This will help us further determine the usefulness, portability, and sensitivity of the evaluation methodologies.

The Pre-test methodologies measure the well-formedness of a translation and the degree to which a translation expresses the content of the source document. While results of the 1992 Evaluation showed that results of the Quality Panel and the Comprehension Test were comparable, a test of the comprehensibility of the translation provides unique insight into the performance of an MT system. Therefore, the 1993 Evaluation will include a Comprehension Test on versions of Original Passages to evaluate the intelligibility of those versions.

7. CONCLUSIONS

The DARPA MT evaluation methodology strives to minimize the inherent subjectivity of judging translations, while optimizing the portability and replicability of test results and accommodating the variety of approaches, languages, and end-user applications.

The two evaluation suites described in this paper sought to accomplish these goals. The comparison among them accordingly is based upon the fidelity of the measurement, the efficiency of administration, and ultimately the portability of the test to other environments. We find, subject to further testing underway, that the second suite is advantageous in all these respects.

REFERENCES

1. Bamberg, Paul. 1992. "The LINGSTAT Japanese-English MAT System" Status Report presented at the 1992 DARPA MT Workshop, Newton, MA August, 1992.
2. Brown, P. F., J. Cocke, S. A. DellaPietra, V. J. DellaPietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P.S. Roossin. 1990. "A Statistical Approach to Machine Translation." Computational Linguistics, vol. 16, pp. 79-85.
3. Nirenburg, S., J. Carbonell, M. Tomita, and K. Goodman. 1991. Machine Translation: A Knowledge-Based Approach. New York: Morgan Kaufmann.
4. Pierce, J., J. Caroll, E. Hamp, D. Hays, C. Hockett, A. Oettinger, and A. Perlis. 1966. "Language and Machines: Computers in Translation and Linguistics." National Academy of Sciences Publication 416.
5. White, J.S. "The DARPA Machine Translation Evaluation: Implications for Methodological Extensibility." Presented at the November 1992 Meeting of the Association for Machine Translation of the Americas. San Diego.