A Quantitative Method for Evaluating Machine Translation Systems

Barbara D. Broome, Ann E. M. Brodeen, Frederick S. Brundick
U.S. Army Research Laboratory

Malcolm S. Taylor
OAO Corporation

Abstract

The development of machine translation of human language has been impeded, initially by over-expectation, and subsequently by a lack of impartial and quantitative methods to measure its effectiveness. In this paper we describe a pilot study designed to explore the use of standardized reading comprehension tests in the evaluation of machine translation systems. In an investigation of the value added by a French-to-English machine translation system, we find a substantial enhancement of ability for non-French-trained participants. Our conclusions are supported by a rigorous statistical analysis. An approach of this type, which we advance as a general methodology, provides a structured method for systems evaluation, and one that is adaptable to more complex evaluation issues.

Introduction and Background

Machine Translation (MT), a computer-based application that seeks to convert the content of a passage provided in one human language to another, preserving as much of the original meaning as possible, was one of the first large applications in Artificial Intelligence (AI) funded by the U.S. government. It received considerable attention until the mid-sixties when, following a negative assessment by the Automatic Language Processing Advisory Committee, government funding was severely curtailed [1]. The over-optimism associated with AI in general, and with MT in particular, subsided, and the translation effort has continued on a much smaller scale with gradual, but marked, advances.

As international interaction has increased, both commercially and politically, the need for translation has increased in tandem. The Army in particular, with its land operations in foreign countries and its use of coalition forces, stands to benefit from translation tools. At the same time, our understanding of what can and cannot be automated has made our expectations of the contribution of AI more realistic. As a result, machine translation, with its capabilities and its limitations, has recently received more favorable attention from both the commercial and defense communities. There are a variety of translation systems on the market, ranging from quick word-to-word translators to deep syntactic and semantic analyzers. Defense efforts like the Army's Forward Area Language Converter (FALCon) program [2] and DARPA's Trans-lingual Information Detection, Extraction

and Summarization (TIDES) program [3] are examples of ongoing efforts to leverage and develop translation tools to support the soldier at all echelons.

The availability of sound MT system evaluation techniques, however, continues to be an issue [4]. The number of languages, the differing technical approaches, and the very complexity of human language itself, all present barriers to assessing the operational effectiveness of MT systems. In earlier studies, a variety of evaluation techniques were employed. Most involved subjective assessments of translation quality, emphasizing the correctness of the output syntax, morphology, and semantics.

In 1993 DARPA funded a three-component evaluation of MT technology, addressing *adequacy*, *fluency*, and *informativeness*. As in previous evaluations, each component was measured subjectively. The third, however, presented an interesting alternative concept. A paragraph was machine translated to English and the participant was given a series of SAT-like questions and asked to assess to what degree the answers could be found in the MT output [5]. Earlier studies concluded that *informativeness* was the least useful of the three components [6]. We, however, find it potentially the most useful, given one critical change: rather than subjectively assess the degree to which information can be found in the translation, simply have the participants actually answer questions based on the MT output. That is, give participants a standard foreign language reading comprehension test that is converted to English via MT.

The strengths of this comprehension-based approach are twofold. First, it focuses on the system as a pairing of the MT tool with its user. Instead of asking how good the machine-produced translation is, it focuses on the more tractable issue of how much better a translation task can be performed when given access to machine translation. Second, it moves evaluation from the subjective to the objective realm, using the number of correctly answered questions on the reading comprehension test as a measure of the system's operational effectiveness.

Our research further explores the use of reading comprehension tests to establish a baseline in the evaluation of MT technology. We consider the specific concerns identified in previous studies, refining the approach to establish a formal evaluation method. Next, we describe the results of a pilot study in which we employ a standardized reading comprehension test to assess the effectiveness of a French MT system.

The Pilot Study

Applying formal statistical techniques to the results of the pilot study, we address three research questions.

Are test scores for untrained participants without access to MT equivalent to random guessing?

If we hope to measure the effect of a French MT in improving reading comprehension for non-French-readers, we need to establish their level of performance absent MT support. While one might assume that participants without MT are randomly guessing at answers, the similarities between French and English may provide cues

(or possibly even miscues) that negate this assumption. In our pilot test, we provide questions in both French and French-to-English-via-MT formats. Future testing would be simplified if we could verify the randomness assumption, eliminating the need to include the original French questions. Using a binomial hypothesis test, we verify that test scores for untrained participants without access to MT are, indeed, indistinguishable from random guessing.

Twenty participants, with no formal training in French, took the reading comprehension portion of the 1995 SAT II: French Subject Test [7].* The reading comprehension portion consisted of 29 multiple-choice questions, each with 4 possible answers. The participants answered a combined total of 151 questions correctly.

If the responses were the result of random guessing, the number of correct responses follows a binomial distribution

$$b(x; n, p) = \binom{n}{x} p^x q^{n-x}, \qquad x = 0, 1, \ldots, n \tag{1}$$

with parameters $n = 20 \times 29 = 580$ and $p = \frac{1}{4}$.

To test the hypotheses

H_0: The results—number of questions answered correctly—do not exceed what one would expect from guessing,
H_a: The results exceed what one would expect from guessing,

it is sufficient to evaluate the sum

$$\sum_{x=151}^{580} b(x; 580, \tfrac{1}{4}). \tag{2}$$

This sum is the probability of observing 151 or more correct responses if the participants are guessing. Evaluation of expression (2) yields a p-value of 0.30, far too large to reject the null hypothesis. The distribution of $b(x; 580, \frac{1}{4})$ shown in Figure 1 suggests the adequacy of a normal approximation. We chose to evaluate expression (2) directly due to its ease of computation.

How do test scores for participants without access to MT compare to their test scores when MT is available?

This question provides insight into the critical evaluation issue of whether or not a selected MT system provides value. Using the Wilcoxon Signed-Rank Test, we find that test scores for untrained participants are significantly higher when given access to MT.

Upon completion of the exercise detailed above, the participants then proceeded to respond to the identical test after it had been submitted to SYSTRAN MT software [8] for translation into English, their native language. At the conclusion of this phase, we had at our disposal a set of 20 paired observations, $(x_i, y_i), i = 1, 2, \ldots, 20$, the number of questions answered correctly before- and after-MT.

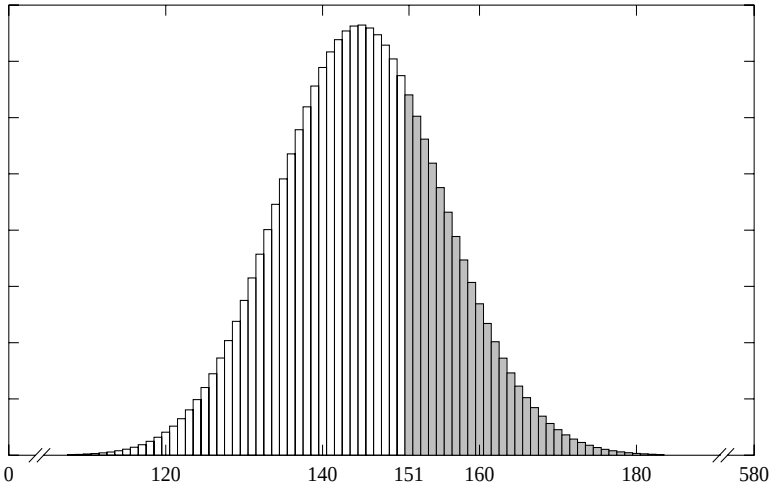


Figure 1. The Distribution Of $b(x; 580, \frac{1}{4})$.

To investigate the impact of the MT system we chose the Wilcoxon signed-rank test with hypotheses

H_0: The MT system has no discernible impact on the ability of the participants to respond correctly,
H_a: The MT system has a discernible impact on the ability of the participants to respond correctly.

Formally, the hypotheses involve comparison of the before-mean $E(X)$ and after-mean $E(Y)$.

H_0: $E(Y) \leq E(X)$
H_a: $E(Y) > E(X)$.

For these data, presented in Table 1, we have $x_i < y_i \ \forall \ i$, with one exception. In that instance, a tie occurred. Rejection of the null hypothesis is thus assured *a priori*. The p-value for the Wilcoxon test applied to these data was determined to be less than 0.005.

How do test scores for untrained participants using MT compare with those of participants who have two or more years of strong French language study?

At this point, we take full advantage of the fact that we have constructed the pilot study from the 1995 SAT II: French Subject Test, for which detailed statistics are available on the test scores of over 3,000 students. The recommended preparation for taking the French SAT is 3–4 years of French language study in high school, or two years of strong preparation. Using a two-sample t-test, we find that test scores for untrained participants with access to MT are significantly higher than those of French-trained participants that took the same French SAT test in 1995.

To compare the level of performance of the test participants with that of students taking this College Entrance Board exam, we chose a two-sample t-test with

Table 1. Number of Questions Answered Correctly

Participant	Without MT	With MT
1	10	21
2	6	14
3	9	18
4	8	21
5	6	21
6	4	17
7	7	19
8	10	21
9	6	17
10	7	15
11	11	17
12	6	22
13	9	9
14	7	21
15	9	23
16	10	20
17	6	18
18	9	20
19	4	14
20	8	9

population variances unknown. We were able to extract from information provided by the testing service an estimate of the mean and variance of student test scores based on a sample of well over 3,000 applicants, more than adequate to support a normal population assumption. The test score for an i^{th} participant may be modeled as $x_{i\cdot} = \sum_{j=1}^{29} x_{ij}$, where x_{ij} is a Bernoulli random variable with parameter p_j. The number of summands, 29, is sufficiently large to support a second normal assumption (under the Central Limit Theorem) for the conceptual population of untrained participants.

We established the hypotheses

H_0: The participants' MT-enhanced performance did not exceed that of students taking the college entrance board exam.
H_a: The participants' MT-enhanced performance exceeded that of students taking the college entrance board exam.

Formally, the hypotheses involve comparison of means of the two populations

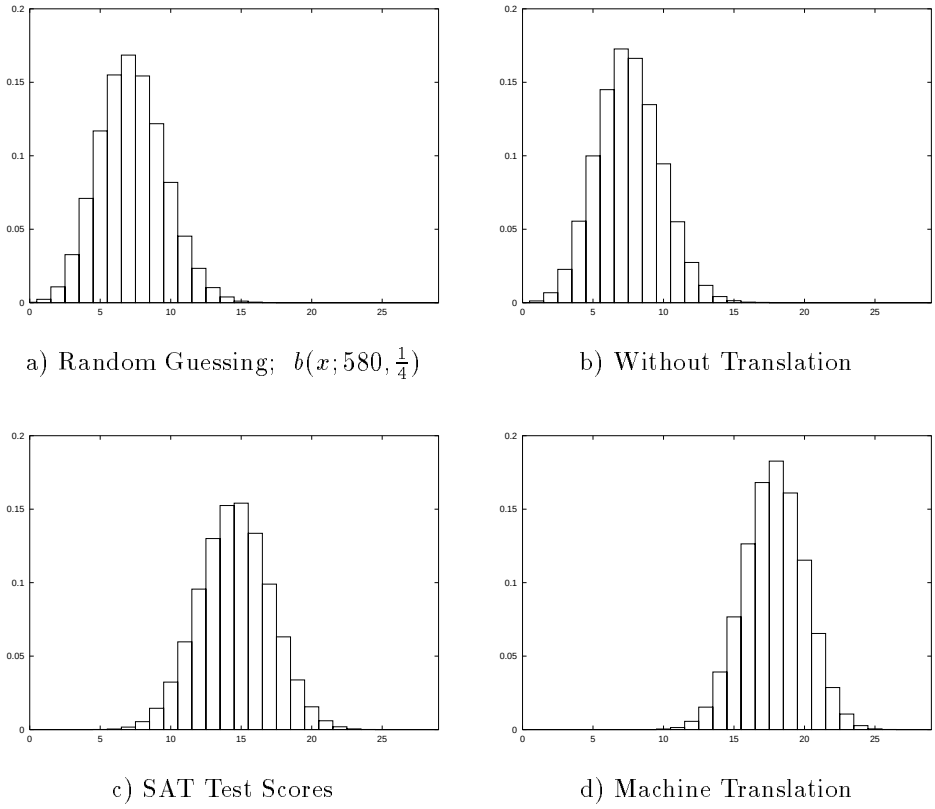
H_0: $E(X) \leq E(Y)$
H_a: $E(X) > E(Y)$

where $E(X)$ denotes the mean of the hypothetical population of participant scores and $E(Y)$ the mean of the population of student scores.

Upon evaluation of the t-statistic with 19 degrees of freedom, the null hypothesis H_0 was rejected with an associated p-value of less than 0.005. These data suggest that untrained participants, with the aid of an MT device, not only rise to the level of, but exceed, the performance of students taking the College Entrance Board exam.

A graphical summary of the three research questions appears in Figure 2.† For the statistician, this additional detail might be superfluous, but for the audience we hope to persuade, it is more likely essential. Someone for whom binomial, Wilcoxon, and t-tests are unfathomable jargon, let alone hypothesis testing and p-values, can still appreciate the striking similarity of Figures 2a and 2b (*question one*); that the histogram in Figure 2d lies to the right of that in Figure 2b (*question two*); and that the histogram in Figure 2d is to the right of that in Figure 2c (*question three*). The raw data underlying Figure 2 is presented in Table 2.

The three tests—the binomial, Wilcoxon, and Student's-t—are *a priori* nonorthogonal contrasts whose significance levels (p-values) must be accounted for at the experimental level. We chose to address this by applying a Bonferroni correction for the hypotheses tested. The Wilcoxon and the t-test, by virtue of their exceedingly small p-values, remain highly significant.



a) Random Guessing; $b(x; 580, \frac{1}{4})$ b) Without Translation

c) SAT Test Scores d) Machine Translation

Figure 2. Graphical Summary of Pilot Study

†In panels b–d the raw data are simulation enhanced.

Table 2. Number of Participants Answering Each Question Correctly

Question Number	Without MT Count	Without MT %	With MT Count	With MT %	SAT %
1	5	25	18	90	88
2	5	25	7	35	75
3	7	35	15	75	73
4	8	40	7	35	38
5	9	45	5	25	41
6	3	15	18	90	69
7	3	15	12	60	53
8	3	15	19	95	49
9	5	25	18	90	63
10	5	25	16	80	57
11	6	30	13	65	61
12	6	30	17	85	57
13	10	50	16	80	63
14	9	45	2	10	17
15	4	20	14	70	39
16	2	10	14	70	55
17	3	15	19	95	67
18	1	5	17	85	31
19	4	20	4	20	47
20	8	40	3	15	28
21	7	35	16	80	57
22	5	25	7	35	52
23	5	25	14	70	20
24	4	20	18	90	51
25	2	10	5	25	26
26	5	25	4	20	54
27	7	35	9	45	26
28	4	20	15	75	44
29	6	30	15	75	53

Conclusion

In conclusion, we submit the method outlined in this study as a general MT evaluation technique. Its emphasis on quantitative measures of effectiveness provides a greatly needed structure for impartial and statistically sound assessments of the effectiveness of MT systems. Further, the approach is flexible enough to expand and adapt to more complex issues, such as evaluating improvements to a single system under development, comparing widely diverse commercial and research systems, and assessing various user characteristics on MT effectiveness.

Acknowledgment

Ms. Broome gratefully acknowledges the helpful support and advice of LTC James Bass, Dr. Guisseppe Forgionne, Dr. Ana Schwartz, Mr. Iftikhar Sikder and Ms. Dawn French in designing the pilot study, which was initiated as a class project at the University of Maryland Baltimore County.

References

[1] Kay, M. "Machine Translation," http://www.lsadc.org/Kay.html.

[2] DeHart, J., C. Schlesiger and M. Holland. "Issues in Optical Character Recognition for Army Machine Translation," Symposium on Document Image Understanding Technology, Annapolis, MD, April 14–16, 1999.

[3] Hovy, E., N. Ide, R. Frederking, J. Mariani, A. Zampolli. "Multilingual Information Management: Current Levels and Future Abilities," a study commissioned by the National Science Foundation in 1998, http://www.cs.cmu.edu/~ref/mlim.

[4] Taylor, K. and J. White. "Predicting What MT is Good for: User Judgements and Task Performance," Proceedings AMTA, October 1998.

[5] White, J. S. and K. B. Taylor. "A Task-Oriented Evaluation Metric for Machine Translation," http://www.cst.ku.dk/projects/eagles2/2ndworkshop/annprog2.html.

[6] O'Connell, T. A., F. E. O'Mara, and K. B. Taylor. "Sensitivity, Portability and Economy in the ARPA Machine Translation Evaluation Methodology," http://ursula.georgetown.edu/mt_web/3Q94FR.html.

[7] Real SAT II: Subject Tests (French, French with Listening), The College Entrance Board and Educational Testing Services, Princeton, NJ, 1998.

[8] SYSTRAN Software, Inc. SYSTRAN PROfessional for Windows. La Jolla, CA, 1998.