Scaling the ISLE Framework: Use of Existing Corpus Resources for Validation of MT Evaluation Metrics across Languages

Michelle Vanni* and Keith Miller†

* U.S. Department of Defense
Fort Meade, MD 20755
USA
mtvanni@afterlife.ncsc.mil

† The MITRE Corporation
7515 Colshire Drive
McLean, VA 22102-7508
USA
keith@mitre.org

Abstract

This paper describes a machine translation (MT) evaluation (MTE) research program which has benefited from the availability of two collections of source language texts and the results of processing these texts with several commercial MT engines (DARPA 1994, Doyon, Taylor, & White 1999). The methodology entails the systematic development of a predictive relationship between discrete, well-defined MTE metrics and specific information processing tasks that can be reliably performed with output of a given MT system. Unlike tests used in initial experiments on automated scoring (Jones and Rusk 2000), we employ traditional measures of MT output quality, selected from the International Standards for Language Engineering (ISLE) framework: Coherence, Clarity, Syntax, Morphology, General and Domain-specific Lexical robustness, to include Named-entity translation. Each test was originally validated on MT output produced by three Spanish-to-English systems (1994 DARPA MTE). We validate tests in the present work, however, with material taken from the MT Scale Evaluation research program produced by *Japanese*-to-English MT systems. Since Spanish and Japanese differ structurally on the morphological, syntactic, and discourse levels, a comparison of scores on tests measuring these output qualities should reveal how structural similarity, such as that enjoyed by Spanish and English, and structural contrast, such as that found between Japanese and English, affect the linguistic distinctions which must be accommodated by MT systems. Moreover, we show that metrics developed using Spanish-English MT output are equally effective when applied to Japanese-English MT output.

1. Introduction

In this paper, we use existing corpus resources to validate MT output quality tests on data produced by MT systems taking as input a language that is structurally dissimilar both to the target language and to a previously tested source language.[1] In earlier work within this research program (Miller and Vanni, 2001), we validated the output quality tests on source language input that was structurally similar to the target output language, namely Spanish. In this work, we validate the tests on a structurally dissimilar language, Japanese. Our research plan, the next stage of which is automation of the tests, will investigate the correlations between score clustering patterns and tasks for which MT output has been previously determined to be suitable.

Our approach to MTE is comprised of distinct stages. These include selection of tests from the ISLE framework, test validation in terms of soundness of design and capacity for replication and automation, approaches to test automation, and experimentation with associating patterns of output quality test scores to those information-processing tasks determined to be performable with the MT output. Once clustering patterns are observed a determination will be made as to whether these patterns are predictive of the type of information processing task previously determined to be performable with the output. Since the intent is to automate the scoring system, this work can also be viewed as the preliminary steps of algorithm design.

The suite of tests is derived from the ISLE framework, and includes coherence, clarity, and measures of syntax, morphology, and lexical coverage. The coherence metric draws on Mann and Thompson's RST (1981), and is based on impressions of the overall dynamic of a discourse. Clarity is measured on a four-point scale, and is differentiated from the coherence metric in that the sentence being evaluated does not need to make sense with respect to the rest of the discourse, nor does the sentence have to be grammatically well-formed, as that feature of the output is discretely measured by the syntax metric. Scores for clarity have been shown to covary with intuitive judgements of output quality. Scores for syntax

[1] Because the corpus resources are already in place, evaluators did not carry out the tests in the blind. Filenames indicated the identities of the systems. However, evaluators endeavoured to avoid any preconceptions about systems in evaluating the output.

are based on the minimal number of corrections needed to render a sentence grammatical; likewise, the morphology scores are based on the rate of strictly morphological errors present in the output text.

We have thus far refined, validated, and tested the measures on Spanish data used in the DARPA 1994 evaluation. In this paper, we report on the suitability of our suite of measures to the output of MT systems whose source language is genetically divergent from that of the English target, namely, Japanese

The crucial characteristic of this methodology is the systematic development of a predictive relationship between discrete, well-defined metrics (a set of quality test scores) and specific information processing tasks that can be reliably performed with the output of a given MT system. We characterize MT output quality in functional terms while responding to the established desiderata for MTE. Thus, the intended outcomes are (1) a system for classifying MT output in terms of the information processing functions it can serve and (2) an indicator for research and development directions in MT designed to serve a specific information processing function. The research described in this paper, that is validating the MT evaluation metrics on Japanese data, provides a basis for the correlation of these metrics with independently-derived measures of usefulness of the output texts for downstream information processing tasks (Doyon, Taylor, & White, 1999).

In this second *test validation* stage, after an introduction to the current state of automated MT Evaluation, we review the tests and the testing process. Our context views validity as a function of (1) the ease with which tests can be applied to varying problematic output, (2) consistency with which the test criteria can be applied; and (3) the extent to which the tests might be automated in later stages of the work.

2. Automated MT Evaluation

Jones and Rusk (2000) broke new ground in automating MTE by comparing, using the K-Nearest Neighbor (KNN) algorithm, a set of linguistic test scores for MT output to a set of the same tests' scores for naturally-occurring target-language text. The goal was to determine to what extent the output was "English-like". The tests used were selected on an *ad hoc* basis, however, and the scores reported on were compared to scores for human-produced text which may not have been of the same type or domain as the text from which the MT output was produced.

Papineni, *et* al. (2001) produced an algorithm which scored MT output as a weighted sum of the counts of 1-grams, 2-grams, and 3-grams which match a synthesized version of four human translations of the input, discounting for sentence length discrepancies. The n-gram-based-scores showed a strong correlation with human judges who rated the outputs on a scale from 1 (very bad) to 5 (very good). A particularly high correlation was found to hold between the BLEU scores and those of the monolingual judges who could rate output only in terms of readability and fluency. Although the BLEU evaluation also used bilingual judges, whose scores correlated somewhat less well with those of the system, it should be noted that in both of these automated MTE systems, the criteria for "goodness" of output only indirectly addressed semantic concerns such as adequacy and informativeness. Moreover, neither provided a diagnostic to direct researchers and developers in improving systems, either generally or for a specific purpose. One recent proposal begins to address that gap.

In Papineni, *et al*. (2002), BLEU was coupled with a named-entity translation evaluation tool called NEE. Comparison of BLEU results on the DARPA 1994 MTE French-English and Spanish-English system output revealed correlations with human judgments of fluency and adequacy in the .94 and higher range. Similar comparisons of NEE results showed correlations of .98 for all but the Spanish fluency score, which exhibited a correlation of .85. Correlations with the informativeness criterion were less pronounced. Note that the translation of named entities is directly related to the use of MT output for information extraction and retrieval purposes. This makes the NEE result an important first contribution in the direction of automating a suite of diagnostic scores that provides a comprehensive picture not only of the performance of the system but also of the tasks performable on the output.

3. Task-Based MT Evaluation

Church and Hovy (1993) proposed that MTE take an approach that gives credit to a MT system for what it does well, with a focus on how it serves follow-on human processing rather than on what it is unlikely to do well. This direction has run a logical course in the Expert Advisory Group on Language Engineering Standards (EAGLES) and the International Standards for Language Engineering (ISLE) proposals for MT evaluation.

Task-based evaluation evolved from the tradition of black-box evaluation as well. This tradition has been most recently instantiated by the DARPA methodology (White and O'Connell 1994) which measured fluency, accuracy, and informativeness on a 5-point scale. Using DARPA evaluation scores and a set of translation-dependent information processing tasks, experiments were performed to rank tasks from more to less tolerant of output errors (White and Taylor 1998; Taylor and White 1998; Doyon, Talbot and White 1999).

Our approach has as its goal to determine what a system "gets right" in its output such that a human information processor or automatic process can perform a specific task with it. We select specific features of MT output proposed in the ISLE framework and we recognize that language-dependent tasks vary in their tolerance of error. We hypothesize that clustering patterns among the sets of scores resulting from the validated tests described in this paper will eventually be shown to reflect variations along these usability dimensions.

4. Data and Methods

4.1. Data

Two raters refined and validated the measures described here by testing them on MT output produced by three different Japanese-to-English systems. Input consisted of one Japanese news article. The Japanese article was chosen from among those used in the DARPA 1994 MT Evaluation and in the snap judgment task of the DARPA MT Scale research (Doyon, *et al.*, 1999).

4.2. Features and Scoring Methods

The ISLE features were selected on the basis of their measurability hypothesized effect of quality of the feature being tested on follow-on information processing tasks, and the perceived likelihood that a test for the feature could be automated in future stages of the research on this methodology. For each feature, we developed an approach to measurement and applied it to actual MT output to test its validity. Our goal was to produce a series of tests that could be applied reliably and consistently.

The features from the ISLE framework that we chose to include in our scoring suite are the following: coherence, clarity, syntax, morphology, and dictionary update/terminology. In the development of these measures, several error classification schemes (Van Slype 1979, Flanagan 1994, and Balkan 1994) were consulted. Features of informativeness, fluency, and fidelity will figure into our measurement suite in subsequent stages of the program. Measurements for these features of our test texts are available as a resource from the DARPA MT evaluation efforts, so it was not necessary to develop new scoring methods for these features.

In order to cross-linguistically validate our selection of ISLE features and our approach to scoring, the two testers worked through the output of three Japanese-to-English MT systems on a single test text in a single domain. Below, we review the scoring method for each feature, details of implementation--some of which were derived from lessons learned during the previous testing process--and guidelines for scoring with linguistic and computational motivations.

5. Validation Runs for Feature Scoring Methods

Based on findings discussed in Miller and Vanni (2001), we chose to reorder the first two tests during this application, such that the Clarity test preceded the Coherence test. It was discovered in the process of validating the tests on the Spanish-English output that the Clarity test should be performed on isolated sentences so that testers could arrive at Clarity judgments in a manner which was independent of discourse structure. Having raters perform the Coherence test, which requires assignment of discourse roles to sentences, before the Clarity test clouded raters' ability to make snap judgements on Clarity of the sentences. This is because determination of discourse function required considerable understanding of the sentence. Avoiding this training effect during this round of test application was accomplished by having the Clarity test precede the Coherence test. Two raters scored the outputs using guidelines developed in previous work.

The systems used to produce the output from Spanish input and the output from Japanese input were different. So, any comparisons of results from the two validation exercises should be viewed only as a measure of the applicability of the test suite to MT output from a source language that is structurally very different from the language on which the tests have previously been validated. Results presented here cannot be used to compare a single system's performance across the two languages.

5.1. Clarity

Our framework merges tests proposed by the ISLE framework for comprehensibility, readability, style, and clarity into a single evaluation feature which we label "clarity." This measure ranges between 0 (meaning of sentence is not apparent, even after some reflection) and 3 (meaning of sentence is perfectly clear on first reading). Since the feature of interest is clarity and not fidelity, it is sufficient that some clear meaning is expressed by the sentence and not that that meaning reflect the meaning of the input text. Thus, no reference to the source text or reference translation is permitted. Likewise, for this measure, the sentence need neither make sense in the context of the rest of the text nor be grammatically well-formed, since these features of the text would be measured by the Coherence and Syntax tests, respectively. Thus, the clarity score for a sentence is basically a snap judgement of the degree to which some discernible meaning is conveyed by that sentence.

In performing the Clarity test on the output of Japanese-to-English systems, scores were generally 0.5 point lower than those given for the Spanish-to-English system output. This is an expected result, which may derive from greater dissimilarity between source and target language in the Japanese-English pair than in the Spanish-English pair. There is still not enough data to formally measure inter-annotator agreement in the same way as for the texts that were used during the test development. Nevertheless, raters' scores for the previously unseen texts produced by the Spanish-to-English systems were very close, and often scores agreed even at the sentence level. Moreover, in the present iteration of the Clarity test, there was never more than 0.25 point difference in raters' scoring.

The results of the Clarity test are shown in Figure 1. As with the Clarity scores for the Spanish-to-English output, short sentences were usually found to yield artificially high Clarity scores. This phenomenon is most pronounced in the highest and lowest ranking Japanese-to-English system outputs in which the five shortest sentences rank generally a point higher than the five longest sentences. For the output of middle-ranking System One, the longest output sentence was nearly 30

words shorter than the longest sentence output by the highest-ranking System Two. The rest of the longest sentences were generally about ten words shorter as well. Since the output on the longer sentences contained fewer words, they were perceived as being clearer and received a commensurate score. It is interesting to note that scores on the longer sentences (translated with fewer words) were relatively similar to scores on the shorter sentences. In addition, the latter did not have the quality of being "artificially high" as they contained fewer (possibly clarifying) words as well. This sentence truncation phenomenon seemed to have a leveling effect on the general clarity of the output as both raters turned in the same Clarity score for System Two. Average sentence scores that do not even reach the 1 level on a 0-3 scale, however, indicate that current Japanese-to-English system output is fairly unclear anyway, and that it may be premature to attribute causative properties to sentence length features. This is consistent with findings from experiments in which Japanese input sentences were split into shorter units prior to machine translation into English (Gerber and Hovy 1998).

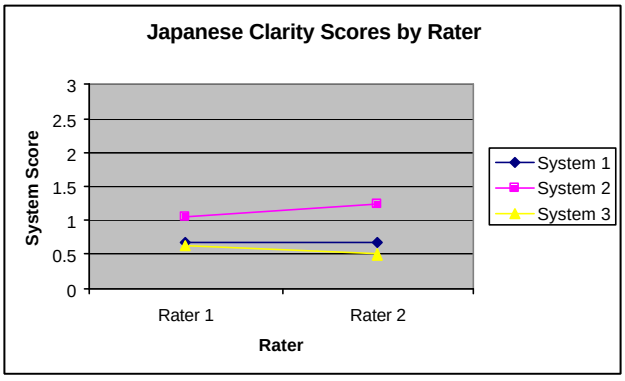

Figure 1. Clarity Scores

5.2 Coherence

Because coherence is a high-level feature that operates at a supra-sentential level, we evaluated it in the Spanish-to-English output by getting a general impression of the overall dynamic of the discourse. Wilks (1978) asserted that there is a low probability that a translation will be at the same time coherent and totally wrong. However, scores reflecting the informativeness of MT output do not show a high correlation with scores that perform surface comparisons of human translations and output (Papineni, *et al.* 2002). So, we separate out this feature of output and evaluate the coherence of the texts with respect to the text as a whole, using a measure that draws on Mann and Thompson's (1981) Rhetorical Structure Theory (RST). As in our previous work with this test, we chose the sentence as the unit of evaluation and scored this feature as the percentage of sentences to which some RST function could be assigned.

In the Spanish validation study, it was difficult for raters to divorce Coherence from meaning. When the sentence was unintelligible, even when discourse cues were present, one was tempted to assign no RST label. Based on this experience, in the present iteration of the methodology on output data from Japanese-to-English systems, we switched the ordering of the Coherence and the Clarity tests. In this way, in addition to avoiding preconceptions about meaning of the sentences during the Clarity test, work on understanding the sentence was done before work on determining the discourse function.

Recall that we apply RST very loosely in our test of coherence. For our purposes, just as for meaning in the Clarity test, it matters only that some logical function be determined for each sentence, not necessarily the "correct" one. RST definitions are used simply to constrain the set of functions that can possibly be assigned to an output sentence. However, function definitions overlap, so systematicity was crucial. In this, guidelines of the type those written by Carlson and Marcu (2001) could be helpful. Some of the distinctions found there, however, are too fine-grained for the coarse MT output.

In the Spanish-to-English output, it was noted that the ability to assign a function to a sentence was largely dependent on the ability of the rater to understand the text surrounding the sentence under consideration. For example, the occurrence of anaphoric references without an actual anaphor often led to low coherence scores. In the present work, however, since, as was noted in 5.1, the text surrounding any given sentence was fairly unclear, clues appearing internal to the sentence were more informative of the function. Since the output was unclear, the surrounding text seemed to play a less significant role.

Sometimes, a lack of clarity was directly related to a rater's inability to assign a discourse function. In Sentence Eight of System Two's output, the expression, "it was a year when," is obviously linked in some way to the previous sentence, but it wasn't clear from the rest of the sentence in what way it was related. So, no RST label was assigned.

In the present study, frequently all the information necessary to assign a discourse label is contained in a given sentence. For example, in the System One output, the raters' discourse function assignment agreed only for Sentence One and Sentence Ten.

Sentence One, System One:

```
It will be certain that this year
became in what age.
```

Sentence Ten, System One:

```
Is 34 years also the fact that
dictatorship power that depends on a
coup ranks with Estonia, Latvia,
Bulgaria also the fact that the right
wing powers make arrangements with
Spain also the fact that Hitler does
strengthen imperialism did birth of.
```

Figure 2. Text One Sentences

Note that, in Sentence One, the future tense and the pleonastic "it" serve to create an expression which is

Expository in nature while the repetition of the expression, "also the fact that," in Sentence Ten, indicates that the sentence is a Coordination. The effect of the heavy reliance on internal cues is a reduction in the amount of interpretation possible at the discourse level. In the Spanish study where the surrounding text lends itself to the understanding of the discourse function of a given sentence, a greater range was found among Coherence scores (0.2 – 1.0) than that found among the scores in the present Japanese-English study (0.3 - 0.5). In both studies, one system nevertheless showed a clear superiority over the others, and it was for this system that the raters' scores agreed most closely.

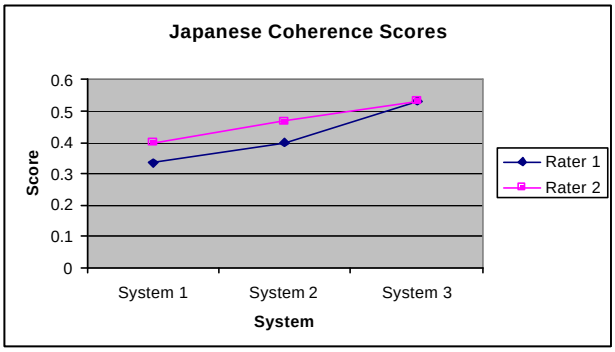

Figure 3. Results of the Coherence Test

Figure 3 illustrates the raters' scores for each system. Note that none of the scores approach 1.0, the highest score for the Spanish-English systems. Although Rater 1 tended toward higher scores in all of the tests, as we will see, both raters were consistent in their giving System 1 a lower rating than System 2 and System 3. Moreover, the difference between raters in relative rankings for the systems on this feature is small enough to lend confidence to the overall design of the test.

5.2. Syntax

As described in previous work, the syntax score is based on the minimal number of corrections necessary to render an MT output sentence grammatical (Miller & Vanni, 2001; Vanni & Miller, 2001). Each evaluator must transform each sentence in the MT output into a grammatical sentence by making the minimum number of replacements, corrections, rearrangements, deletions, or additions possible. The syntax score for each sentence is then defined as the ratio of the number of changes for each sentence to the number of words in the sentence.

Due to the greater divergence in syntactic structure between Japanese and English than between Spanish and English, it was much more difficult to apply the syntactic test to the Japanese output than it had been to apply it to the Spanish output. Nevertheless, we find that raters' scores tracked well with one another, producing the same relative ranking of systems. Furthermore, the difference in scores between raters for a given system was similar to that of the Spanish systems, with an approximately 3-percentage point maximum difference for any single Spanish system, and a 4 percentage point maximum difference for any single Japanese system. The most notable distinction between the results of application of the syntax test to the Japanese data versus the application of the test to the Spanish data was the relatively wider range of scores for the three Japanese systems. As shown in Figure 4, there was approximately a 17 point difference between the highest-ranked and lowest-ranked Japanese system, whereas this difference was approximately half that for the Spanish systems. This differentiation between systems coincides with intuitive judgements of the variability in quality of the three systems' output. As with the application of this test to the Spanish data, even when raters had the same score for a given sentence (that is, they have the same total number of changes), it is likely that they chose a different combination of the four operations to arrive at their final grammatical sentence. Although it was not done as part of this effort, further work could consider at a finer level the implications of patterns of numbers of individual syntactic test operations (i.e. deletions, substitutions, additions, and rearrangements) used by raters to make the texts grammatical.

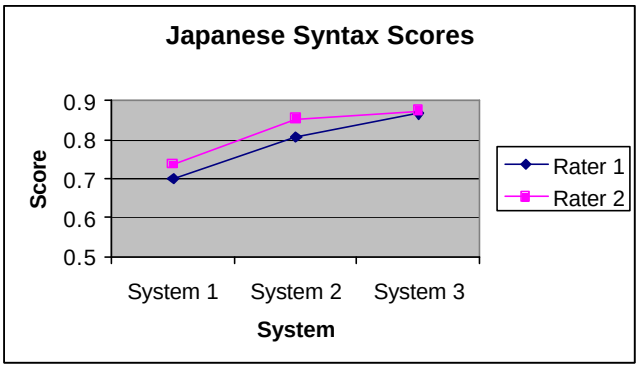

Figure 4. Results of the Syntax Test

While this work showed that it was possible to apply the same syntactic test to the output of the Japanese-English MT systems as was applied to that of the Spanish-English systems, it was also noted that it was much more labor-intensive to apply these tests to the Japanese output. Future work on the automation of this test, or of an automated procedure that correlates highly with this syntactic test will be well worth the effort in terms of labor-savings, and will be valuable in providing an easily and rapidly repeatable measure of syntactic quality that can be applied iteratively as improvements are made to MT systems.

5.3. Morphology

The morphological score is calculated as the number of morphological corrections to the MT output, divided by the total number of inflectable words in the output text. While, as with the Spanish-English output, it was at times difficult to separate purely morphological effects from

those that had their roots in syntax, this was less of a problem with the Japanese-English output in which errors in number, such as verb form errors (e.g., *make -> makes*), were the most common.

Unlike the output of the Spanish-English systems in which many sentences were found to have no morphological errors, the output of at least two of the Japanese-English systems was wildly varying in numbers of morphological errors. In the System One output, there was an error in nearly every sentence while in the System Two output, only three sentences contained errors. Moreover, although the range of final scores for each set of output was upwards of 0.9, none of the output of the Japanese-English systems ever achieved a score close to 1.0, as one of the Spanish-English systems did.

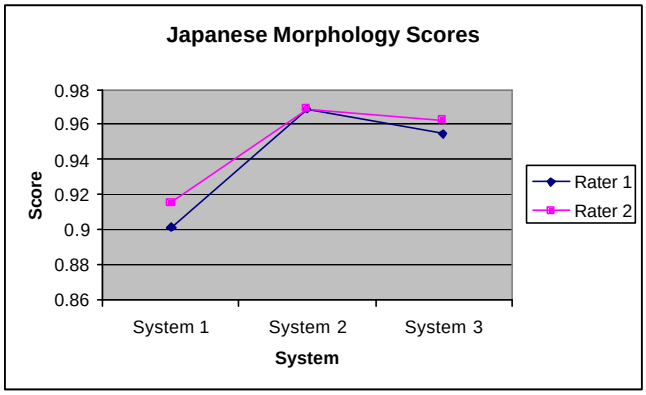

Figure 5. Results of Morphology test

In carrying out the present study, it was important for the raters that they not allow markings on texts from syntactic tests to influence scoring of morphological features. Morphological quality had to be assessed on the text as it was output by the system and not as it was reformulated in the syntax test to be grammatical. Moreover, in this test, raters encountered adjectival forms which were counted as inflectable words even when they were not, e.g., *unique*. This was done to simulate an automatic process which would hypothetically have only part-of-speech information for calculating the total number of inflectable words. Finally, the notion of *word* was not limited to whitespace-delimited lexical units: for example proper nouns containing internal whitespace, such as *United States*, were counted as one inflectable word.

As is demonstrated in Figure 5, raters showed agreement in the morphological ranking of systems and gave the same score to the highest ranking system.

In Sections 5.5, 5.6, and 5.7, complementary measures of lexical coverage and correctness have been developed and validated: two concern themselves primarily with general and domain-specific lexical coverage, and the other with the handling of named entities. The latter is believed to be crucial in determining the suitability of MT output for use in downstream information extraction tasks.

5.5 Dictionary Update

Although there are many ways that a dictionary update measure could be calculated, two objective and easy-to-observe features of MT output are the number of words not translated and the number of domain-specific words that are correctly translated. It is these two features that we previously tested as dictionary update measures in our set of evaluation measures. The non-translated word score is calculated as the percentage of non-translated words appearing in the target language document, and the domain terminology score was calculated as the percentage of correctly translated domain terms.

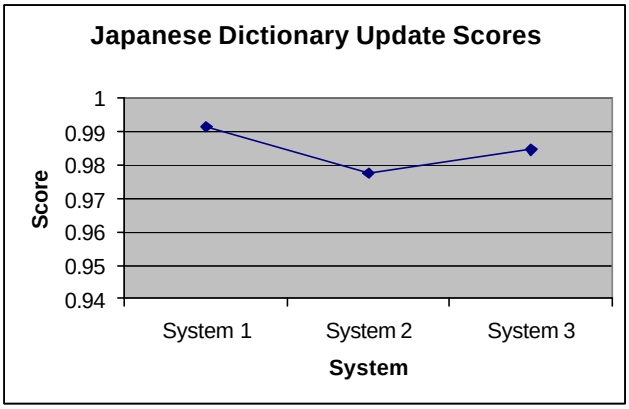

Figure 6 Results of Dictionary Update Test

As was true in the application of these tests to the Spanish-English output texts, the application of the tests to Japanese-English MT output was fairly straightforward. Terms such as *ayakariakunaino* were readily identified and accounted for. It was interesting to note that the system in question did, however, produce a romanization of the untranslated words, and did not leave them in the native script. Also included in this count were particles and other bits of non-English material, which may or may not have been the result of romanization of text found in the source. Examples of this include *na*, *re*, *X*, and *inu*. This was similar to the case of a non-word, *soed*, found in the Spanish-English output, which was also taken into account as part of the dictionary update score. Also as for the Spanish, but perhaps to an even greater extent, there were output sentences that were completely unintelligible but in no way due to untranslated words. Thus, this test could clearly not be used in isolation to provide a picture of overall MT quality, whether quality is defined along the lines of clarity, fluency, adequacy, or coherence.

Since this is a deterministic test –the presence or absence of a not-translated-word in a sentence is clear-cut– the two raters scores are identical. Hence, only one set of dictionary update scores is displayed in Figure 6. Finally, that the scores display only approximately one percentage point of difference between the highest-ranked system and the lowest suggests that the differences in scores between the systems are not significant.

5.6 Domain Terminology

This test is also clear-cut, producing a single score for each system, with no variability between raters. The procedure for this test is as follows: First, a list of key term translations is extracted from the human translation. To accomplish this, raters individually select key terms from the human translation, and then the separate key term lists are reconciled before application of the test to the MT systems' output. During the test application, systems receive a point for each term for which the translation matches exactly, and no point otherwise. The final score is the percentage of exactly-matched translations of key terms.

There are two divergent directions in which this test could be developed in the future. First, it could be made more sensitive to acceptable variation in translation of key terms by application of the ACME cloze test methodology as described in Miller (2000). Another further development of the test could involve the automated extraction of domain term list from the human translation. The application of this test to Japanese-English MT output was found to be similar to its application to Spanish-English MT output. The scores for the three systems were .50 for System 1 and .58 for Systems 2 and 3.

5.7 Names

As a special instance of a terminology score, we separately calculate the percentage of proper names correctly translated. This is directly in line with the spirit of both ISLE-based and Task Based MTE, in that scores in the named-entity translation metric should correlate strongly with the usability of the output of a machine translation system for information extraction tasks. As is the case for domain specific terms, the proper names are first identified in the reference translation. Evaluators then examine the output of each machine translation system, awarding a point for each instance of a correct translation of these proper names. As with the Terminology test, specific guidelines for the test resulted in identical scores by both raters, as seen in Table 1.

	Spanish Sys.		Japanese Sys.	
Names	Rater 1	Rater 2	Rater 1	Rater 2
Systems1	0.53	0.53	0.38	0.38
Systems 2	0.72	0.72	0.41	0.41
Systems 3	0.59	0.59	0.38	0.38

Table 1. Results of Names test

Note that the scores for the Japanese systems were quite a bit lower than those for the Spanish systems. This is probably due in part to the fact that if a Spanish MT system does not contain a proper name in its lexicon, it will likely appear in the output as an untranslated word, and there is a possibility that this will be the correct 'translation' of the name. A hypothetical Japanese system, however, must provide a romanization of the unknown name in order to be counted as correct, and even then, it is not certain that the romanization will match that found in the human translation. For example, one system produced the romanization *Kyuushuu*, which did not match the form produced by the human translator (*Kyushu*). Hence, we see that the romanization, and more importantly, variation and normalization of names is a problem that must be accounted for in the application of a named-entity translation score. The Japanese systems tended to correctly translate the names of big corporations and countries: all systems correctly translated *Mitsubishi*, *Fuji*, *Japan*, and *Bulgaria*, but at least one of the three systems did not correctly translate *Yunishi*, *Iwate*, *Datsun*, and *Estonia*. It is worth noting, too, that regardless of its international status, *United States* was incorrectly translated by all systems, which chose *America* over the human translator's two-word translation. This is one example of a situation in which a named entity translation evaluation metric that is sensitive to variation in the representation of named entities would be beneficial. As for the domain terminology metric, this variation could be accounted for by incorporating the ACME cloze testing MTE metric into the named entity testing procedure.

Our future plans in automating the tests will consider the use of the NEE automated measure of named entity translation, as described in Papineni *et al*. (2002). We believe that this measure, if applied to our texts should provide scores that are very compatible with those we arrived at manually.

5.8 Test Ordering

In previous work on validating the tests on Spanish-English MT output, an attempt was made to determine an ordering for the tests that would attenuate the training effect, such that a test on one aspect of the output would not interfere with a tester's ability to objectively assess a subsequent feature being evaluated. Based on lessons learned during these previous evaluations, the ordering of the Coherence and Clarity tests was reversed for this round of evaluations. It is felt that this reordering did produce the desired effect of minimizing, if not completely eliminating, the training effect. In fact, preceding the Coherence test with the Clarity test had a positive impact on the ability of the raters to perform both tests in an unbiased way. While this will not likely be an issue for automated versions of the tests, it was an important factor in the performance of the tests by human evaluators.

6 Conclusions and Future Work

The goal of our research program is to map objective, replicable measures for ISLE MT evaluation features to tasks for which MT output may be used (as defined in Doyon et al. (2000)) and to automate the MTE process where possible. As a crucial step in that direction, we plan to make use of the valuable resource of the DARPA MT evaluation output for which such usability data is available. We will apply our evaluation metrics to this data, and determine whether a correlation exists between patterns of scores on the MTE tests and the tasks for

which the output has been determined to be useful. As a preparatory step, in this paper, we have performed and reported on a verification run on a set of MT outputs for Japanese-English MT systems. This was an important step, because the systems represented in the DARPA task-based MTE data are Japanese-English MT systems, and our metrics had as yet been untested on any non-Romance source language. The Japanese data for this validation run is also comprised in the DARPA MTE data set, but is contained in a part of the set for which task usability data was not generated. Thus, we are conserving the crucial resource of task-tagged MT output, which is the result of a large human-intensive effort, for the final step in our research program, validating measures on other data before applying them to the actual task-based data set. We have taken account of lessons learned in previous validation runs, and now feel that the tests are ready for the final two steps: automation, to the extent possible, and application to the task-tagged translation data.

It is our belief that certain of the tests lend themselves to complete automation while the labor involved in some of the other tests could be greatly reduced by some level of automation. In particular, some of the word-based metrics (e.g. domain terms, names) could derive some level of automation as well as benefit from some added flexibility through the implementation of Miller's (2000) ACME methodology, based on cloze testing. Additionally, automated methods such as the NEE evaluation method will be examined for inclusion as the Named Entity Translation metric. Finally, since automated n-gram metrics such as BLEU have been shown to correlate with Fluency and Adequacy scores, consideration will be given to including those metrics in suite as well. This will be particularly valuable if additional task-tagged translation data is made available, for which there will not likely be analogs of the DARPA fluency and adequacy scores, which were human-intensive to generate. It is our belief that a robust, well-rounded test suite, focusing on linguistic features of the output as well as on features such as Adequacy will provide a tool that is highly predictive in determining the tasks for which the output of a given MT system may be used.

7 References

Balkan, L. 1994. Test Suites: Some issues on their use and design. Machine Translation Ten Years On, Conference at the University of Cranfield. 26-1.

Carlson, L. and D. Marcu. 2001. Discourse Tagging Reference Manual. ISI Technical Report, forthcoming.

Chaumier, J., Mallen, M.C., and Van Slype, G. Evaluation du systeme de traduction automatique SYSTRAN; evaluation de la qualite de la traduction. 1977. CEC. Report number 4. Luxembourg.

Church, K. and E. Hovy. 1993. Good applications for Crummy Machine Translation. Machine Translation 8:239-258.

Doyon, J., Taylor, K., and J. White. 1999. Task-Based Evaluation for Machine Translation. Proceedings of MT Summit 7. Singapore.

Flanagan, M. 1994. Error Classification for MT Evaluation. In Technology Partnerships for Crossing the Language Barrier: Proceedings of the First Conference of the Association for Machine Translation in the Americas, Columbia, MD.

Gerber, L. and E. Hovy. 1998. Improving Translation Quality by Manipulating Sentence Length. Lecture Notes in Artificial Intelligence 1529: Machine Translation and the Information Soup : Proceedings of the Third Conference of the Association for Machine Translation in the Americas. 448-460.

Hovy, E. 1999. Toward Finely Differentiated Evaluation Metrics for Machine Translation. Proceedings of the EAGLES Workshop on Standards and Evaluation. Pisa, Italy.

International Standards for Language Engineering. 2000. (http://www.isi.edu/natural-language/mteval) The ISLE Classification of Machine Translation Evaluations, Draft 1, October, 2000. Proceedings of the Hands-on Workshop on Machine Translation Evaluation. Association for Machine Translation in the Americas, Cuernavaca, Mexico.

Jones, D. and G. Rusk. 2000. Toward a Scoring Function for Quality-Driven Machine Translation. In Proceedings of COLING-2000.

Mann, W., and S. Thompson. 1988. Rhetorical Structure Theory: Toward a functional theory of text organization. Text 8:3.243-281.

Miller, K. 2000. The Machine Translation of Prepositional Phrases. Unpublished PhD Dissertation. Georgetown University. Washington, DC.

Miller, K. and M. Vanni. 2001. Scaling the ISLE Taxonomy : Development of Metrics for the Multi-Dimensional Measurement of Machine Translation Quality. In Proceedings of MT Summit VIII. Santiago de Campostela, Spain.

Papineni, K., S. Roukos, T. Ward, and W. Zhu. 2001. BLEU : A Method for Automatic Evaluation of Machine Translation, IBM Research Report, RC22176, September 2001.

Papineni, K., S. Roukos, T. Ward, J. Henderson, and F. Reeder. 2002. Corpus-Based Comprehensive and Diagnostic MT Evaluation : Initial Arabic, Chinese, French, and Spanish Results. Proceedings of the Human Language Technology Conference.

Polvsen, C., N. Underwood, B. Music, and A. Neville. 1998. Evaluating Text-type Suitability for Machine Translation a Case Study on an English-Danish MT System. Proceedings of ELRA Conference, Granada, Spain.

Taylor, K. and J. White. 1998. Predicting What MT is Good for : User Judgments and Task Performance. Proceedings of the 1998 conference of the Association of Machine Translation in the Americas. 364-373.

Van Slype, G. 1978. Second Evaluation of the English-French SYSTRAN Machine Translation System of the Commission of the European Communities. 1978. CEC. Final Report. Luxembourg.

Van Slype, G. 1979. Critical Study of Methods for Evaluating the Quality of Machine Translation. Prepared for the Commission of European Communities Directorate General Scientific and Technical Information and Information Management. Report BR 19142.

Vanni, M. 2000. Lessons for Text-Differentiated MT. Proceedings of the Hands-on Workshop on Machine Translation Evaluation. Association for Machine Translation in the Americas, Cuernavaca, Mexico.

Vanni, M. and K. Miller. 2001. Scaling the ISLE Taxonomy : Validating Tests of Machine Translation Quality for Multi-Dimensional Measurement. In Proceedings of the MT Evaluation Workshop of the MT Summit VIII, Santigao di Campostela, Spain.

Voss, C. and F. Reeder, eds. 1998. Proceedings of the Workshop on Embedded Machine Translation: Design, Construction, and Evaluation of Systems with an MT Component. Association of Machine Translation in the Americas Annual Meeting, Langhorne, PA.

Voss, C. and Van Ess-Dykema. 2000. When is an Embedded MT System "Good Enough" for Filtering? Proceedings of Embedded Machine Translation Systems. ANLP/NAACL 2000 Workshop. Seattle, Washington.

White, J.S. and T.A. O'Connell. 1994. The ARPA MT Evaluation Methodologies: Evolution, Lessons, and Further Approaches. Proceedings of the 1994 Conference of the Association for Machine Translation in the Americas. Columbia, MD.

White, J.S. and K. Taylor. 1998. A Task-Oriented Metric for Machine Translation. Proceedings of the First Language Resources and Evaluation Conference. Granada, Spain.

White, J., Doyon, J., and Talbott, S. 2000. Task Tolerance of MT Output in Integrated Text Processes. Proceedings of Embedded Machine Translation Systems. ANLP/NAACL 2000 Workshop. Seattle, Washington.