Scaling the ISLE Framework: Validating Tests of Machine Translation Quality for Multi-Dimensional Measurement

Michelle Vanni[1] and **Keith J. Miller**[2]

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

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

Abstract

Work on comparing a set of linguistic test scores for MT output to a set of the same tests' scores for naturally-occurring target language text (Jones and Rusk 2000) broke new ground in automating MT Evaluation. However, the tests used were selected on an ad hoc basis. In this paper, we report on work to extend our understanding, through refinement and validation, of suitable linguistic tests in the context of our novel approach to MTE. This approach was introduced in Miller and Vanni (2001a) and employs standard, rather than randomly-chosen, tests of MT output quality selected from the ISLE framework as well as a scoring system for predicting the type of information processing task 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.

1 Introduction

Work on comparing a set of linguistic test scores for MT output to a set of the same tests' scores for naturally-occurring target language text (Jones and Rusk 2000) broke new ground in automating MT Evaluation. The tests used were selected on an ad hoc basis and the scores reported on were compared to scores for humanly-produced text which may not have been of the same type or domain as the text from which the MT output was produced. In this paper, we report on work to extend our understanding, through refinement and validation, of suitable linguistic tests in the context of our novel approach to MTE. This approach was introduced in Miller and Vanni (2001a) and employs standard, rather than randomly-chosen, tests of MT output quality selected from the ISLE framework as well as a scoring system for predicting the type of information processing task performable with the output. Since the intent is to automate the scoring system, this work can also be viewed as the necessary, preliminary steps of algorithm design.

This methodology is an effort to characterize MT output quality in functional terms while responding to the established desiderata for MTE. Our research program entails a systematic development of the relationship between the evaluation metric (a set of quality test scores) and specific tasks performable on MT output It is comprised of distinct stages, to include test selection from the ISLE framework, test validation in terms of soundness of design and capacity for replication and automation, approaches to test automation, and the mapping of patterns of test scores to those information-processing tasks performable with the MT output. In this paper, we focus on the stage of *test validation* with an overview of the tests themselves and details of 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) whether the tests can be repeated by others with consistency; and (3) the extent to which the tests might be automated in later stages of the work.

2 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 the 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.

The other direction from which task-based evaluation evolved is the tradition of black-box evaluation. 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 (and eventually a computational NLP algorithm) 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 characteristics of the sets of scores resulting from the validated tests described in this paper will eventually be shown to reflect variations along these usability dimensions.

3 Data and Methods

3.1 Data

Two testers refined and validated the measures described here by testing them on MT output produced by three different Spanish-to-English systems. Input consisted of one Spanish original news text article. This material was used for the 1994 DARPA evaluation. Future work will experiment on material used in the MT Scale research.

3.2 Features and Scoring Methods

The ISLE features were selected on the basis of their measurability 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 which 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. However, scores for these features of our texts are available from the DARPA MT evaluation efforts. So for them it was not necessary to develop new scoring methods.

In order to validate our selection of ISLE features and our approach to scoring, the two testers worked through the output of three machine translation systems on a single test text in a single domain. We describe the scoring method for each feature, details of implementation discovered in the testing process, and guidelines for scoring with linguistic and computational motivations.

4 Validation Runs for Feature Scoring Methods

Although two raters scored the outputs for each feature, this step in the development of the method is not meant to be an actual inter-rater reliability study. For example, the tests were not performed completely in isolation, as would be done in such a study. On the contrary, raters were encouraged to confer, discuss the methodology, and reflect on the scoring process used to arrive at their scores. The validation procedure was carried out in much the same way as the development of guidelines for creating marked-up text as ground truth data for named-entity extraction.

4.1 Coherence

Because coherence is a high-level feature that operates at a super-sentential level, we evaluate it by getting a general impression of the overall dynamic of a discourse. Wilks (1978) asserted that there is a low probability that a translation will be at the same time coherent and totally wrong.[1] So, we evaluate the coherence of the texts with respect to the text as a whole with a measure that draws on Mann and Thompson's (1981) Rhetorical Structure Theory (RST). 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. The steps we took for this test included counting the sentences in the text, reading each sentence and attempting to assign it an RST function, assigning a score of 1 if a function could be determined and, if not, a score of 0, and, finally, adding up sentence scores for each text and dividing by the number of sentences in the text. The result of the division then was the final coherence score for the text.

This is a very loose application of RST. For our purposes, it matters only that some logical function can be determined for each sentence. It is not necessary that the MT system convey the "correct" RST function with respect to the source text or human translation. We use RST definitions simply to constrain the set of functions that can possibly be assigned to a sentence in the MT output.

Because the function definitions overlap, it was crucial that the rater be systematic in applying the RST functions. The rater had to know what distinguished the functions from each other. In this, guidelines, such as those written by Carlson and Marcu (2001), had they existed at the time, would have been helpful. For this particular application of the RST, however, it may also be that some of the distinctions were, in fact, too fine-grained for application to MT output which, in fact, is rather coarse. For future iterations, it may be desirable to select or define a subset of the functions modulo the danger of being too restrictive for use with a wide variety of (as yet unseen) domains and text types.

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. That is, a function was most assignable to a sentence when the sentence followed a sequence of other intelligent, coherent sentences. One specific example was the occurrence of anaphoric references without an actual anaphor to refer back to in the preceding text. More than once this led to the inability to assign an RST discourse label. It was also found that the RST function *Background* required greater clarity than the others used for assignment since raters needed to draw a distinction between new and old information and how both types related to the rest of the text. Other functions, such as *Elaboration*, for example, did not require the same level of clarity since embellishment of already-established information did not need to be especially clear in order to be recognized as performing that discourse function. Yet

[1] as cited in Van Slype (1979: 34)

other functions were found to be signaled by discourse cues in such a way that the clarity of the rest of the sentence was not a factor in making the coherence judgment.

Basically, it was difficult 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, future iterations of the methodology will experiment with switching the ordering of the Coherence and the Clarity tests. In this way, work on understanding the sentence can be done before work on determining the function. With the Coherence test performed before the Clarity test, the ability to make a snap judgment on Clarity was hindered.

When discourse functional distinctions between sentences were not clear, raters were advised to make the determination of whether the sentence could easily be one of several functions or whether the difficulty lay with justifying the sentence as an instance of some one of several possible functional categories. When the problem fit the former description, the sentence was to be assigned a "1" but in the latter case, it would be assigned a "0". Even though the Coherence test in its current form is a bit cumbersome and labor-intensive, it is adequate for our investigation which has two primary goals with respect to Coherence. The first is to determine whether we can develop a valid consistent measure that is reflective of this aspect of the output text. The second is to determine whether or not this Coherence feature is strongly correlated with the ability to use the MT output for the MT Scale tasks or some other relevant follow-on processing. If it turns out that Coherence is relevant to task performance, then we will revisit the test validation stage in order to develop a more tractable measure of Coherence.

Figure 1 illustrates the raters' scores for each system. Although, as we will see, Rater 1 tended toward higher scores in all of the tests, 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.

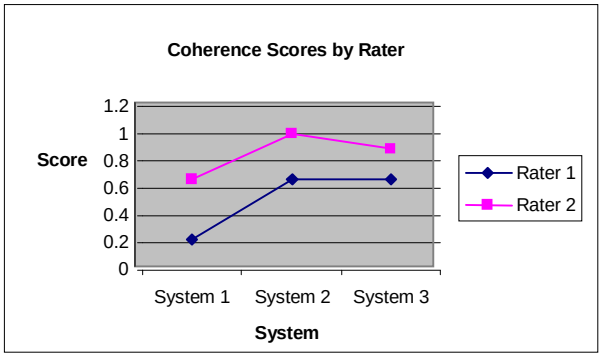

Figure 1. Results of the Coherence Test

4.2 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 is arrived at by assigning a score between 0 (meaning of sentence is not apparent, even after some reflection) to 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 "makes 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 meaning is conveyed by that sentence. The clarity score for the entire text is the mean sentence Clarity score. It is worth noting that while 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, the authors' scores for the previously unseen rating verification texts were very close, and often scores agreed even at the sentence level.

It is not surprising that short sentences were found to yield artificially high Clarity scores since the phenomena which make sentences longer, such as embedded sentences, quotations, and relative clauses, tend to complicate structure and necessitate a higher quality of translation to ensure clarity. For complex sentences to score well on this feature, the relationship between sentence parts had to be explicit.

In a comparison of rater notes, two phenomena stood out. The first was that for the exact same title output, one rater gave a score of "0" and the other rater gave a score of "2". Conversely, both raters tended to agree on their scores for sentences to which each gave a score of "3". In other words, scores converged on intuitively "better" output. We observed that when there was bad output, there was more room for interpretation or "reading into" the text for some meaning. For this reason, raters would be more lenient in their scoring of texts at the bottom end of the scale. That is, raters were more likely to agree on a score of "0" for a sentence in an otherwise good translation than they would be on a score of "0" in a lower quality translation. Thus, for the Clarity test, particularly if it is performed after the Coherence test, it was discovered that considering each sentence in isolation and independent of the discourse structure is an important element of the test design. To simulate the effect of sentences in isolation for this iteration of the tests, raters were encouraged to select sentences at random to read and score. Still, it was almost impossible to eliminate the training effect while the same evaluators were reading all the test passages.

The results of the Clarity test are shown in Figure 2. In this test, both raters had the same highest score for System 3. This confirmed what we had observed about the varying levels of interpretation of "bad" output and the resulting tendency to be lenient at the lower end of the scale. Otherwise, the correlation between raters' scores scale with their scores on Coherence, with similar differences between systems' scores for each rater. For both tests, the scores move in tandem for System 1 and 2 and the identical score which both Raters gave to System 3 on the Clarity test was a result which supported the reliability of the clarity test design.

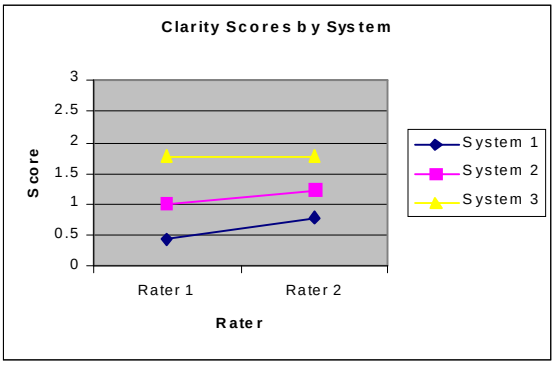

Figure 2. Results of the Clarity test

4.3 Syntax

The ISLE MT evaluation framework cites measures in Van Slype (1979) from the very high-level to the very fine-grained. Our measure produces a rather coarse-level score, and is of intermediate complexity to apply. It is an adaptation of that proposed by Chaumier, Mallen, and Van Slype (1977)[2].

The score is based on the minimal number of corrections necessary to render an MT output sentence grammatical. Each evaluator is tasked with transforming each sentence in the MT output into a grammatical sentence by making the minimum number of replacements, corrections, movements, deletions, or additions possible. These changes are then scored following the scheme of Chaumier et al. (1977) and Van Slype (1978), with the exception that corrections and replacements are counted as a single category. The syntax score for each sentence is then calculated as the ratio of the number of changes for each sentence to the number of words in the sentence; the overall syntax score for the text is calculated in an analogous manner.

Recalling the preceding discussion about the struggle to maintain a separation between evaluations of Clarity and those of Coherence, it was sometimes difficult to draw the line between purely syntactic errors and errors that crossed into other linguistic categories. Thus, we stipulated that only syntactic changes (to the particular exclusion of semantic and morphological changes) would be permitted. For this reason, if a sentence was syntactically correct but semantically anomalous, it was counted as completely correct for purposes of this feature. Likewise, a sentence with only morphological errors was counted as correct. Finally, since suppletive forms (as with case in English pronouns, `he'/`him') represent errors at the level of sentence structure, they were not taken into account in the Morphology test but instead were accounted for in the Syntax test. Note that while the Syntax test did not count, for example, Person/Number errors even when the forms in question (e.g., `are'/`is') were irregular, it did count any errors that effected a change in word category.

The raters, armed with the guidance to make the fewest changes possible and not to look at the human translation, faced two issues in particular. The first regarded questions about the nature of grammar rules and how they differ from rules of style. This quandary affected the performance of the test because, depending on where the line was drawn, a well-formed sentence would have undergone many more or many fewer changes to be arrived at. Changing MT output to create a stylistically well-formed sentence requires many more changes than does the creation of a simply grammatically correct sentence. The second, related issue concerned the determination of the fewest number of changes. If raters proceed to read a sentence of output, correcting as they go along, they arrive at the end of the sentence with possibly a larger number of changes than they might have had, had they started making changes after reading the sentence a few times. To resolve these competing priorities, raters developed strategies that involved the reading of each output sentence several times in order to formulate an idea of its possible sense. Then, reading through one more time, they could craft a basic meaningful sentence from the words available, changing the order or inserting or deleting as they go. In this way, the style question was mitigated since the emergence of the new sentence was based on an idea the rater could express with the elements present in the output and not simply on the application of some set of rules.

Figure 3 shows the results of the Syntax test. Note that the raters gave the systems the same relative ranking. In fact, the absolute scores between raters are very close. It should be pointed out as well that 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 sentence.[3]

[2] as cited in Van Slype (1979: 131)

[3] For example, for one sentence, the raters each had 7 changes. Rater 1 used 2 Replacements, 1 Rearrangement, 3 Deletions, and 1 Addition while Rater 2 used 1 Replacement, 5 Deletions, and 1 Addition.

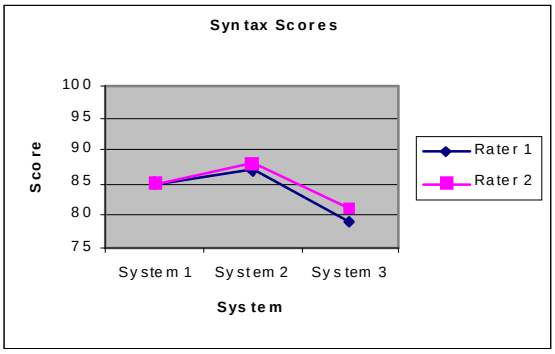

Figure 3. Results of the Syntax test

4.4 Morphology

Again, several sets of criteria were considered for possible implementation in our study; it is our aim that the measure finally chosen be objective, and thus replicable, and that there be the prospect for its partial automation in the foreseeable future. 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. It was at times difficult to separate purely morphological effects from those that had their roots in syntax. It was decided, as noted in the Syntax discussion in 4.3, that suppletive case-marking forms of English pronouns (e.g., `who'/`whom', `him'/`he') were to be counted as syntactic and not morphological errors.

Many sentences were found to have no morphological errors. Although errors such as `are'/`is', `be'/`are', `his'/`its', infinitive/inflected form and cardinal/ordinal (`11'/`11th') were all counted, there was nevertheless a concern among raters that they were being too lenient in their scoring of the output.

Figure 4 shows the results of the Morphology test. For reasons we will investigate, scores for System 3 were somewhat divergent. On the other hand, both raters had, in fact, similar scores for the other two systems and the same relative ranking for the three systems. Rater 2 found more of a distinction in performance between the two but that difference (about a hundredth of a point) is likely not to be statistically significant.

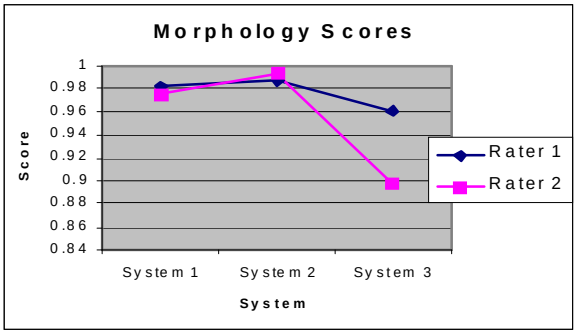

Figure 4. Results of Morphology test

4.5 Dictionary Update

Dictionary update is suggested as an MT evaluation measure in the ISLE framework. 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 chose for the dictionary update measure in our set of evaluation measures. Other possible measures, such as the number of incorrectly translated words, were left for future consideration, due to the difficulty in arriving at a precise and objective definition of such a measure. The non-translated word score is calculated as the percentage of non-translated words appearing in the target language document.

This was a fairly straight-forward test. Terms such as *iglesia del sagrado corazon* were readily identified and accounted for. Except for a couple of exceptions, such as *Cataluna*, a non-English, non-Spanish word, and a non-word, *soed*, there were few, if any, ambiguous situations. One frustration however involved an output sentence which was completely unintelligible but in no way due to untranslated words. So, the problems could not be reflected in the score for this test. It received the same score as that of the other, more intelligible sentences.

That the scores for this test, shown in Figure 5, are close and covary suggests a reliable test.

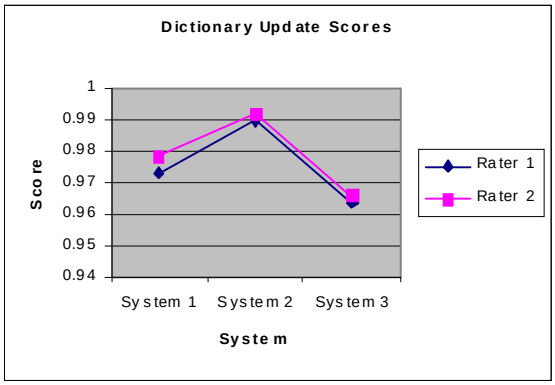

Figure 5. Results of the Dictionary Update test

4.6 Domain Terminology

Voss and Van Ess-Dykma (2000) developed an MT evaluation measure based on the percentage of domain-specific words from the source text that were correctly rendered in the translation. They further showed that it was possible to set a threshold for this measure in order to determine the utility of the machine-translated output for use in their filtering task. We thus adopt this practical measure, in the hopes that it will also correlate with results of other task-based evaluation methodologies, such as that presented by White, Doyon, & Talbott (2000). We calculate this measure as the ratio of the number of domain terms appearing correctly in the translation to the

total number of domain terms in the human reference translation.

Scanning the list of domain terms extracted from the human reference translation for the test articles (which were drawn from different domains), it is easy to see why a measure of the accuracy of translation of domain-specific terminology might correlate with the usability of a machine translation system for a task like filtering or triage. The domain of the articles could easily be determined simply by scanning the term list, without any reference to the article itself.

The important consideration for this test was that domain terms be exact. Therefore variations counted wrong for the purposes of this test included those stemming from the occurrence of non-English forms, not-translated forms, synonym usage, such as `thinkers' for `intellectuals' or `pictures' for `illustrations', misspellings, wrong ordering for phrasal constituents of terms, errors of category or form, e.g., `sculptures' for `sculpture', etc. would all count as being wrong for the purposes of this test

By contrast, formatting errors, such as lower case for proper names would be accepted in this test because for searching or extraction purposes, this aspect of the rendering would be accounted for in, for example, case insensitive searches for information retrieval .

Other issues surfaced as well in the implementation of this test. These involved features of the human translation key which were possibly differently but just as effectively rendered in the MT output, e.g., use of accents on names, and different legitimate adjectival forms such as `Argentinean' v. `Argentine'. Raters had to mark anything other than an exact match as wrong even though the variations encountered may have been correct from the language perspective. These are important issues for an algorithm designed to characterize a system's terminology-handling capability. It will have to be designed to accept as correct legitimate variations in form. Because the guidelines for this test were precise, raters could be strict in their implementation of them. For this reason, the test results in Figure 6 show very close ratings.

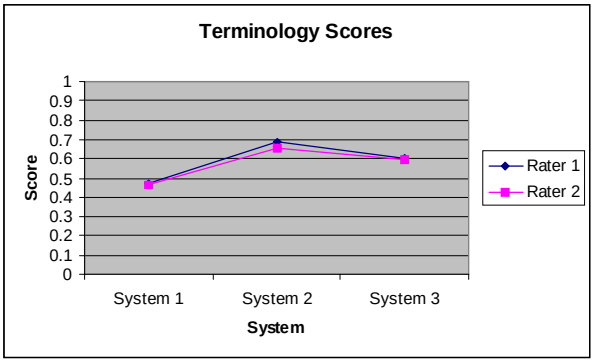

Figure 6. Results of Terminology test

4.7 Names

As a special instance of a terminology score, we separately calculate the percentage of proper names correctly translated. As for domain specific terms, the proper names are first identified in the reference translation. Evaluators then examine the output of each machine translation system, marking each instance of these proper names in the translation as correct or incorrect. Proper names appearing in the reference translation but missing from the machine translation are counted as incorrect. As with the Terminology test, specific guidelines for the test resulted in identical scores by both raters, as seen in Figure 7.

Figure 7. Results of Names test

Names	Rater 1	Rater 2
System 1	0.53	0.53
System 2	0.72	0.72
System 3	0.59	0.59

4.8 Test Ordering

The tests were ordered to achieve an attenuation of 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. We learned in the validation process that perhaps our ordering did not achieve this objective in some places. For example, after developing the Coherence test, we hypothesized that it was the most unlikely to affect the results of other tests and the most likely to be affected by the results of the other tests. This was true. What we failed to consider was that it may have been desirable to precede the Coherence test with tests which would assist the rater in its performance, such as the Clarity test. As mentioned in Section 4.1, we plan to experiment with just such a reordering in future work

5 Conclusions and Directions for Future Work

The goal of our research program is to map objective, replicable measures of ISLE MT evaluation features to tasks for which MT output may be used (as defined in Doyon et al. (2000)) and to automate the process where possible, we plan to apply our evaluation metrics to the DARPA MT evaluation output for which such usability data is available. Before using this data, however, we have performed and reported on here a verification run on a separate set of MT outputs. This run has pointed up issues to be addressed for adjusting and fine-tuning the test suite to be reflective of the different linguistic characteristics of MT output, such as test ordering, random sentence evaluation for Clarity, strategies for identifying the minimal number of changes to be made to sentences for the Syntax test, and detailing the nature of morphological errors, among other things.

Our next step is to run the suite of tests on MT output which has already been judged to be of a certain quality

for the performance of specific language-dependent information processing tasks. We will be testing our hypothesis that patterns of ISLE framework test scores on MT output equate to suitability of that output for information processing task performance. While exploring that question, we hope to discover which of the features is most predictive of the usability of MT output in the performance of each specific task.

In addition, 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. It is our plan to automate the tests in the suite to the extent that this is practical. 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.

Acknowledgements

We are grateful to Kathi Taylor, Jennifer Doyon, John White, Flo Reeder, and Lori Gerber for their valuable assistance in assembling the materials needed for this study.

Bibliography

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.

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. 2001a. 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.

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.

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.S., 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.