Scaling the ISLE Taxonomy: Development of Metrics for the Multi-Dimensional Characterisation of Machine Translation Quality

Keith J. Miller, Michelle Vanni

The MITRE Corporation
1820 Dolley Madison Boulevard
McLean, VA 22012
USA
keith@mitre.org

US Department of Defense
mtvanni@afterlife.ncsc.mil

Abstract

The DARPA MT evaluations of the early 1990s, along with subsequent work on the MT Scale, and the International Standards for Language Engineering (ISLE) MT Evaluation framework represent two of the principal efforts in Machine Translation Evaluation (MTE) over the past decade. We describe a research program that builds on both of these efforts. This paper focuses on the selection of MT output features suggested in the ISLE framework, as well as the development of metrics for the features to be used in the study. We define each metric and describe the rationale for its development. We also discuss several of the finer points of the evaluation measures that arose as a result of verification of the measures against sample output texts from three machine translation systems.

Keywords

Machine Translation Evaluation (MTE), ISLE Taxonomy, MT Scale, Task-based evaluation

Introduction

Attempts at measuring MT output quality have included the comparison of a set of test scores for MT output to a set of the same tests' scores for naturally-occurring target language text (Jones and Rusk 2000). This work broke new ground in automating MT Evaluation (MTE). However, the tests used were selected on an ad hoc basis and the scores reported on were compared to scores for human-produced text, not necessarily relevant to the text from which the MT was produced.

We propose a novel approach to MTE which employs standard, rather than randomly-chosen, features of MT output quality selected from the ISLE framework. This approach also involves a scoring system that has as its goal to predict the type of information processing tasks performable with the output.

This methodology is an effort to characterize MT output quality in functional terms while responding to the established desiderata for MTE. These include the capacity to automate and replicate the process as well as to produce results fine-grained enough to be useful to stakeholders such as users, researchers, and developers.

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, such as triage, detection, filtering, extraction, and gisting. 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. The issues of score-to-task mapping and validating test selection are crucial to our research program. This paper, however, focuses on the key stage of *test development*: the selection of MT output features from the ISLE framework and the development of tests to measure system performance with respect to these features, informed by previous approaches.

Task-Based MT Evaluation

Traditional approaches to MT evaluation do not account for the differences in the strengths of humans versus those of computers. For this reason, it was proposed by Church and Hovy (1993) that MT evaluations 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. Because the results of such methods were widely declaimed as being fairly unhelpful to developers or to users, a different tack was pursued. Using scores developed from the DARPA evaluations and a set of translation-dependent information processing tasks, experiments were performed to establish an order among the tasks performable on the output which ranked them from more to less tolerant of errors (White and Taylor 1998; Taylor and White 1998; Doyon, Talbot and White 1999).

This work takes cues from both of these directions. From the former, we have set as our 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. Furthermore, we use specific features of MT output proposed in the ISLE framework, acknowledging that "tasks performable on output" vary in their tolerance of error. We hypothesize that characteristics of the sets of scores resulting from the tests described in this paper will eventually be shown to reflect variations along these usability dimensions.

Data and Methods

Data

The measures defined and developed here were tested on MT output produced by three different Spanish-to-English systems. Input consisted of two Spanish original news texts. This material was used for the 1994 DARPA evaluation. Future work will experiment on material used in the subsequent MT Scale research.

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 many features, while several methods for measurement had been proposed, they had not been applied to actual MT output (Van Slype, 1978; ISLE, 2000). Thus, our goal was to adapt a single approach or synthesize several approaches in order to produce a method 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 addition to the criteria mentioned above, we were guided by the perceived likelihood of features to have an impact on the utility of MT output and by the ease with which feature measures could be adapted or merged, as will be discussed below. 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 also 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 it was not necessary to develop new scoring methods.

Results

As part of the process of deciding which ISLE features to include in our test suite and developing a method for scoring those features, we worked through the output of three machine translation systems on two test texts in different domains. Below, we describe the scoring method for each feature that resulted from this testing process, along with our motivations for choosing the feature and scoring method in question.

Features and Scoring Methods

Coherence

As a potential evaluation measure, coherence is attractive in that it is a very high-level feature, operating at a super-sentential level. Thus, it should be possible to evaluate coherence by getting a general impression of the overall structure of a text, without delving too deeply into the syntactic and morphological features of the output for individual sentences. Furthermore, while coherence is a monolingual phenomenon, and can thus be evaluated by a monolingual speaker referring only to the target language text, Wilks (1978) asserts that there is a low probability that a translation will be at the same time coherent and totally wrong.[1] If this is true, high-level coherence measures may correlate with fidelity measures, and possibly even with measures of clarity (see below).

In order to evaluate the coherence of the texts, which ISLE defines as "the degree to which the reader can define the role of each individual sentence (or group of sentences) with respect to the text as a whole," we devised a measure that draws on Mann and Thompson's (1981) Rhetorical Structure Theory (RST). We chose the sentence as the unit of evaluation for this coherence measure, in keeping with the spirit of this definition of coherence. The coherence score for a text is the percentage of sentences to which some RST function can be assigned, and is arrived at in the following manner:

(1) Count the number of sentences in the text.
(2) For each sentence, read the sentence, and attempt to assign an RST function to that sentence in light of the rest of the text. At no time during the coherence test may the evaluator look at either the source text or the reference (human) translation of the text. If an RST function can be determined, the sentence scores 1 for coherence; if not, the sentence score is 0.
(3) After all sentences have been scored, add the sentence coherence scores, and divide by the number of sentences. The result is the final coherence score for the text.

It is worth noting that this is a very loose application of RST: in RST, the unit of interest does not necessarily have to be a sentence, and the individual functions themselves are important. For our purposes, it matters only that some logical function can be determined for each sentence, such that a coherent structure for the overall text is evident. It is not necessary that the MT system has conveyed the "correct" RST function with respect to the source text or human translation; imposing this constraint would raise the possibility that not only the coherence, but also the fidelity of the translation is directly affecting the score. Thus, we use RST function definitions simply to constrain and define the set of functions that can possibly be assigned to a sentence in the MT output.

Clarity

In reviewing the tests proposed by the ISLE framework for comprehensibility, readability, style, and clarity, we noted that the criteria appeared similar. It was thus decided that these features be merged into a single evaluation feature, for which we chose the label "clarity." A condensed version of several of the scales cited in Van

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

Slype (1979), our clarity 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). Here again, since the feature of interest is clarity and not fidelity, it is of no consequence if the meaning conveyed by the text being evaluated is not compatible with the meaning of the original source text; it is sufficient that in the evaluator's opinion, some clear meaning is expressed by the sentence. Thus, no reference to the source text or reference translation is permitted. Likewise, it is of no import whether the sentence "makes sense" in the context of the rest of the text or if the sentence is grammatically well-formed. Those features of the text are measured by the coherence and syntax features, respectively. In sum, the clarity score for a sentence is based upon 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 not enough data to formally measure inter-annotator agreement, the authors' scores for the trial texts were very close, and often scores agreed even at the sentence level.

Syntax

Several schemata indicated by the ISLE framework as possible methods for assessing syntactic quality were considered. In particular, the ISLE MT evaluation framework cites measures in Van Slype (1979) from the very high-level to the very fine-grained. We chose a measure that produces a rather coarse-level score, and is of intermediate complexity to apply. The measure is an adaptation of that proposed by Chaumier, Mallen, and Van Slype (1977)[2].

Our syntax evaluation score is based on the minimal number of corrections necessary to render the MT output grammatical. More precisely, each evaluator is tasked with transforming each sentence in the MT output into a grammatical sentence by making the minimum number of replacements, corrections, additions, 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 corrections for each sentence to the number of words in the sentence; the overall syntax score for the text is calculated in an analogous manner.

As with the struggle to maintain a separation between evaluating clarity and evaluating fidelity as discussed above, it was sometimes difficult to draw the line between purely syntactic errors and errors that crossed into other linguistic categories. Thus, for purposes of this test, we stipulated that only syntactic changes (to the particular exclusion of semantic and morphological changes) are permitted. For this reason, if a sentence is syntactically correct but semantically anomalous, it is counted as completely correct for purposes of this feature. Likewise, a sentence with only morphological errors is counted as correct. Finally, since suppletive forms (for case of English pronouns) are not taken into account in the morphological score (see below), they are accounted for in the syntactic score.

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, for example, that suppletive case-marking forms of English pronouns (e.g., who/whom, him/he) were to be counted as syntactic and not morphological errors.

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.

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 in (White, Doyon, & Talbott, 2001). 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.

Names

As a special instance of a terminology score, we separately calculate the percentage of proper names

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

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.

A Note on Test Ordering

The ordering of tests was determined on the basis of attenuation of the training effect. When it was perceived that a test on one aspect of the output would interfere with a tester's ability to objectively assess a subsequent feature being evaluated, ordering of the tests was rearranged to avoid such interactions.

For example, after developing and evaluating the test measure for coherence, we hypothesized that the coherence test is the most unlikely to affect the results of other tests and the most likely to be affected by the priming effect of carefully examining the MT output as other tests are performed. It was thus decided that this test should be performed first in the evaluation sequence.

The result is a top-down ordering by which tests which use larger units of measure, (e.g., coherence, a sentence-based measure) are performed before tests which use smaller units of measure (e.g. domain terms, a word-based measure). In Van Slype's (1979) terms, nearly all macroevaluation measures precede micorevaluation measures for purposes of avoiding the training effect.

Conclusions and Directions for Future Work

Recalling that 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)), 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 believe that since the test suite and ordering of tests has just become stabilized, we would benefit by performing a verification run on a separate set of MT outputs. It is our hope that this run will resolve any major irregularities remaining in the test suite.

Following the verification of the tests, 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.

Finally, objective measures for some other features suggested by ISLE, such as style, proved elusive. Thus, while such features may also be useful in determining the task-usability of MT output, we have left development of metrics for these features, should they prove necessary, for future work.

Acknowledgements

We are grateful to Kathi Taylor, Jennifer Doyon, John White, Flo Reeder, and Lori Gerber for their valuable insights, discussions of prior work, and for providing the data used in this research.

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.

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.

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., 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.

Wilks, Y. 1978. The value of the monolingual component in MT evaluation and its role in the Battelle report on SYSTRAN. Luxembourg CEC Memorandum.