[*Translating and the Computer 20*. Proceedings from Aslib conference, 12 & 13 November 1998]

Evaluating MT Systems: Testing and Researching the Feasibility of a Task-Diagnostic Approach

Michelle Vanni
U.S. Department of Defense
mtvanni@afterlife.ncsc.mil

1. Introduction

In this paper, I present a method of Machine Translation Evaluation (MTE) which aligns the goals of the evaluation to those of both an information processing (IP) environment and a research and development activity. A small-scale probe into the feasibility of implementing such an evaluation is also reported on. This probe, comprising a series of tests and assessments consistent with the methodology described found correlations between (1) a system's handling of complex sentence structure and the Relevance Judgment task and (2) a system's handling of domain-specific terminology and the Topic Identification task.

2. Previous Work in MTE

Goals of MTE can differ. The U.S. government has typically sponsored large-scale evaluation programs with the goal of making research funding decisions. But, in commercial and other official settings, more pragmatic goals have motivated the pursuit of other, varied approaches to MTE. The goals of these approaches have included, among others, improvement of system output, assessment of data for suitability as input to a particular system, and selection of the appropriate engine for a well-defined real-world task.

Moreover, as MT technologies improve and niche applications become apparent, approaches to evaluating them necessarily reflect the specific characteristics of the application and thus become more complex. The chronicle of MTE reveals a progression from questions of whether or not MT is possible and how it is best possible to what its best purpose is and what approach is best employed to serve that purpose. What follows is an attempt to place the present methodology in a context of scientific pursuit of MTE which follows this progression.

Traditional Approaches to Evaluation

In the early years of MT research, evaluation was aimed at determining the feasibility of the technology itself in terms of the quality-efficiency trade-off. While MT researchers decry the negative impact of the ALP AC report on progress in MT in the U.S. in the ensuing 30 years, the emphasis which it placed on the study of linguistics and computational linguistics resulted in major advances in parsing and automatic semantic processing.[1] By contrast, in more recent MTEs, such as the evaluation component of the DARPA MT initiative, determining the

[1] In 1966, the Automatic Language Processing Advisory Committee of the National Academy of Sciences and National Research Council published its report, Language and Machines, on systems tested for intelligibility and fidelity which found MT in its then-current state not to be a cost-effective technology.

promise of a particular approach has been the focus.[2]

In both cases, the quality of MT system output was compared with that of renderings produced by a human translator and used a uniform source of input, usually journalistic or technical text, in order to maintain a systematic comparison. An important lesson learned from the last series of DARPA Evaluations performed in that style was that size of knowledge source, whether corpora for the statistical systems or lexicons for the rule-based systems, can be determinant (White 1995). The term, *black box* has been used to refer to an evaluation which does not observe the internal workings of a program but only the effects and effectiveness of the algorithm. Given the focus, in both studies, on the MT output's intelligibility/fluency and fidelity/accuracy, these approaches can correctly be classified as black box methodologies.

Modern Approaches

More recently, a major on-line service conducted an evaluation of French-to-English and English-to-French MT systems in which the focus was on input text and system efficiency (Flanagan 1994). Another corporation is currently involved in research associating the levels of quality of MT output in terms of human translators' capabilities judged by their handling of foreign language material of a well-defined level of complexity (White 1997). Yet another methodology has been proposed in Spalink (1994) which takes a user's perspective on how characteristics of an input text affect not only a system's performance but also its potential improvability.

In these studies, the characteristics of the source language input served to point up a system's capabilities in terms either of the types of input it handles, e.g., imperatives, complex sentences, etc. or the tasks to be performed with its output by the information processor, e.g., the Internet user. They can be contrasted with previous approaches by their emphasis on discovery of the internal workings of systems and on system effectiveness for particular applications. In these assessments, the application of MT technology is compared with no translation at all rather than on an ideal human translation. These approaches want the proverbial glass to be half-full, not half-empty, as did albeit unwittingly, previous methodologies. While glass-box evaluation has been traditionally used for individual system testing, in these evaluations and proposals, the "glass box" was used to compare different systems. In the on-line service study, speed, a very application-driven feature of most systems, was crucial. When systems handled input text characteristics equally well, speed was the deciding factor.

New Ideas

Hovy and Church (1991) advocated the identification of niche applications for MT systems in the spirit of allowing them to do what computers do best, e.g., repetitive tasks, number and word crunching, etc. and leaving to humans what they do best, e.g., interpreting subtle shades of meaning, untangling complex syntactic structures, etc. (Kay 1980). Hovy

---

[2] The 1992 MTE methodology tested and compared the quality and efficiency of systems using three distinct approaches to translation engine design: symbolic, statistical, and hybrid symbolic-statistical.

(1998) boils the situation down to the *process* and the *purpose* of the translation task. Taking the idea a step further, Resnik (1997) actually compared system performance on different tasks. In these new perspectives, there is a categorical shift from comparing outputs with human rendering to comparing the effective use of the outputs on well-defined tasks. In MTE research, the focus has shifted to the uses to which the translations produced will be put.

Perspective

Methodologies can be contrasted in three areas: (1) the input used, (2) what the output is judged against, and (3) the task to be performed with the translation produced. Past methodologies used homogeneous input in terms of text complexity, compared output with human quality translations, and gave no consideration to the task for which the translation was being done. Present-day MTE efforts consider varying types of input, compare the output with no translation at all, and consider system features as they would affect a single task. Contemporary investigators focus on the task itself without consideration of input complexity or output quality except in terms of the ability of a user to perform a task with it. What is crucial to realize is that the elements chosen to be included in any particular approach to MTE methodology grow out of the aim of both the translation and the evaluation itself. This concept could be termed a Task-Diagnostic approach to MTE.

The nature of analytical work in many organizations, such as international banks, media outlets, travel agencies, and governments entails the processing of many different types of texts in languages unknown to the monolingual IP professional. A series of decisions are made based on the information gleaned from these texts: filtering, selection, grouping, databasing, and reporting. Any evaluation methodology which seeks to deal with the types of translation issues faced by IP professionals in such organizations has to involve the system's handling of different types of input text and the effectiveness of systems in terms of the production of output of a quality high enough that specific decisions can be accurately made on the basis of it. What we propose is a methodology which focuses on input characterization and the performance of specific IP tasks on the basis of the output.

3. Present Methodology: The Task-Diagnostic Measure

The methodology adopted and developed in this study is one by which task-based judgments of MT output quality are assessed and diagnostic determinations of an MT system's performance on specific input characteristics are given. The assessments of the task-based judgments and the diagnostic determinations are then compared for possible associations. The aim of the evaluation is not only to predict the tasks performable by an IP professional using a given system but also to indicate areas where researchers might want to improve their MT's functionality in order to make their systems effective for a given task. The results are thus intended to reflect the requirements of decision-makers in IP and MT research. The method entails:

- identification of input in terms of linguistic, format, and terminological complexity
- quantitative assessment of system output with respect to input characteristics
- assessment of system output with respect to IP task performance
- comparison of assessment results to determine possible correlations

The goal of our probe was, for each MT system/task pair, to define a 'translatability index' to predict system effectiveness (Gdaniec 1994) and, given input with a particular set of linguistic characteristics and an IP task which required a particular level of MT output, to provide criteria for the selection of MT engine. Results were also meant to point out those areas where basic research might focus in order for MT systems to be used for specific purposes.

Our research design consisted of both MT-independent and MT-dependent analysis metrics (Spalink 1994). The former involves the assignment of linguistic, terminological, and format characteristics to data while the latter entails determining system features with judgments on the handling of particular phenomena. Determining usability of the system output in performance of an IP task predicts the value of the system for use on a specific type of text, and vice versa. Linguistic characteristics deal with grammatical phenomena and other linguistic behavior of a syntactic nature, such as ellipsis, e.g., 'John rides a bike better that Jim [rides a bike]' v. 'John rides a bike better than [John rides] a scooter' and reduced relative constructions, e.g., ‘Jane fed the dog [which had been] left at the door.' Terminological characteristics consist of lexical items representing a particular semantic domain (e.g., 'pie', 'cake','brownie','cookie'), fixed expressions, and particular usages. Format characteristics describe the arrangement of the text on the page, e.g., list, table, chart, paragraph, etc. The characteristics reflect a level of text-processing difficulty involved in the handling of the input and exclude features such as word processing codes and encoding schemes.

Documentation on MT systems is frequently not detailed enough to assist in the discovery of how the system would handle a particular text feature. For example, a brochure might state that a system handles personal names but does not indicate how it handles those names which are included in the names of corporate entities, e.g., 'Mr. Robert Jones' v. 'Bob Jones University.' It is often the case that a test suite of texts will be developed in order to determine system features by abstraction. This approach is not reliable if characterization of the text does not reflect the linguistic model on which the system is based or if delays in processing cause the system to stall before it reaches the parameter in question (Spalink 1994). But, regardless of manner of determination, the relation between input text and specific MT system is crucial to an evaluation of this nature in which the result is aimed at a matching of translation environments and MT systems featuring specific capabilities.

The system output quality aspect of the study resembles more standard methodologies in that the system is measured in relation to the input text and then on its own, in this case, in terms of the degree to which IP tasks can be performed on its output. We measure output quality in terms of the ability of an IP professional to identify topics; make relevance judgments; and reason about how the decision was reached. This stage of the study is a test of output understandability. The usual black box measures (judgments of grammaticality, degree of fluency, and the ability to answer questions from the text) are replaced with measures of the performance of these IP tasks. The standard measurements are reflected in the degree to which the potential user can reason from or make relevant decisions based on the MT system output. The tester's focus on the task attenuates the affective filter present in the making of direct judgments.

An input/output test measures how well a system performs a specific text characteristic. Unlike previous evaluations of this type, consideration is given to the characteristics of both the input texts and the MT systems being tested. These differences must be considered if a comparison is to be made between comparable things, such as factors, texts, systems, or algorithms. Matches between text characteristic and system capability scores should predict equivalent input and output texts while mismatches between text characteristics and system capabilities should predict non-equivalent input and output texts. In this study, we assemble sets of characteristics which pair loosely with particular input texts and real-world IP tasks. The following section describes the steps taken to carry out this pilot plan.

4. Glass Box and Black Box Testing and Results

Four Spanish-English MT systems were used: Gister, an in-house word-for-word glossing program, the Globalink Translation System (GTS), Systran, and PanLite, an evolving MT research engine. Feature testing was done using a corpus consisting of non-standard, heterogeneous Spanish texts, approximately 54K words (433Kbytes), taken from newspapers, on-line services, and other commercial material. A team of four Spanish linguists identified input text characteristics and graded the outputs, and ten testers performed IP tasks on the outputs.

All input was characterized as having one of six particular characteristics, so each system's performance on the handling of each characteristic was observed. The advantages of using this approach instead of artificially assembling a test corpus for this purpose were efficiency, since only the relevant characteristics were assessed, and reliability, since by using naturally-occurring language rather than synthetic test sets, the systems' strengths were not compromised by the inevitable infusion of English-like constructions. An unexpected result was the uncovering of text feature *Sets* which were loosely correlated with *text type*.

Standard black-box assessments, e.g., multiple-choice questions on the content of the output, were replaced with a series of IP tasks consisting of identifying topics, judging relevance, and expressing reasoning to support the relevance determination. The value of the tasks was twofold. High accuracy in task performance would, we thought, reflect a level of MT output quality representing a measure of the technology's value to an IP environment. The subjectivity of the assessment of how well the tasks were performed on the output coupled with the objectivity of the input characteristic assessment would allow for a linking of text characteristics with IP tasks, to answer the question, "What does a system have to do well in order to be useful in the performance of IP task X?"

While the MT-independent characterization of the input data was being performed, the lists of text features drawn up, and the selection from that inventory being made for determination of the MT-dependent characteristics, it was observed that distinct sets of features from each of the three categories tended to co-occur. Two sets of three contrasting features were turned up and the texts exhibiting them identified. The text features in the data which tended to be mutually exclusive were three-clause v. one-clause sentence complexity, specific domain terms v. alpha-numeric terms, and prose formatting v. list formatting. The feature sets, Set A consisting of one-clause complexity, alpha-numeric terminology, and list-formatting, and Set B of three-clause complexity, specific domain terminology, and prose formatting, only loosely

correlated with specific text types, e.g., Set A with advertisements, personals, and announcements and Set B with straight reporting, editorials, and letters to the editor. Since texts vary in too many different ways for any more defining correlation to be made, further inquiry into feature-type correlations was determined to be beyond the scope of the study.

Scores were assigned by system and characteristic. Each linguist observed the data and gave a score to each system's handling of each of the characteristics according to the criteria in Table One. The number of times that a characteristic was noted in the texts was a factor in the scoring only to the extent that it gave the system more or fewer opportunities to perform either well or poorly on the characteristic. Usually, additional tokens of a system's performance on a characteristic did not greatly affect the linguist's score averages, reported on in Table Seven. Rating criteria used for human translators deduct points for mistakes in the translation. A more manageable range of quality could be observed among the characteristics handled well than among mistakes in the output, so metrics based on the former were more informative. Criteria were thus drawn up for the determination of MT-dependent characteristics. A four-point rating scale, shown here, was used to judge each system's handling of each characteristic studied.

Table One: Rating Criteria for MT-dependent Characteristics

SCORE	CRITERIA FOR SCORE
0	**not translated**
1	**completely wrong**
2	**very literal, missing something, only possibly comprehensible**
3	**correct but poorly translated, awkward**
4	**well done, a good rendering, colloquial, dynamic equivalent**

The goal of this aspect of the study was to test systems with respect to the IP tasks which could be performed adequately with their output. IP tasks were viewed as a series of decisions of increasing complexity to be made on the MT output, not inconsistent with recent research on approaches to evaluating the effectiveness of MT systems in general (Resnik 1996). In this phase of the study, the performance of IP tasks of Topic Identification, Relevance Judgment, and Reasoning to Support Relevance Judgments was examined for each system to probe how and which MT systems might assist an IP professional in the process of *discernment,* described below.

To judge relevance, linguistic skills are augmented with knowledge of both subject matter and the interests of the information consumer. Since the level of these types of knowledge varied among testers, it was noted that a good deal of subjectivity could be involved in the testing. A process, termed discernment, was identified as the process of making of a series of decisions about information in a text, how that information matches a set of known criteria, how that match compares with other information-criteria matches, and the relative importance of the pieces of matched information. Our probe focused on how MT output might be used in the discernment process. To deal with this, the tasks of Relevance Judgement and Reasoning to

Support Relevance Judgements were added to the fundamental IP task of Topic Identification.

Texts were scanned and IP tasks performed for 100 texts, 50 from Set A and 50 from Set B, to serve as answer key. The texts were then run through the MT systems. An ordered arrangement of five versions (outputs of four MT systems and one Spanish original) was assembled for five testers for each of the Sets to process the material in a standard manner, determining the topic of the text and its relevance, and commenting on the reasoning process used to arrive at decisions. Their answers were compared to those in the answer key, and a score was assigned on the basis of the rating criteria presented in Tables Two, Three, and Four. The results of this assessment, reported in Table Eight were then associated with those from the glass box tests.

Criteria for the Rating of System Output for Information Processing Task Testing

Table Two: Topic Identification Scoring Criteria

SCORE	CRITERIA
2.0	Answer matches key
1.5	Answer matches but is uncertain, e.g., with question mark or "DK"
1.0	Answer incompletely matches key
0.5	Incomplete, non-content-based or no match but non-selections match those of key
0.0	Answer does match key

Table Three: Relevance Judgement Scoring Criteria

2.0	Answer matches key
1.5	Answer does not match key but includes accurate content information
1.0	Answer does not match but includes accurate decision-making information
0.5	Answer is blank or uncertain but includes accurate content information
0.0	Answer blank uncertain, neither matches key nor reflects content

Table Four: Reasoning Process Scoring Criteria

2.0	Expressed reasoning process accurately reflects content and other scores are 1.0+
1.5	Expressed reasoning process accurately reflects content and RJ score is 1.5 +
1.0	Expressed reasoning process not content based and other scores are 1.0-
0.5	Reasoning process not expressed and RJ score is 2.0
0.0	Reasoning process does not reflect content and other scores are 0.5-

The rating scale valued each response at two points with partial credit for each of the tasks tested. Separate but related criteria were developed for each task, making certain combinations of scores impossible in order to correct for inter-analyst disagreement and give an explicit read-out of what was being evaluated, especially with respect to the interaction between the Relevance and Reasoning responses. For example, if the Relevance response did not match the

key but the Reasoning response indicated that inaccurate content in the MT output was not the cause, then more credit was given than would have been given had the judgment been based on poor MT output. In the Topic Identification task, some responses, based on information gleaned from non-content-related text characteristics, such as formatting styles and length, were clues to accurate identification of the subject matter of the text. For the testers judging the Set B data, Topic Identification was easier because the topics were indicated in the title. Hence, scores for this category are higher for this Set. In the Relevance Judgment task, there was a tendency to rate as relevant outputs which had the more fluent rendering rather than those which expressed appropriate text-content information. For example, reasons for not judging relevant a particular text were given as "too garbled" or "too stilted to understand."

Information given by testers in the Reasoning Process task was considered along with that in the other two responses. Possible score combinations and their relevance are given in Tables Five and Six. If the Reasoning Process supported a Relevance Judgment which matched the key, it received the full two points. But if it did not, partial credit was assigned depending on the amount of accurate content information that was used in arriving at the decision. Set B Reasoning scores are considerably lower than those given by Set A's testers because Set B testers generally based their comments on the quality of the output and not on content discernable from the output or the real-world information that led to their decisions.

Table 5: Possible Black Box Score Combinations

RELEVANCE JUDGMENT SCORES	REASONING PROCESS SCORES				
	2.0	1.5	1.0	0.5	0.0
2.0	X			X	
1.5		X			
1.0			X		
0.5			X		
0.0			X		X

Table 6: Significance of Score Combinations

SCORE	EXPLANATION
4.0	Relevance Judgment based on accurate interpretation of text content
3.0	Relevance Judgment based on factors unrelated to accurate text content
2.5	Relevance undetermined, despite possible accurate interpretation of text content
2.0	Relevance Judgment based on factors unrelated to inaccurate text content
1.5	No Relevance Judgment due to factors unrelated to inaccurate text content
1.0	No Relevance Judgment: applying factors unrelated to inaccurate text content
0.0	No or unclear Relevance Judgment based on inaccurate text content only

Scoring reflected an interdependence between Relevance Judgment and Reasoning Process responses. It was necessary to determine how testers arrived at their relevance decisions in order to judge whether or not the decision was based on accurate information gleaned from the output text, presupposing that such information would be reflected in the Reasoning task response. Relevance Judgments based on readability only could not be given extra credit. By combining responses, we rewarded with points that part of the decision-making process that was performed on the basis of a system's accurate rendering of input content. Tables Seven and Eight show Text Characteristic and IP Task Testing results:

Table Seven: Summary Results of Text Characteristics Evaluation

TEXT FEATURES	MT SYSTEMS			
	GISTER	GTS	SYSTRAN	PANLITE
SET A DATA				
Three-clause S-complexity	3	2	2	3
international relations terms	4	1	4	2
prose format	1	0	1	1
SET B DATA				
One-clause S-complexity	1	3	4	2
alpha-numeric terms	3	1	1	1
list format	1	1	3	2

Table Eight: Results of Information Processing Task Evaluation

IP TASKS	MT SYSTEMS			
	GISTER	GTS	SYSTRAN	PANLITE
SET A DATA				
Topic	.69	.59	.49	.49
Relevance	.69	.73	.61	.75
Reasoning	.55	.62	.64	, -47
Mean Total	.64	.65	.58	.57
SET B DATA				
Topic	.81	.91	.86	.89
Relevance	.46	.61	.62	.42
Reasoning	.28	.34	.36	.23
Mean Total	.52	.62	.61	.51

Formatting was a problem for all of the systems. Prose seemed to be less of a challenge for the more sophisticated systems, PanLite and Systran, but Gister slipped into a mode by which both Spanish and English fragments would appear, one following the other, making the output difficult to read. GTS ignored tabs, left justifications, and indentations. Gister performed well with respect to the types of terminology observed. This is undoubtedly because of the ease of access which users have to the lexicons. Nevertheless, in examining the clause complexity scores, Systran's high score for three-clause sentence complexity stands out. The Systran lexicon, containing such entries as subordinating conjunctions, transition phrases, and figures of speech, permitted a smooth rendering.

In IP task testing, the relatively high marks for Set B's Topic Identification would not be construed as superior performance by the MT systems. It should be kept in mind that in this Set, the title was available which made that task easier with that data. In addition, the Set B testers tended to judge relevance on the basis of output quality rather than on content, disallowing any credit to be given to the system for their recognition of additional elements of information conveyed by the system which contributed to their decision. For this reason, the combination scores cannot be compared across Sets. The more striking result is the combination Relevance and Reasoning scores across systems within one Set. While Gister outperformed other systems on Topic Identification for the Set A data, it is Systran which was consistently high on the combined Relevance and Reasoning scores for both types of data. This would seem to indicate that, although not so effective in actually parsing the less complex inputs, the extensive types of linguistic renderings in its lexicon contribute to the passing on of additional information not necessarily conveyed by syntactic means.

Systran's IP scores might seem to indicate that the system (Systran) with the more robust grammar performs better on data involving more linguistic and terminological complexity and less format complexity. Why then did the PanLite engine perform as it did, since it had a more robust grammar than Gister or GTS? The answer lay in a feature of its processing which provides for fallback onto word-for-word mode if an exact linguistic formulation is not present in the input. This mode's similarity to Gister processing explains why these scores are comparable to those of Gister.

5. Conclusion: Value of Task-Diagnostic MTE for Large Organizations

Some researchers have explored a comparative measurement of a system's effectiveness on a series of IP decisions (Resnik 1997). While these approaches measure the extent to which an MT system can help a user make IP decisions, they do not measure other variables which are relevant for system enhancement in general and for more effective usability in particular. For system enhancement purposes, it is necessary to have some kind of glass box view of system processing so that researchers and developers can be encouraged to work on topics which will eventually improve system output in relevant ways. The proper set of input text characteristics may not yet be completely clear and, in fact, this may be an area for MTE researchers to explore. But it is certain that a determination of the processing factors that affect output quality in ways that are relevant for the performance of IP tasks is essential for increased usability of MT systems.

Similarly, comparison-only approaches do not report on the performance of systems on different types of data. So, an IP center may discover that System Y is better than System Z in producing output that helps their IP analysts identify topics, but it will still be unclear how that system might operate on the data or tasks another IP environment has to perform. Differentiated input data types are a necessary feature of MTE to support large organizations. Because the diagnostic aspect of our approach observes generalities in both system strengths and weaknesses and in differentiated data types, it serves the purpose of providing not only guidelines for selecting MT systems for particular text types but also feedback to developers regarding areas in need of improvement in their systems.

This study was a first attempt to test some hypotheses about associations between system performance on specific text-input characteristics and the ability of IP professionals to use system output on specific IP tasks. Future probes should test new input text features and combinations; measure more subtle and meaningful correlations between input characteristics and IP tasks; assess performance on different tasks; increase the quantities of data studied; and include data on system algorithms.

References

Church, K. and Hovy, E. 1991. Good applications for crummy machine translation. In Proceedings of the Natural Language Processing Systems Evaluation Workshop. Ed. by Jeannette Neal and Sharon Walter. Calspan-UB Research Center.

Flanagan, Mary. 1994. Machine Translation Evaluation: A Strategy for Compuserve. Compuserve Technical Report.

Gdaniec, C. 1994. The LOGOS translatability index. In Technology Partnerships for Crossing the Language Barrier: Proceedings of the First Conference of the Association of Machine Translation in the Americas, Columbia, MD.

Hovy, E. 1998. Creating Useful Evaluation Metrics for Machine Translation. Talk given at the First International Conference on Language Resources and Evaluation, Granada, Spain.

Kay, M. 1980. The proper place of men and machines in language translation. Xerox PARC Technical Report CSL-80-11.

Resnik, Philip. 1997. Evaluating multilingual gisting of web pages. UMIACS Technical Report. University of Maryland Institute for Advanced Computer Studies.

Spalink, K. 1994. Proposal for a differentiated text-related machine translation evaluation methodology. MT News International 9.

White, John. 1995. Approaches to Black Box Evaluation. In Proceedings of the MT Summit, Luxembourg, July 10-13, 1995.

White, John. 1997. Single measure machine translation evaluation. MS.