Is That Your Final Answer?

Florence Reeder
George Mason Univ./MITRE Corp.
1820 Dolley Madison Blvd.
McLean VA 22102
703-883-7156
freeder@mitre.org

ABSTRACT
The purpose of this research is to test the efficacy of applying automated evaluation techniques, originally devised for the evaluation of human language learners, to the output of machine translation (MT) systems. We believe that these evaluation techniques will provide information about both the human language learning process, the translation process and the development of machine translation systems. This, the first experiment in a series of experiments, looks at the intelligibility of MT output. A language learning experiment showed that assessors can differentiate native from non-native language essays in less than 100 words. Even more illuminating was the factors on which the assessors made their decisions. We tested this to see if similar criteria could be elicited from duplicating the experiment using machine translation output. Subjects were given a set of up to six extracts of translated newswire text. Some of the extracts were expert human translations, others were machine translation outputs. The subjects were given three minutes per extract to determine whether they believed the sample output to be an expert human translation or a machine translation. Additionally, they were asked to mark the word at which they made this decision. The results of this experiment, along with a preliminary analysis of the factors involved in the decision making process will be presented here.

Keywords
Machine translation, language learning, evaluation.

1. INTRODUCTION
Machine translation evaluation and language learner evaluation have been associated for many years, for example [5, 7]. One attractive aspect of language learner evaluation which recommends it to machine translation evaluation is the expectation that the produced language is not perfect, well-formed language. Language learner evaluation systems are geared towards determining the specific kinds of errors that language learners make. Additionally, language learner evaluation, more than many MT evaluations, seeks to build models of language acquisition which could parallel (but not correspond directly to) the development of MT systems. These models frequently are feature-based and may provide informative metrics for diagnostic evaluation for system designers and users.

In a recent experiment along these lines, Jones and Rusk [2] present a reasonable idea for measuring intelligibility, that of trying to score the English output of translation systems using a wide variety of metrics. In essence, they are looking at the degree to which a given output is English and comparing this to human-produced English. Their goal was to find a scoring function for the quality of English that can enable the learning of a good translation grammar. Their method for accomplishing this is through using existing natural language processing applications on the translated data and using these to come up with a numeric value indicating degree of "Englishness". The measures they utilized included syntactic indicators such as word n-grams, number of edges in the parse (both Collins and Apple Pie parser were used), log probability of the parse, execution of the parse, overall score of the parse, etc. Semantic criteria were based primarily on WordNet and incorporated the average minimum hyponym path length, path found ratio, percent of words with sense in WordNet. Other semantic criteria utilized mutual information measures.

Two problems can be found with their approach. The first is that the data was drawn from dictionaries. Usage examples in dictionaries, while they provide great information, are not necessarily representative of typical language use. In fact, they tend to highlight unusual usage patterns or cases. Second, and more relevant to our purposes, is that they were looking at the glass as half-full instead of half-empty. We believe that our results will show that measuring intelligibility is not nearly as useful as finding a lack of intelligibility. This is not new in MT evaluation – as numerous approaches have been suggested to identify translation errors, such as [1, 6]. In this instance, however, we are not counting errors to come up with a intelligibility score as much as finding out how quickly the intelligibility can be measured. Additionally, we are looking to a field where the essence of scoring is looking at error cases, that of language learning.

2. SIMPLE LANGUAGE LEARNING EXPERIMENT
The basic part of scoring learner language (particularly second language acquisition and English as a second language) consists of identifying likely errors and understanding the cause of them. From these, diagnostic models of language learning can be built and used to effectively remediate learner errors, [3] provide an excellent example of this. Furthermore, language learner testing

seeks to measure the student's ability to produce language which is fluent (intelligible) and correct (adequate or informative). These are the same criteria typically used to measure MT system capability.[1]

In looking at different second language acquisition (SLA) testing paradigms, one experiment stands out as a useful starting point for our purposes. One experiment in particular serves as the model for this investigation. In their test of language teachers, Meara and Babi [3] looked at assessors making a native speaker (L1) / language learner (L2) distinction in written essays.[2] They showed the assessors essays one word at a time and counted the number of words it took to make the distinction.

They found that assessors could accurately attribute L1 texts 83.9% of the time and L2 texts 87.2% of the time for 180 texts and 18 assessors. Additionally, they found that assessors could make the L1/L2 distinction in less than 100 words. They also learned that it took longer to confirm that an essay was a native speaker's than a language learner's. It took, on average, 53.9 words to recognize an L1 text and only 36.7 words to accurately distinguish an L2 text. While their purpose was to rate the language assessment process, the results are intriguing from an MT perspective.

They attribute the fact that L2 took less words to identify to the fact that L1 writing "can only be identified negatively by the absence of errors, or the absence of awkward writing." While they could not readily select features, lexical or syntactic, on which evaluators consistently made their evaluation, they hypothesize that there is a "tolerance threshold" for low quality writing. In essence, once the pain threshold had been reached through errors, missteps or inconsistencies, then the assessor could confidently make the assessment. It is this finding that we use to disagree with Jones and Rusk [2] basic premise. Instead of looking for what the MT system got right, it is more fruitful to analyze what the MT system failed to capture, from an intelligibility standpoint. This kind of diagnostic is more difficult, as we will discuss later.

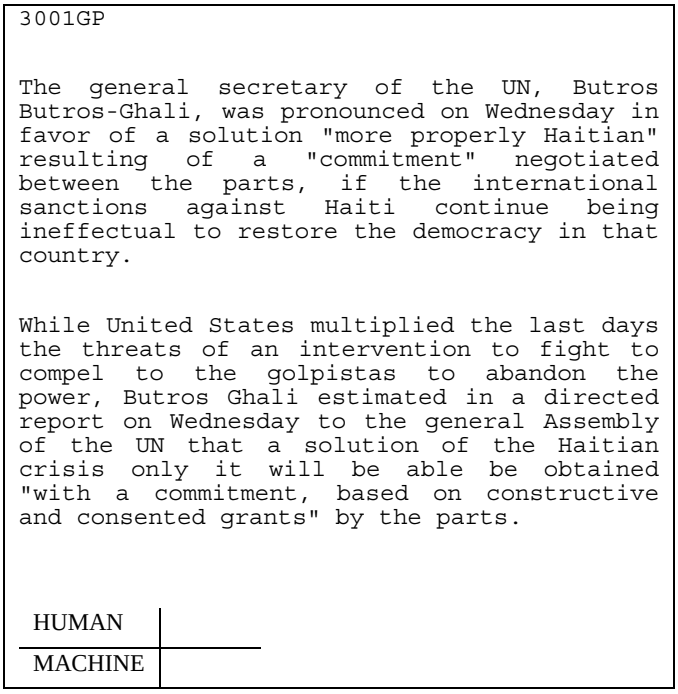
We take this as the starting point for looking at assessing the intelligibility of MT output. The question to be answered is does this apply to distinguishing between expert translation and MT output? This paper reports on an experiment to answer this question. We believe that human assessors key off of specific error types and that an analysis of the results of the experiment will enable us to do a program which automatically gets these.

[1] The discussion of whether or not MT output should be compared to human translation output is grist for other papers and other forums.

[2] In their experiment, they were examining students learning Spanish as a second language.

3. SHORT READING TEST

We started with publicly available data which was developed during the 1994 DARPA Machine Translation Evaluations [8], focusing on the Spanish language evaluation first. They may be obtained at: http://ursula.georgetown.edu.[3] We selected the first 50 translations from each system and from the reference translation. We extracted the first portion of each translation (from 98 to 140 words as determined by sentence boundaries). In addition, we removed headlines, as we felt these served as distracters. Participants were recruited through the author's workplace, through the author's neighborhood and a nearby daycare center. Most were computer professionals and some were familiar with MT development or use. Each subject was given a set of six extracts – a mix of different machine and human translations. The participants were told to read line by line until they were able to make a distinction between the possible authors of the text – a human translator or a machine translator. The first twenty-five test subjects were given no information about the expertise of the human translator. The second twenty-five test subjects were told that the human translator was an expert. They were given up to three minutes per text, although they frequently required much less time. Finally, they were asked to circle the word at which they made their distinction. Figure 1 shows a sample text.



3001GP

The general secretary of the UN, Butros Butros-Ghali, was pronounced on Wednesday in favor of a solution "more properly Haitian" resulting of a "commitment" negotiated between the parts, if the international sanctions against Haiti continue being ineffectual to restore the democracy in that country.

While United States multiplied the last days the threats of an intervention to fight to compel to the golpistas to abandon the power, Butros Ghali estimated in a directed report on Wednesday to the general Assembly of the UN that a solution of the Haitian crisis only it will be able be obtained "with a commitment, based on constructive and consented grants" by the parts.

HUMAN	
MACHINE	

Figure 1: Sample Test Sheet

4. RESULTS

Our first question is does this kind of test apply to distinguishing between expert translation and MT output? The answer is yes. Subjects were able to distinguish MT output from human translations 88.4% of the time, overall. This determination is

[3] Data has since been moved to a new location.

more straightforward for readers than the native/non-native speaker distinction. There was a degree of variation on a per-system basis, as captured in Table 1. Additionally, as presented in Table 2, the number of words to determine that a text was human was nearly twice the closest system.[4]

Table 1: Percentage correct for each system

SYSTEM	SCORE
GLOBALINK	93.9%
LINGSTAT	95.9%
PANGLOSS	95.9%
PAHO	69.4%
SYSTRAN	87.8%
HUMAN	89.8%

Table 2: Average Number of Words to Determine

SYSTEM	AVG. # WORDS
PANGLOSS	17.6
GLOBALINK	25.9
SYSTRAN	31.7
LINGSTAT	33.8
PAHO	37.6
HUMAN	62.2

The second question is does this ability correlate with the intelligibility scores applied by human raters? One way to look at the answer to this is to view the fact that the more intelligible a system output, the harder it is to distinguish from human output. So, systems which have lower scores for human judgment should have higher intelligibility scores. Table 3 presents the scores with the fluency scores as judged by human assessors.

Table 3: Percentage Correct and Fluency Scores

SYSTEM	SCORE	FLUENCY
PANGLOSS	95.9	21.0
LINGSTAT	95.9	30.4
GLOBALINK	93.9	42.0
SYSTRAN	87.8	45.4
PAHO	69.4	56.7

Indeed, the systems with the lowest fluency scores were most easily attributed. The system with the best fluency score was also the one most confused. Individual articles in the test sample will need to be evaluated statistically before a definite correlation can be determined, but the results are encouraging.

[4] For those texts where the participants failed to mark a specific spot, the length of the text was included in the average.

The final question is are there characteristics of the MT output which enable the decision to be made quickly? The initial results lead us to believe that it is so. Not translated words (non proper nouns) were generally immediate clues as to the fact that a system produced the results. Other factors included: incorrect pronoun translation; incorrect preposition translation; incorrect punctuation. A more detailed breakdown of the selection criteria and the errors occurring before the selected word is currently in process.

5. ANALYSIS

An area for further analysis is that of the looking at the details of the post-test interviews. These have consistently shown that the deciders utilized error spotting, although the types and sensitivities of the errors differed from subject to subject. Some errors were serious enough to make the choice obvious where others had to occur more than once to push the decision above a threshold. Extending this to a new language pair is also desirable as a language more divergent than Spanish from English might give different (and possibly even stronger) results. Finally, we are working on constructing a program, using principles from Computer Assisted Language Learning (CALL) program design, which is aimed to duplicate the ability to assess human versus system texts.

6. ACKNOWLEDGMENTS

My thanks goes to all test subjects and Ken Samuel for review.

7. REFERENCES

[1] 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.

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

[3] Meara, P. & Babi, A. 1999. Just a few words: how assessors evaluate minimal texts. Vocabulary Acquisition Research Group Virtual Library. www.swan.ac.uk/cals/vlibrary/ab99a.html

[4] Michaud, L. & K. McCoy. 1999. Modeling User Language Proficiency in a Writing Tutor for Deaf Learners of English. In M. Olsen, ed., Computer-Mediated Language Assessment and Evaluation in Natural Language Processing, Proceedings of a Symposium by ACL/IALL. University of Maryland, p. 47-54

[5] Somers, H. & Prieto-Alvarez, N. 2000. Multiple Choice Reading Comprehension Tests for Comparative Evaluation of MT Systems. In Proceedings of the Workshop on MT Evaluation at AMTA-2000.

[6] Taylor, K. & J. White. 1998. Predicting What MT is Good for: User Judgments and Task Performance. Proceedings of AMTA-98, p. 364-373.

[7] Tomita, M., Shirai, M., Tsutsumi, J., Matsumura, M. & Yoshikawa, Y. 1993. Evaluation of MT Systems by TOEFL. In Proceedings of the Theoretical and Methodological Implications of Machine Translation (TMI-93).

[8] White, John, et al. 1992-1994. ARPA Workshops on Machine Translation. Series of 4 workshops on comparative evaluation. PRC Inc. McLean, VA.

[9] Wilks, Y. (1994) Keynote: Traditions in the Evaluation of MT. In Vasconcellos, M. (ed.) MT Evaluation: Basis for Future Directions. Proceedings of a workshop sponsored by the National Science Foundation, San Diego, California.