In One Hundred Words or Less

Florence Reeder

MITRE Corporation
7515 Colshire Drive, McLean, VA, 22102
USA
freeder@mitre.org

Abstract

This paper reports on research which aims to test the efficacy of applying automated evaluation techniques, originally designed for human second language learners, to machine translation (MT) system evaluation. We believe that such evaluation techniques will provide insight into MT evaluation, MT development, the human translation process and the human language learning process. The experiment described here looks only at the intelligibility of MT output. The evaluation technique is derived from a second language acquisition experiment that showed that assessors can differentiate native from non-native language essays in less than 100 words. Particularly illuminating for our purposes is the set of factor on which the assessors made their decisions. We duplicated this experiment to see if similar criteria could be elicited from duplicating the test using both human and machine translation outputs in the decision set. The encouraging results of this experiment, along with an analysis of language factors contributing to the successful outcomes, is presented here.

1. Introduction

It has been said that machine translation (MT) evaluation techniques are more prolific than techniques for MT system development (Wilks, 1994). Through the long and painful history of MT evaluation, the measurements have failed to meet the desired properties of replicability, scalability and informativeness for users and developers. For instance, in the DARPA 1994 evaluations (White, et al, 1994), human raters rated each text along a five-point scale for fluency or intelligibility. While this resulted in a relative ranking of the systems, it did little to inform either users or developers about the linguistic abilities of the system. Therefore, the search continues for meaningful metrics which correlate with an overall score of success while informing specific linguistic theories, criteria or needs. One possible area where some of these metrics and the tests which accurately measure them occur is in the evaluation of language learners, particularly second language learners. We will first take a look at a recent MT evaluation performed in this paradigm, followed by a description of a language learner evaluation experiment. Finally, we will present the results of our testing based on the language learning experiment and make recommendations for automating the scoring process.

2. Measuring MT Intelligibility

Machine translation evaluation and language learner evaluation have been associated for many years (i.e., Tomita, et al., 1993; Somers & Prieto-Alvarez, 2000). One attractive aspect of language learner evaluation is the expectation that the produced language is not perfect, well-formed language. Language learner evaluation (LLE) systems are geared towards determining the specific kinds of errors that language learners make. Additionally, language learner evaluations, more than many MT evaluations, seeks to build models of language acquisition that could parallel (but not directly correspond to) the development of MT systems. These models frequently are feature-based, and they may provide informative, objective metrics which can be applied to diagnostic evaluation for system designers and system users. Finally, in the language teaching community, a large amount of study has been devoted to the finding of objective, measurable, minimal scoring effort tests which correlate with a language learner's ability. These goals of LLE make it a field which may be utilized for MT evaluation.

In a recent experiment along these lines, Jones and Rusk (2000) present a reasonable idea for measuring MT output intelligibility: they try to score the English output of translation systems using a wide variety of metrics developed from automated natural language processing (NLP) software. They look at the degree to which a given output is English and compare this to human-produced English. Their goal is to find a scoring function for the quality of English that can enable the learning of a good translation grammar. To accomplish this, they use existing NLP applications on the translated data and come up with a numeric value indicating the degree of "Englishness". They utilized syntactic measures including word n-grams, number of edges in the parse[1], log probability of the parse, execution time of the parse, overall score of the parse, etc. Their semantic criteria were primarily based on WordNet and incorporating the minimum hyponym path length, path found ration, and percentage of words with a sense in WordNet. Other semantic tests measured mutual information (a la information retrieval) for differing translations.

Two problems can be found with this approach. The first is that the data was drawn from dictionaries. Usage examples in dictionaries, while they provide practical 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 their work views the linguistic glass as half-full instead of half-empty. By focusing on the positive aspects of language, they miss the real value in analyzing the errors generated by systems. That is, unlike with language learners who benefit most from positive

[1] Both the Collins parser and the Apple Pie Parser were used for these measures.

language examples, negative exemplars are very indicative of MT improvement needs.

We believe that our results show that measuring intelligibility is not nearly as useful as finding a lack of intelligibility. This is not a new idea in MT evaluation: as numerous approaches have been suggested to identify translation errors (e.g., Flanagan, 1994). In our case, however, we are not counting errors to come up with an intelligibility score so much as finding out how quickly the intelligibility can be measured and the kinds of criteria that can be used in this judgment. Furthermore, we are basing the judgment of intelligibility on features of language learner tests which are designed to support error-filled input. Finally, it is the case that the criteria we arrive at will be used to support an overall model of MT quality, however their combination will not be a simple counting.

3. Language Learner Evaluation

The basic part of scoring learner language, particularly in second language acquisition (SLA) and English as a second language (ESL) courses, consists of identifying likely errors and understanding their cause. From these, diagnostic models of language learning can be built and used effectively to remediate learner errors (i.e., Michaud & McCoy, 1999). Furthermore, language learner testing seeks to measure a student's ability to produce language that is fluent (intelligible) and correct (adequate or informative). These correspond with the criteria typically used to measure MT system capability.[2] Finally, LLE has the goals of finding objective, consistent tests which accurately correlate with a student's abilities – a desired goal of MTE.

In looking at different SLA testing paradigms, one experiment stands out as a useful starting point for this investigation. In their test of language teachers, Meara and Babi (1999) looked at assessors making a distinction between native speakers (L1) and language learners (L2) for written essays.[3] They showed the assessors student essays one word at a time and counted the number of words it took to make the distinction.

Their first result was 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 discovered that it took longer to confirm that an essay was a native speaker's (L1) than a language learner's (L2). It took, on average, 53.9 words to recognize an L1 text and only 36.7 words to accurately distinguish an L2 text.

They ascribe the fact that L2 took less words to identify to the notion that L1 writing "can only be identified negatively by the absence of errors, or the absence of awkward writing." While the test subjects did not readily select features, lexical or syntactic, that could be consistently used in assessment, the writers hypothesize that there is a "tolerance threshold" for low quality writing. In essence, once the pain threshold has been reached through various kinds of errors, missteps or inconsistencies, each with a different weight, then the assessor could confidently make the proper attribution. While the researchers' purpose was to rate the language assessment process, the results are intriguing from an MT evaluation perspective.

With this experiment in mind, we believe that MT intelligibility assessments can be viewed similarly and take this as a starting point for rating MT intelligibility. The first question we wish to answer is: Does this kind of test apply to distinguishing between expert translation (ET, corresponding to L1) and MT output (corresponding to L2)? The second question is: Does the ability for subjects to differentiate ET from MT correlate with the intelligibility scores for the text as assigned by human raters? The final question is: Are there characteristics of the MT output which enable the decision to be made quickly and can these characteristics be used to design an automated test for MTE? This experiment is a step towards answering these questions.

4. Reading Test

We started with publicly available data that was developed during the 1994 DARPA Machine Translation Evaluation (White, et al., 1994), focusing on the Spanish language evaluation first.[4] We selected the first 50 translations for each system as well as for the two human translations. 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 that they represent a different style of language than essays and could serve as distracters.

The participants, all native speakers of English, were recruited through the author's workplace, the author's neighborhood and other locations. Each subject was given a set of six extracts – a mix of different machine and human translations where no articles were duplicated within a test set. Different subjects had a different mix of the number of machine translations versus the number of human translations. The participants were told to read line by line until they were able to make a distinction between the possible types of authors of the text – a human translator or a machine translation program. Twenty-five test subjects were given no information about the expertise of the human translator, while twenty-five test subjects were told that the human translator was an expert. To enforce a snap-judgment decision, subjects were given only 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 test sheet.

[2] The discussion of whether or not MT output should be compared to human translation output is a moral one: from our standpoint, human translation represents the best that can be done at this time.

[3] In their experiment, they examined students learning Spanish as a second language.

[4] Currently available at: http://issco-www.unige.ch/projects/isle/mteval-april01/

3002PA

Umberto Bossi, chief of the federalist Northern League, one of the three parties of the new majority of right in Italy, induced Wednesday the interruption of the negotiations conducted by the new president of the Council, Silvio Berlusconi, in order to form the new Italian government.

In the afternoon, in a note transmitted to the Italian national Assembly, the federalist movement, that from the beginning of the political consultations required that is attributed him the key ministry of the Interior, had already stated that demanded the suspension of the conversations held with the national Alliance and with Forza Italia for the formation of the government."

HUMAN	
MACHINE	

Figure 1: Example test sheet for PAHO system

5. Results

In general, the results were better than expected. It should be noted that this only addresses the intelligibility question and not the fidelity question. Translators (professional or student) would never think of committing some of the kinds of intelligibility errors MT systems do.[5] In looking at the scores, then we must limit ourselves to viewing how this reflects intelligibility judgments alone.

It could be argued that this is more of a Turing test than a measure of MT quality. While it does have the flavor of a Turing test[6], what we are trying to get at is the kinds of errors that contribute to the perception of lack of quality. It is worthy to say that most participants, particularly those in engineering, attributed the highest quality of work to humans and expected the machines to make the mistakes. That is, their expectation was of L1 quality from humans and L2 quality from MT. Although interesting from a sociological point of view, it digresses from our main topic.

Subjects were able to distinguish MT output from human translations 88.4% of the time overall. This determination was more accurately made for these readers than the L1/L2 distinction for language testers. Table 1 shows the results where the percentage given is the number of times the document's generation was correctly attributed, as broken down by document sources.

[5] Thanks to the review who pointed out that translation students produce expert sounding, but totally wrong translations.

[6] And it will be interesting to compare this to Loebner results.

SOURCE	% CORRECT
Paho	69.4%
Systran	87.8%
Human	*89.8%*
Globalink	93.9%
Lingstat	95.9%
Pangloss	95.9%

Table 1: Percentage of Correct Attribution by Source

From this data, the first question to be answered is: Does this kind of test apply to distinguishing between expert translation and MT output? The simple answer is yes. Users can make an L1/L2 type of determination between sources of a document in a relatively small number of words. Because of the indication of potential success, we advance to the next questions which look at the measures to be inferred more closely.

We examine question 2, does the ability for subjects to differentiate ET from MT correlate with the intelligibility scores for the text as assigned by human raters? To determine this, again at the average level for systems, we look at the correct attribution scores charted against the fluency (intelligibility) scores as determined in the DARPA tests. Table 2 shows this in terms of numbers and Figure 2 shows it pictorially.

SOURCE	% CORRECT	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
Human	*89.8*	*89.2*

Table 2: Correct Attribution and Correlating Fluency

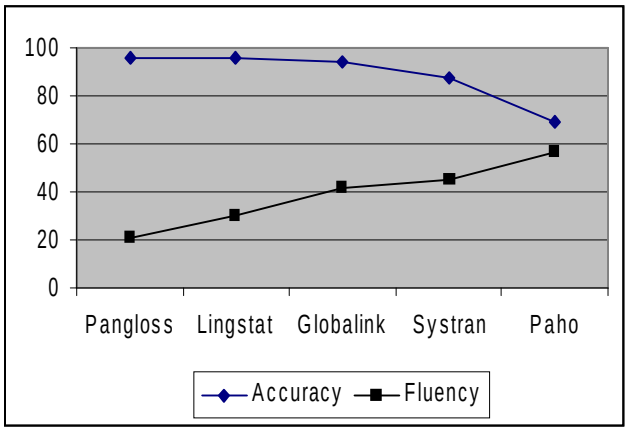

Figure 2: System Attribution Accuracy and Fluency

If one assumes that the higher the fluency score, the more intelligible the system output, the harder it is to distinguish, then there exists a correlation between this measure and fluency. As system fluency increases for

each system, the accuracy of correctly attributing its source decreases. Indeed, the systems with the lowest fluency scores were most accurately attributed.

Another measure of the ease of attributing a system is in the word count. That is, a system that is less intelligible would, according to the test, take fewer words for assigning it to the correct category. Table 3 reports the average number of words for category assignment per system, with the human score also included.

SOURCE	AVG. # WORDS
Pangloss	17.6
Globalink	25.9
Systran	31.7
Lingstat	33.8
Paho	37.6
Human	*62.2*

Table 3: Average Number of Words for Each Source

This shows that indeed, the number of words does increase as the system fluency increases, as shown in figure 3[7]. The implication is that the automated tests developed from these results will not need to analyze masses of text, at least at the lowest fluency levels.

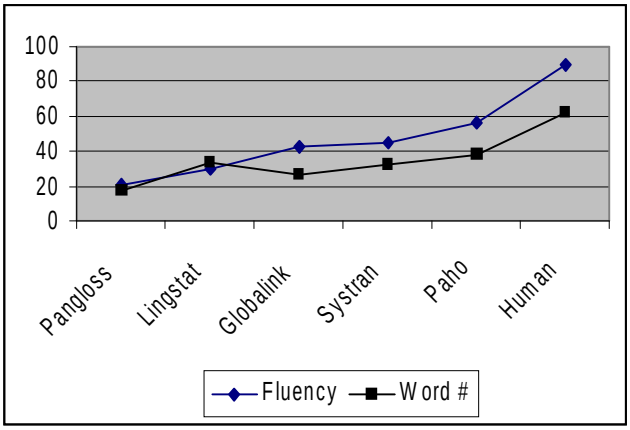

Figure 3: Fluency Scores and Number of Words

The final question for the experiment: Are there characteristics of the MT output which enable the decision to be made quickly? The initial results lead us to believe that this is so. Individual articles in the test sample will need to be evaluated statistically before a definite correlation can be determined, but the results are encouraging.

The factors which contribute to this quick decision are many and varied. A preliminary analysis has shown that not-translated words (other than proper nouns), incorrect pronoun handling, inconsistent preposition handling and incorrect punctuation were generally immediate clues as to the fact that a system produced the text. We will examine each of these in order.

The not-translated word effect is not surprising, as no translator would think of putting a source text word in the final text, preferring to omit information or add information to compensate (see Loehr, 1998) for an interesting description of this. This implies that vocabulary acquisition research should be examined as a source of more accurate scoring for systems.

Incorrect pronoun translation is another known area of deficiency for MT systems. Again, this is not new to the MT community, but the importance of it as a possible evaluation criteria is useful. One system in particular utilized the "every possible translation" strategy for pronouns[8] which was a dead giveaway to readers, and in fact changed the minds of a few.

Inconsistent preposition translation also was mentioned in post-interviews as a source of error that gave away translations as MT. In particular, some subjects had a threshold of the number of consecutive prepositions that once hit marked the translation as machine.

Incorrect punctuation – everything from misplaced commas to lack of capitalization of proper nouns – is another major source of determination mentioned. While not surprising as well (see Thompson & Brew, 1996), the human subject results indicate that Thompson & Brew were on the right track. Intuitively, we do use the most straightforward cues. Automating the testing of this (finding the right metrics) may be different matter as they learned in their study.

Another area for further analysis is the details of the post-test interviews. These have consistently shown that the deciders utilized error spotting techniques, although the types and sensitivity thresholds varied from subject to subject. Some errors were serious enough to make the choice obvious, while others had to occur more than once to push the decision above the threshold point.

6. Future Directions

We believe that we have shown that, for intelligibility at a minimum, the approach of designing a set of simple, yet indicative tests a la language learning evaluation is a feasible exercise. Of course, more work is necessary to design a framework which corresponds to a learner model and to choosing criteria which would then feed the model for an overall score. What follows is an idea of how this might occur.

Usual methods of rating the quality of MT output have relied on human judges assigning scores on a graded scale (such as 1→5; 1→7 or 1→9). Each notch in the graded scale is described for the raters, inter-rater reliability is measured and a system is assigned a score accordingly. This holistic kind of scoring is subject to human factors issues such as item ordering, yet it does reliably capture information about the relative quality of the output. The biggest difficulty is that the general ratings cannot give

[7] Note, the author understands the relative mismatch between Y-axis labels – more analysis will provide a better picture of correspondence – although it can be seen that the relative shape of the curves is the same.

[8] He/she/it as example

reasonable indicators of the factors which make one translation better than another, nor do they capture what about the MT output is meaningful to an end user. Therefore, it is reasonable to take a slightly different look at the MTE problem.

Given that we have a body of data for which there are human ratings, our slightly different look at MTE resembles the work done in educational testing measurement (e.g., Burstein & Chodorow, 1998). In educational testing measurement, essays are graded by humans, again on a sliding kind of scale. This rating is accepted as the "gold standard" of measurement. The problem then becomes not designing new measurement sets, but instead of trying to identify and measure the criteria which contribute to the ratings – a classical machine learning problem of categorizing items based on usually objective and reasonably measured criteria. In this way, MTE becomes a much more replicable, automatable task while at the same time continuing to capture the human intuition of quality output. To avoid "gaming" the system, we will need to continually check the framework and points in the framework for indication of overall quality.

Acknowledgements

Thank you to the anonymous reviewers for their very helpful comments on the presentation of results.

Note

The views expressed in this paper are those of the author and do not reflect the policy of the MITRE Corporation.

Bibliographical References

Burstein, J. & M. Chodorow. 1999. Automated Essay Scoring for Nonnative English Speakers. 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. 68-75.

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.

Loehr, D. 1998. Can Simultaneous Interpretation Help Machine Translation. In Proceedings of the Conference of the Association for Machine Translation in the Americas, AMTA-98. Langhorne, PA.

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

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

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

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.

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).

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

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.