How Are You Doing? A Look at MT Evaluation

Michelle Vanni & Florence Reeder

1820 Dolley Madison Blvd., McLean VA 22102
mtvanni@afterlife.ncsc.mil & freeder@mitre.org

Machine Translation evaluation has been more magic and opinion than science. The history of MT evaluation is long and checkered – the search for objective, measurable, resource-reduced methods of evaluation continues. A recent trend towards task-based evaluation inspires the question – can we use methods of evaluation of language competence in language learners and apply them reasonably to MT evaluation? This paper is the first in a series of steps to look at this question. In this paper, we will present the theoretical framework for our ideas, the notions we ultimately aim towards and some very preliminary results of a small experiment along these lines.

1. Introduction

Machine Translation Evaluation (MTE) has been more magic and opinion than science. The notion of evaluating MT products results in too broad of a scope for reasonable evaluation – everything from interface, to scalability, to faithfulness of translation, to mean-time-between-failures of the system are fair game for the evaluation of MT systems. Yet, it is necessary to have a method to measure the usefulness of a system to users and equally desirable to point to places where system designers and researchers can improve system outcomes. Bowing to the notion that evaluating MT in a vacuum is like evaluating a sports team that never plays a game, the trend towards task-based evaluation provides guidelines and constraint on what to evaluate, how to evaluate and what context to use for evaluation. The long history of MTE will be described more thoroughly in the next section, but the holy grail is to have an automated evaluation method that is objective, gives reasonable measures of utility and does not rely on casts of thousands to reproduce. Therefore, we look for ways to constrain and decompose evaluation so that it provides measures that are both meaningful to developers and users and ones that indicate not only where systems will be useful but also how they can be improved.

If we consider the history of MTE, we are coming full-circle by looking at the evaluation of language learners as a source for techniques in MTE. Language learner evaluation has had a similarly checkered career – methods for accurately measuring language competence have changed to reflect trends of pedagogy and computing ability. Language learner evaluation research, however, has developed some simple tests which have shown strong correlations to language ability and are good indicators of language competence. These are exactly the kinds of measures we are seeking for

MT evaluation. The history of language learner evaluation and an outline of the principles which may be applied to MTE will be discussed in section 3.

Taking these two ideas in concert, then, we begin a program of looking at the utility of applying language learner evaluation strategies to MTE. The first step in this will be described in this paper. The next section addresses MTE research and why it has been a difficult challenge. The following section briefly highlights the evaluation of language learners and language skills. After that, we describe an initial experiment which will help the process of determining the granularity of measure appropriate for automating MTE. Finally, we will discuss the results of the experiment and look to future tests which may prove useful.

2. Overview of MTE Research

Machine Translation (MT) Evaluation (MTE) is a long-standing issue with many approaches and formalisms having been proposed throughout the years. What to evaluate, how to evaluate and what context to use in evaluation are problematic issues. Unlike some other Natural Language Processing (NLP) problems, there is no gold-standard evaluation possible. This lack of "ground truth" makes the task of automating evaluation even more challenging. The lack of agreement on the assessment of what makes a good translation, even when human translators are involved, hampered initial efforts in MTE, which compared the output of systems to renderings produced by professional translators (ALPAC, 1966; DARPA, 1994). The results of tests of adequacy, informativeness, and fluency as performed on system output were compared to the results of those performed on the human renderings. While this notion of focusing on the outcome of the translation process is a reasonable one, the implementation of the tests proved difficult and somewhat detrimental to the field. In order to assemble the amount of data necessary, such MTE programs were expensive, time-consuming and human-intensive. Requirements listed below multiplied the cost in dollars, time and human involvement exponentially:

(1) expert renderings for each of the input texts
(2) several tests performed on each system's output
(3) testing by several individuals for each criterion evaluated
(4) diagnostic tests, performed by language experts, of each system's output (DARPA, 1992)
(5) the production of back translations from English for systems handling non-English input (DARPA, 1992)

Moreover, these programs measured only one broad aspect of the translation output at a time. Developers were left with little to go on in the way of help to improve their systems and users were left with little which would help them select an appropriate system to meet their requirements. For example, one finding of the DARPA 1994 evaluation was that larger knowledge sources were correlated with better performance (White, 1994) – a useful piece of information in a general sense, but not particularly helpful for specific system designers or users.

Even before the DARPA studies were completed, there was a sentiment in the community that perhaps black-box evaluations looked at the glass as half-empty

rather than half-full (Church & Hovy, 1993). Since then, plans for large-scale evaluations have become more functionally oriented. For example, the MT Scale plan (White & Taylor, 1998) sought to associate the diagnostic scores assigned to the output used in the DARPA evaluation with a scale of language-dependent tasks such as scanning, sorting, and topic identification. Linking the breakdown in a user's language-based performance of a function to some phenomenon in system output extended the usefulness of this approach (Taylor & White, 1998). Similar types of associations were explored even further with experiments in correlating systems' handling of a set of text features with users' performance on information processing tasks (Vanni, 1998) and measuring a system's performance on new text types (Povlsen, et al., 1998). Consideration of variables such as the function of MT output and the complexity of MT input continued to be explored by researchers with the recognition and description of the role of the user's purpose and process (Hovy, 1996).

The direction of these endeavors seems to be toward streamlining the evaluation process and equipping users with tools for carrying out their own evaluations, assessments of MT systems which are tailored to what the user requires from the MT system output. One feature of any such test will be a description of what linguistic features the system can handle reliably. Another possibility suggested during the time of functional evaluation was to look at the language models developed for language acquisition, particularly second language acquisition (SLA) errors (Connor-Linton, 1995). Research in SLA and also cognitive skills development provides us with a potential model for identifying a constellation of such features useable diagnostically to characterize the performance of a system.

3. Models of Language Learner Evaluation

Like MT, language learner evaluation has gone through a long and varied history – a reflection on pedagogical, cognitive and other changes in language learning development. Yet, it is this long history that may yield useful ideas in MTE – as we understand the language learning process better, we have developed measures of what it means to "know" a language. These measures, and the insight into language sills they provide, will lead to useful methods for measuring system abilities and, hopefully, illustrate the ways in which system performance can be improved. Before we discuss the first in a series of experiments to demonstrate this, we will highlight some aspects of language learning and learner evaluation.

Language teaching in the 18th and 19th centuries focused on the form of language rather than the function of it. Features of a language such as grammar rules and vocabulary lists were taught such as the long tables of Latin conjugations. These were memorized and translations were of texts that had existed for centuries. Greek and Latin, the primary languages taught, were not in use beyond academics and their study reflected this lack of contextualization. Some of the principles of this form of teaching exist in language pedagogy today as reflected by the drill-and-practice exercises that still proliferate.

With the expansion of language learning and the idea that language learning could provide benefit beyond academic exercise, there was a movement based on the idea that one could demonstrate a certain useful command of a language without knowing which prepositions take dative form or even what dative means.[1] This was the beginning of language learning in context which viewed foreign language abilities as they developed and identified ways of teaching useful language skills without as much emphasis on language mechanics. The "communicative language approach" (Asher, 1977) caused a reevaluation of teaching and testing methods – a trend away from the traditional, rote methodologies occurred. Instead of testing conjugation with the filling in of tables, learners were evaluated on their abilities to communicate a given point in a given situation with differing levels of sophistication.[2]

Neither trend – rote learning or totally communicative learning – is sufficient to support all levels of language learning. Additionally, neither reflects what we know about language acquisition, particularly second language acquisition. In recent years, the trend is to view language acquisition as a continuum of related skills that build upon each other. Evaluation of language learners focuses on measuring competence and performance while supporting a model of which languages features are learned in what order. One popular theory that has been computationally useful is the notion of Zones of Proximal Development (ZPD) as described in Michaud & McCoy (1999). In this theory, language learning is seen as a scaffolded series of abilities where some abilities are at the same level and others are needed to reach the next level of development. Specific aspects of language can be tested and the correlation between these tests and the level of the student is good. This notion draws on the drill-and-practice testing methods developed under early language teaching methodology, but also attempts to also characterize the ability of the student to use language effectively. Another measure growing from language development research is the Interagency Language Roundtable (ILR) scale (Child, et al., 1993) used to assess government linguists. For purposes noted later, we chose the ILR scale as the first measure for our experiments.

At this point, we will discuss the commonalties between testing of language learners and MTE which lead us to the series of experiments we are enacting. The most attractive feature of learner evaluation is the multitude of automated tests, both standardized and non-standardized, which exist. If we can draw a correlation between the language skills these measure and the language capabilities translation engines provide, we have a less human-intensive measure for translation engines. Additionally, these tests can inform translation engine developers about the kinds of language features which could be improved for better translation quality. The objection may be raised that measuring language learning skills is not the same as measuring translation ability. Generally, though, advanced foreign language handling skills are associated with the students' ability to translate as well. That is, to

---

[1] Quick – can you define dative? As defined (Crystal, 1992): "the dative mainly affects nouns, along with related words (such as adjectives and pronouns), and signals a range of meanings typically expressed in English by the prepositions *to* or *for*…"

[2] Interested readers are directed to Levy, 1997, for a more detailed description of language teaching evolution.

understand what language is to be used in a situation and generate appropriate responses.

4. Experimental Design

At the roughest grain, we can look at the grading of MT system outputs as if they were language learners. While we recognize that this is too coarse a grain to provide much in the way of meaningful indicators of usefulness or areas for future development, it is a starting point to give a baseline of these measures. To this end, we found a widely-used set of criteria for evaluating foreign language students which focuses on the coherence and competency of the produced text. The ILR scale (Child, et al., 1993) identifies both levels of language competency and also the kinds of tasks and kinds of materials which might be mastered by a student at each of these levels. Table 1 roughly describes some each level.

In our preliminary experiment, we identified five 100 word foreign language texts (examples in Tables 2-6) of different complexity levels. We then produced expert translations of each of these texts. Following this, we submitted the initial text to two (or more) MT systems and then applied the ILR scoring methods on a 100-point scale. Table 7 shows the grading scheme for scoring.

Level 0+	**Level 1**	**Level 2**	**Level 3**	**Level 4**
Survival	Orientation	Instructive	Evaluative	Projective
Traffic signs	Forms	Instructions	Analyses	Think-pieces
Calendars	Menus	News reports	Critiques	Commentary

Table 1. ILR Rating Scale

Various	Expositions	Displays	Conferences
Reading	Cinema	Music	Opera
Circus	Theatre	Dance	Meetings
Cultural	Appointments		

Table 2. Level 0 Translation Exercise Example

Hello, how are you?	What's your name?
Very well, thank you and you?	My name is Jeanne?
Everything is going well.	And them?
Have a nice day.	Their names are Jacques and Jules.

Table 3. Level 1 Translation Exercise Example

Bill Clinton was awakened Friday, December 31, at 5:00 in the morning to learn of Boris Yeltsin's decision to leave office.

According to the White House spokesperson, in the course of a 20 minute conversation, the outgoing Russian president stated to his American counterpart that the Russians will remain faithful to their constitution, to democracy, to arms control and to the market economy.

Table 4. Level 2 Translation Exercise Example

The annual report of the Inter-Ministerial Mission Against Sects (IMLS), which was expected by the end of the year, will not be submitted to the Prime Minister before January 15th.

Officially, the delay would only be due to unimportant technical adjustments.

In reality, the advisors hope that the IMLS reviews its copy while correcting the wording of certain anti-sect proposals in order to avoid exaggerated reactions to the diplomatic plan.

Table 5. Level 3 Translation Exercise Example

Life goes on. The news is not good. That's probably what one must call the domino effect which is nothing other than the news. The hull cracks here, the weather there and the cold front continues. Reason finally snaps regarding this Cuban child. Because here we have this little one traveling in the belly of the political whale from now on, a symbol of the clash among those who diplomatically seek to create happiness for children through that of nations. We have to save the child, a soldier of an American-Cuban guerilla war. The child has become a hostage in the struggle which pits the Republican-majority American congress against its president and his attorney general, Janet Reno. The child is summoned to appear before the State Court of Florida on February 10th. Total absurdity. Liberty has a strong back which transports everywhere in the global village the image of a laughing child and you see how the breath of liberty is good for his complexion and his smile. One imagines the hearing if it's necessary to have one someday. Then, my little one, speak without fear. Who do you prefer? Your father or Liberty? You are free to decide – all your orphaned liberty or your six years under the influence.

Table 6. Level 4 Translation Exercise Example

Error Type	**Points Subtracted**
Major syntactic errors significantly altering the meaning	4
Minor syntactic errors causing meaning distortions	3
Lexical, grammatical affixation	2
Stilted usage, disfluencies	1

Table 7. Error Assessment Scale

The first stage of the experiment is the selection of materials. These were randomly selected from newspapers used for the teaching of French translation. Second, the materials were run through three translation engines of varying degrees of sophistication and completeness.[3] The resulting translations were then given to a teacher for scoring. We are reporting on this scoring, understanding that the next step is to have other teachers score the materials.

5. Results and Discussion

The most interesting initial result is that the levels of learning on which the machine translation engines performed best were levels 02 and 03. This was consistent across all translation engines, regardless of methods of development. Levels 02 and 03 represent full sentences with developed grammars, but without inference and descriptive analytical power. This is not surprising for a number of reasons. First, translation engines work best on well-formed input text. Levels 0 and 01 represent many stages of ill-formed or under-developed language use. Level 04 represents a level of sophistication and cross-sentence processing that most translation engines do not possess. More detailed analysis will be necessary to determine which specific language features of these text levels make them more amenable to translation engines or automated processing.

Another criticism that may be leveled is that this still represents a human-intensive evaluation technique for MT. Especially when more than one teacher is needed for scoring and, to be complete, we would need translation students exercises mixed in with translation engine outputs. We recognize this and hope to use the results not to develop a new human-intensive evaluation methodology, but to show us if the language learning track is worth pursuing for MTE.

6. References

Asher, A. 1977. *Learning Another Language Through Actions: The Complete Teachers Guidebook*. Los Gatos, CA: Sky Oaks Productions.

Child, James, Clifford, Ray, and Pardee Lowe, Jr. 1993. Proficiency and Performance in Language Testing. *Applied Language Learning*. 4:1-2. 19-54.

Church, Ken & Hovy, Eduard. 1993. Good Applications for Crummy Machine Translation. *Machine Translation* 8. 239-258.

Connor-Linton, J. 1995. Cross-cultural comparison of writing standards: American ESL and Japanese EFL. *World Englishes*, 14.1:99-115. Oxford: Basil

Crystal, D. 1992. *An Encyclopedic Dictionary of Language and Languages.* Blackwell Publishers, Oxford, UK.

[3] While we will not specifically name them here, two are commercially available and the third was a government developed system.

Hovy, Eduard. 1994. Why Core Technology Evaluation Doesn't Work. Talk given at the Second Conference of the Association for Machine Translation in the Americas. Montreal, Quebec, Canada.

Levy, M. 1997. *Computer Assisted Language Learning : Context and Conceptualization*. Oxford University Press.

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

Pierce, J., Chair. 1966. *Language and Machines: Computers in Translation and Linguistics*. Report by the Automatic Language Processing Advisory Committee (ALPAC). Publication 1416. National Academy of Sciences National Research Council.

Povlsen, Claus, Nancy Underwood, Bradley Music, and Anne Neville. 1998. Evaluating Text-Type Suitability for Machine Translation a Case Study on an English-Danish System. *Proceedings of Language Resources and Evaluation Conference, LREC-98*, Volume I. 21-27. Granada, Spain.

Taylor, Kathryn B. and John S. White 1998. Predicting what MT is Good for: User Judgments and Task Performance. *Proceedings of Third Conference of the Association for Machine Translation in the Americas, AMTA98*. Philadelphia, PA.

Vanni, Michelle. 1998. Evaluating MT Systems: Testing and Researching the Feasibility of a Task-Diagnostic Approach. *Proceedings of the Conference of the Association for Information Management (ASLIB): Translating and the Computer 20*, London, England.

White, John S. and Kathryn B. Taylor. 1998. A Task-Oriented Evaluation Metric for Machine Translation. *Proceedings of Language Resources and Evaluation Conference, LREC-98, Volume* I. 21-27. Granada, Spain.

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