Task-Based Evaluation for Machine Translation

Jennifer B. Doyon, John S. White
Litton PRC
1500 PRC Drive
McLean, VA 22102
(doyon_jennifer,
white_john}@prc.com

Kathryn B. Taylor
QSI Inc.
8201 Greensboro Drive
Suite 1200
McLean, VA 22102
taylorkb@erols.com

Abstract

In an effort to reduce the subjectivity, cost, and complexity of evaluation methods for machine translation (MT) and other language technologies, task-based assessment is examined as an alternative to metrics-based in human judgments about MT, i.e., the previously applied adequacy, fluency, and informativeness measures. For task-based evaluation strategies to be employed effectively to evaluate language-processing technologies in general, certain key elements must be known. Most importantly, the objectives the technology's use is expected to accomplish must be known, the objectives must be expressed as tasks that accomplish the objectives, and then successful outcomes defined for the tasks.

For MT, task-based evaluation is correlated to a scale of tasks, and has as its premise that certain tasks are more forgiving of errors than others. In other words, a poor translation may suffice to determine the general topic of a text, but may not permit accurate identification of participants or the specific event. The ordering of tasks according to their tolerance for errors, as determined by actual task outcomes provided in this paper, is the basis of a scale and repeatable process by which to measure MT systems that has advantages over previous methods.

1 Introduction

It is by now well known that machine translation (MT) evaluation is significantly different from evaluation of other language processing technologies, because of the fact there is no single "right" translation of any expression, and thus no single ground truth for comparison. In general, the techniques devised to mitigate this difficulty have involve: using target-native speakers to make judgments about the fidelity and intelligibility of MT output (White et al. 1994; White 1995; Doyon et al. 1998). These methods have a disadvantage of requiring large numbers of judgments, texts, and raters to control against biasing effects.

Meanwhile, the context of MT (and all language processing) has shifted in the last decade from the presumption of turnkey, single function systems to end-to-end production environments that integrate multiple automatic language processing systems into a single process flow. While the intelligibility and fidelity measures remain valuable in comparing MT systems, they do not directly indicate the contribution (or degradation that an MT system makes in the context of other processes such as topic detection, information extraction, gisting, summarization, and so on. New evaluation methods must be able to take optimum advantage of human judgments without the heavy resource requirements, and at the same time measure something useful for the new context of operation.

While an evaluation strategy is now a required part of the development of any technology, there are few standards available which measure either the progress or usefulness of any individual MT system. Self-reporting by system developers is not always a cost-effective or reliable method to measure the progress of a research project and may divert resources away from research. There is also a difference in the motivation for the internal testing that system developers conduct, and the internal diagnostic or other procedures employed do not address the same kind of performance issues that the community at large has.

However, there must be a clear indicator of progress during system development and a measure of what benefit the technology provides when operational. The various participant groups interested in MT technology (theoreticians, programmers, managers, contract officers, etc.) all have different evaluation requirements. It is crucial to evaluate any technology with objective metrics that answer questions that prospective users of the technology have, usually "how does the use of this technology benefit me?" Practically speaking, the answer must translate into a reliable estimate of cost savings realized from technology use.

There is also no straightforward way to relate such measures such as adequacy, fluency, and informativeness, long-accepted metrics for MT, to the benefits an MT system might provide in the working environment described above. There is an immediate need for a clearly documented approach for applying measurements to technologies for which there is no "right" output, just a number of human judgments of performance, which vary in their consistency. For example, text summarization is a technology that shares many evaluation issues with machine translation.

As mentioned above, the key to a meaningful evaluation of any technology is creating the proper context for the technology. This translates to the ability to envision a process to which the technology benefits in some way. It may be faster, more accurate than an existing technology, in which case it can be measured against what it replaces. In the case in which a totally new capability is being developed, the measurements taken become a baseline. A context is then developed for the capability, i.e., what are its major characteristics and whom does it benefit?

We are used to seeing technology products reviewed side-by-side, consumer reports style, with common features such as price or "click-to-clunk" performance side-by-side, with some common operational context (each system is put through the same paces, i.e., given some objective to achieve) to produce any performance measurements. This approach seems very sensible, judge a technology's performance by how it works under actual conditions. The key is to develop a set of conditions and operational tasks which users will recognize and relate to as tasks that they, too, will be performing in their normal use of the system.

Successful task-based testing requires:

- a comprehensive knowledge of user needs
- knowledge of what objectives the technology is meant to achieve
- the purpose of the evaluation

as well as:

- selection of a suitable task(s) for testing
- test construction
- ground truth
- creation of a convenient test forum
- ongoing validation (does the data gathered support a determination of whether the technology accomplishes its objective (Norris 1999).

A task-based assessment of technology becomes an option when a well understood model of the uses to which the technology will be put exists, and the model has been generalized into classes of tasks. An objective metric upon which to base these answers does not exist for MT, and work has been ongoing to devise that metric.

In recognition of both the seeming applicability of second-language learning assessment tools and the trend away from subjectively defined and scored performance scales the US Federal Intelligent Document Understanding Laboratory (FIDUL) has been developing a new metric to rate MT by which other language processing tasks it may facilitate. (White and Taylor, 1998, Taylor and White, 1998). This MT Functional Proficiency Scale project has collected results of users performing a variety of exercises with raw Japanese-to-English MT output, to discover which translation problems make output less useful for other language processing tasks. These translation problems are then collected in a simple diagnostic test set which can be run on any Japanese-to-English MT system. This diagnostic test set will indicate the suitability of an MT system's output for any of the "downstream" tasks in the user's production process.

2 Proficiency Scale Development

Different language processing tasks will be variously able to use degraded text typical of contemporary MT systems. For example, "filtering," the act of discarding irrelevant texts based on a very cursory examination of the title and a small sample of the body, is likely to be effective even with relatively poor MT output. Scientific editing, by contrast, must have a professional-level translation to work from. This "task tolerance" is the key to measuring MT output for the functional proficiency scale. Before beginning data collection with users, a hypothetical ordering of text-handling tasks by tolerance (from least tolerant to most tolerant of translation errors) was assembled and can be seen in Table 1 below.

Task	Description
Publishing·	Produce a technically correct document in fluent English
Gisting	Produce a summary of the document
Extraction	For documents of interest, capture specified key information
Triage	For documents determined to be of interest, rank by importance
Detection	Find documents of interest
Filtering	Discard irrelevant documents

Table 1 - Preliminary Ranking of Test-Handling Tasks

2.1 Text-Handling Task Exercises

The first step in developing the proficiency scale is to determine this "tolerance" order, which is elicited by a series of exercises performed by people who perform monolingual text handling tasks with translated material in their ordinary work. With such an order established, it is then possible to predict that MT output suitable for a particular task is also good enough for tasks requiring lower quality text (i.e., are "more tolerant"), and not useful for tasks requiring higher quality text ("less tolerant").

During this series of exercises, users have performed three separate exercises for the purpose of eliciting judgments about the usefulness of particular translated text. Two of the exercises elicit the usefulness of a set of translated texts for a task that a user typically performs, and a third helps to indicate the translation phenomena that actually affect the usefulness of the texts.

- In the *"Snap Judgment"* exercise, users make quick, intuitive judgments about a body of translated texts, collecting those that the user believes might be of sufficient quality to be of further use in the text handling task they typically perform. The user sorts 15 English translations of Japanese newspaper articles into three groups (those that they could use, those that they might use, and those that they cannot use). Correlation among users who perform the same tasks will suggest a preliminary ranking of tolerance.

- *"Task-Specific"* exercises have users perform a particular activity that resembles the actual task they do in their regular assignments. The following are examples *of some of these exercises:*

- *Filtering.* The user is given a set of 15 texts and a subject area (e.g., "crime"). The user groups each text according to whether the text is directly relevant to the subject area, irrelevant to the subject area, or of unknown relevance (because the text is unintelligible). Whether a text is grouped correctly or incorrectly is an indicator of the text's usefulness for filtering tasks. The measures of recall and precision will be used. For example, of the texts related to the subject area, what percentage did the user identify correctly (recall). and of the texts the user identified as being related to the subject area, what percentage were actually related (precision). The precision and recall metrics will establish the task tolerance level of filtering.

- *Detection.* The user is given a set of 15 texts and three subject areas (e.g., "crime," "economics," and "government and politics"). The user groups each text according to whether the text is directly relevant to the category of "crime," relevant to the category of "economics," relevant to the category of "government and politics," irrelevant to all three categories, or of unknown relevance (because the text is unintelligible). Whether a text is grouped correctly or incorrectly is an indicator of the text's usefulness for detection. The measures of recall and precision will be used. For example, of all the texts related to each of the three subject areas, what percentage did the user identify correctly (recall), and for all those identified by the user as being related to a subject areas, what percentage were actually related (precision). The precision and recall metrics will establish the task tolerance level of detection.

- *Triage.* The user is given 15 documents grouped into three sets of texts: crime, economics, and government and politics. The user organizes each group in terms of the most germane to a specific problem statement (e.g., "rank the texts in terms of their relevance to criminal conspiracies"). The results of the triage task will be scored by comparing the users' ordering of the texts against a previously generated ground truth ordering of the same three text sets. As with filtering and detection, ranked texts (as well as texts of too poor a quality to be ranked) will identify the tolerance level of text triage activities.

- *Extraction.* Users perform the extraction exercise by identifying named entities in texts. In this exercise, users mark the elements of the text that fill particular information slots (e.g., persons, locations, organizations, dates, etc. - this task is based on the 2Named Entity" task of the U.S. government's Message Understanding Conference (MUC). The results will be scored according to the number of correct fills as compared to the number of fills identified in ground truth created from the expert human translations of the same text. As with the other exercises, the percentage of correct fills will establish the task tolerance of extraction.

- *Gisting.* The bracketed version of the DARPA expert human translation used to evaluate the adequacy of a translation is edited so that only those brackets that contain information relevant to a summary of the document remain. The user is asked to apply the DARPA l-to-5 adequacy scale ratings to indicate to what extent the information in the selected brackets is present in the aligned paragraphs of a translated version of the text. As with the other exercises, text ratings will help to indicate the task tolerance of the gisting task.

- The third exercise in the set, *"Rating Reasons, "* has users identify phenomena in the exercise texts that they found particularly problematic. This exercise helps to focus certain phenomena as diagnostic, i.e., that problems of those sorts constitute the difference between being able to use a translation for a particular task or not.

These exercises have the effect of arranging the output texts by the tasks that can be done with them. Each text is identified as being suitable for certain tasks and not suitable for others or for none of the tasks. Since every task-specific exercise uses the same text corpus, the arrangement of texts resulting from these exercises implies the tolerance level of each of these tasks. The result is the order of text handling tasks by task tolerance. New MT output is measured on this tolerance scale. If the output can be determined to be suitable for one task, then it is also suitable for all the tasks that are more tolerant, and none of the tasks that are less tolerant. It may be that there is not a single ordering of tasks. For example, document detection may be less tolerant than extraction for certain types of data, and more tolerant for others. However, the general principle remains as long as there is convergence of all possible orders (e.g., that publication is always less tolerant than filtering).

2.2 Results

Execution of the three exercises of Snap Judgment, Specific-Task, and Rating Reasons began in March 1999 and was completed in May 1999. Seventeen users from varying analyst groups across U.S. government agencies completed the five exercises of filtering, detection, extraction, triage, and gisting. Findings from the Snap Judgment exercise seem to show that a ranking of text handling tasks does exist. This ranking is a result of an analysis performed over each task group's (each group containing a number of users performing the same task) judgment of which of the 15 texts in the Snap Judgment corpus were of a tolerance level appropriate for their task. Table 2 shows the resulting scores and ranking of text handling tasks by tolerance.

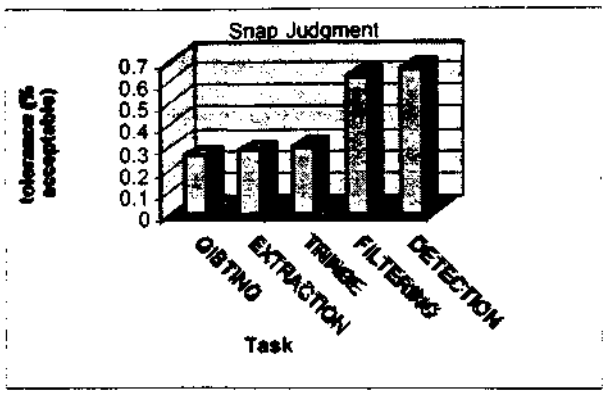

Table 2 - Snap Judgment Results

The task-specific exercise findings (results analyzed for each individual in a user group) do not completely correspond to the results found in the Snap Judgment analysis. The results of the individual task-specific exercises appear to support most, but not all, of the ordering of the Snap Judgment's ranking of the text-handling tasks. Table 3 shows the task-specific scores and ranking of text-handling tasks by tolerance of translation errors.

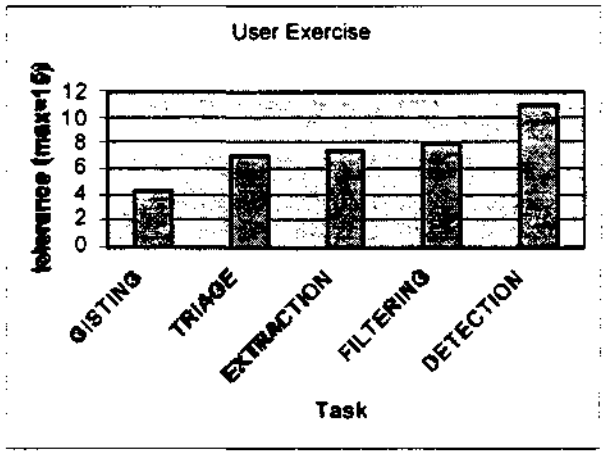

Table 3 -Task-Specific Exercise Results

Results from the Snap Judgment and Task-Specific exercises are surprising because both suggest that the detection task is more tolerant of translation errors than the filtering task, contrary to the original hypothesis. Additionally, the results from the two exercises suggest that the ordering of the triage and extraction tasks be exchanged. Further analysis of data collected from users and verification of ground truths for all exercises is underway in order to validate these findings.

3 Diagnostic Test Set Development

The next step is the development of a diagnostic test set that will enable the mapping of new MT output onto the tolerance scale. This is accomplished by analyzing certain translation phenomena (linguistic, lexical, formatting, punctuation, etc.), categorizing them, and representing them in a set of Japanese test patterns. The problem of which phenomena to extract as diagnostic is accomplished by isolating which phenomena seem to make the difference between a mapping at one tolerance level and another. When analyzing the exercise results, there will be "border texts," i.e., texts which are just good enough for one task and not quite good enough for the next, less tolerant, task. The translation phenomena that occur at these tolerance borders are distilled from the texts, and classified according to generally accepted contrastive descriptions of the source and target language. Descriptions of the sort used in language teaching are quite useful here, because they are descriptive, consistent, and not bound to theories germane to particular issues in MT (Connor-Linton 1995).

Source language passages that represent each of the identified translation phenomena are prepared, and included in a simple text diagnostic set. From this point on, MT systems need only run the diagnostics to determine the suitability of an MT system to produce output that is usable for a particular text-handling task.

4 Conclusion

The MT Functional Proficiency Scale project is developing metrics for perhaps the most pressing single issue in MT today, namely, the actual usefulness of MT in an automatic, end-to-end process of multilingual information processing. At the same time, the methodology will incorporate the knowledge gained from human judgments, without the effort and size of such evaluations.

References

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

Doyon, J., Taylor, K. B., and White, J. S. (1998). "The DARPA Machine Translation Evaluation Methodology: Past and Present." Philadelphia: AMTA98.

Norris, John. (1999). "Seminar for Task-Based Assessment for Language Learning." Georgetown University, Washington DC.

Taylor, Kathryn B., and White, J. S. (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, AMTA'98.

White, J. (1995). "Approaches to Black Box MT Evaluation." Paper presented MT Summit, Luxembourg. July 1995. Luxembourg: European Parliament and the International Association for Machine Translation.

White, J., and O'Connell, T.A. (1994). "The ARPA MT evaluation methodologies: evolution, lessons, and future approaches." Proceedings of the 1994 Conference. Association for Machine Translation in the Americas.

White, J., and Taylor, K. B. (1998). "A Task-Oriented Evaluation Metric for Machine Translation." Proceedings of LREC-98, Volume I. 21-27.