

Determining the Tolerance of Text-handling Tasks for MT Output

John White, Jennifer Doyon, Susan Talbott

Litton PRC
1500 PRC Drive, McLean, Virginia, USA
{white_john, doyon_jennifer, talbott_susan}@prc.com

Abstract

With the explosion of the internet and access to increased amounts of information provided by international media, the need to process this abundance of information in an efficient and effective manner has become critical. The importance of machine translation (MT) in the stream of information processing has become apparent. With this new demand on the user community comes the need to assess an MT system before adding such a system to the user's current suite of text-handling applications. The MT Functional Proficiency Scale project has developed a method for ranking the tolerance of a variety of information processing tasks to possibly poor MT output. This ranking allows for the prediction of an MT system's usefulness for particular text-handling tasks.

1. Introduction

From the introduction of the field of machine translation (MT), there have been two strongly held impressions: MT output is far from perfect, and there are probably certain tasks that can be accomplished even with bad MT output that might make it worth using. There are some ways to measure each of these impressions. Coverage, intelligibility, and fidelity are measurable attributes that can help determine how far from "perfect" translations are, and various measures of time, cost, or quality improvements have been done with respect to particular business environments. However, a significant gap in evaluation lies in the inability to predict, on the basis of an MT systems output, the tasks for which that output might be useful.

The U.S. government's MT Functional Proficiency Scale project has developed a method for task-based MT evaluation, which has resulted in a scale of text-handling tasks, such as topic detection, gisting, document filtering, and entity extraction, in terms of the ability to perform these tasks with possibly poor MT output.

Development of the scale involved government users who typically perform one or more text-handling functions in their native language, often using translated material. These users were given sets of exercises to elicit the acceptability of English translations of Japanese newspaper articles for each of the text-handling tasks investigated. Ease users found in performing tasks with this output correlates to the tolerance of each of the tasks to MT output. In this paper, we will discuss the identification of the user-performed text-handling tasks, the development and execution of the user exercises, and the computation of the exercise results, which produce the task tolerance scale. We will also offer possible avenues for future research that exploit the lessons learned and tolerance scale resulting from this project.

2. Scale Development

Potential users and other related sources were interviewed to identify and define text-handling tasks.

From these definitions the task-based exercises were developed. These exercises were then used to measure the quality of a corpus of translations. Proficient judgments provided by skilled users helped to validate the subsequent scale.

The corpus used for the MT Functional Proficiency Scale project was a subset of translations from the Defense Advanced Research Projects Agency (DARPA) "3Q94" evaluation. These are Japanese-English translations from a variety of MT systems and expert human control texts. Different sets of this corpus were used for the Snap Judgment and Task exercises.

2.1 Task Identification

Preliminary interviews with experienced text-handling users helped to define text-handling tasks, as well as provide information that would help to determine the task-based exercise for which the user would be most suited.

2.2 Exercises

The purpose of the exercises was to ascertain if users could complete text-handling tasks with adequacy using translations of varying degrees of quality. User judgements made during these exercises indicated to what degree each translation held value for a particular task(s).

2.2.1 Snap Judgment

In this first exercise, each user was asked to make quick, intuitive judgments about the usability of 15 translations. This judgment was to be made with the user's assigned task exercise in mind. Thus, a user who typically performs document detection considered, for each text, whether it could be of use in the detection activity.

2.2.2 Task Exercises

Each user was given one of the five task exercises: Filtering, Detection, Triage, Extraction, or Gisting, corresponding to the task they specialise in or typically perform.

Filtering is the process of discarding text not related to a given topic of interest. In the filtering exercise, 15 translations were sorted into three piles: relevant to the topic of interest, not relevant, or cannot be determined. The same set of translations was used for all of the task exercises.

Detection is the process of sorting texts into various given topics of interest. In the detection exercise, 15 translations were sorted into 1 of 5 topics of interest. A sixth category was provided for translations whose relevance to one of the given topics could not be determined.

Triage is the process of sorting texts, within a common topic of interest, by level of relevance to a given problem statement. In the triage exercise, 15 translations were pre-sorted into 3 topics of interest, then ordered by level of relevance to a problem statement within each group.

Extraction is the process of pulling key words/ information from a text. In the extraction exercise, entities found in 7 translations were color coded with labeled highlighter pens. The entities included: Persons; Locations; Organisations; Dates; Times; and Money/Percentage. Entity guidelines were based upon the U.S. Government’s Message Understanding Conference (MUC) definitions (Chinchor & Dungca, 1995).

Gisting is the process of summarising the key points of a text. In the gisting exercise, 7 translations were judged, on a scale of 5 to 1, by how much meaning of an expert human translated segment of text could be found in the “machine translation” version of the same news text.

2.3 Human Factors

A variety of human factors issues were considered in the development of the exercise sets. For instance, if asked, “Can you do your job with this text?”, a biased answer may follow (Taylor and White, 1998). However, if phrased in a way that makes it clear that it is the translation being judged rather than the user’s ability, a more accurate answer may be returned.

3. Results

3.1 Compilation of results data

As discussed above, users who specialized in at least one of the text-handling tasks were given a set of exercises to elicit their task-related judgments about the usefulness of the texts. In the snap judgment exercise, the users were asked to look at 15 translations and judge each for whether they could be used for the text-handling task each user typically performs. The user sorted each document into simple yes/no categories, and the snap-judgement usefulness of each document was determined by the average of the “yes” responses for the users of each task.

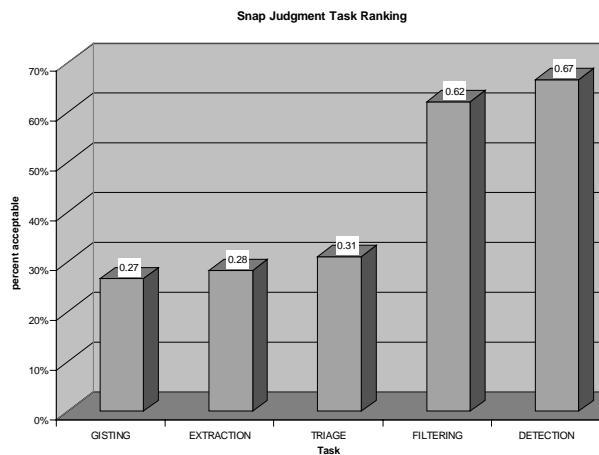


Exhibit 1 shows the ordering of tasks by their tolerance as indicated by the snap judgment exercise.

The task-specific exercises were each computed in terms of the metrics relevant to the particular task: for example, recall for detection and filtering; recall and precision for extraction; and fidelity for gisting. In the detection exercise, recall was computed across documents for each user, for each of the categories (crime, economics, government/politics) into which the documents were sorted. The average recall value was used to determine the acceptability cutoff point within each category, and the number of acceptable documents across all three categories resulted in the percent acceptable for the entire task. The same process was used to determine the acceptability cutoff for filtering, where the documents with acceptable recall for both the “in” and “out” sortings was the acceptability percentage. A similar process was used for extraction, except that an average precision measure was also factored in.

Computing the acceptability for gisting is also similar, except gisting judgments were collected with a fidelity test (namely the "adequacy" measure referred to above); thus each text for each user has an average of the scores, on the 5-1 scale, for the decision points in that text. In turn, the average of these average scores gives the cutoff for acceptability for gisting.

Acceptability in the triage task is measured by comparing ordinal rankings given by the users with ordinal rankings from the ground truth set. For this measure, a uniformity of agreement value was established as the mean of the standard deviations for each text in each problem statement. Then the mean user ranking for each text was compared to the ground truth ranking, plus-or-minus the uniformity measure. A text is acceptable if it matches the ground truth within the uniformity measure.

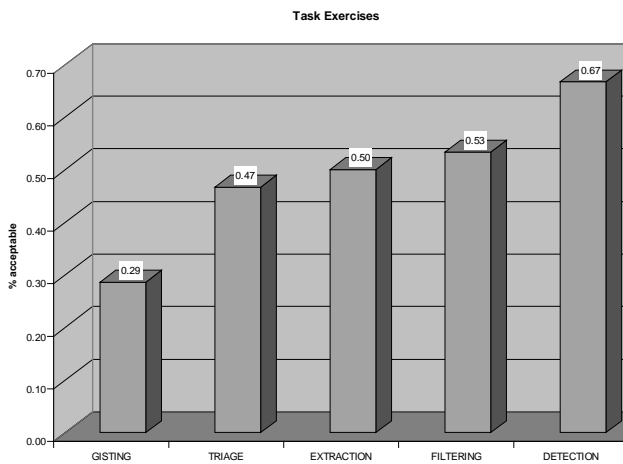


Exhibit 2 shows the tolerance for each task to MT output, based on the task-specific user exercise.

3.2 Analysis

The tolerance charts for snap judgment and task-specific exercises agree in the MT-tolerance order of the text-handling tasks: detection, filtering, extraction, triage, and gisting, in order from most tolerant to least tolerant of MT output. The agreement between the two exercises indicates a consistency of judgment across two views and two different data sets, lending confidence to the tolerance ranking. As an initial heuristic we had hypothesized a different tolerance order in which filtering was the most tolerant and triage much more tolerant than it turns out to be. It appears, from interviews subsequent to the exercises, that both tasks involve a deeper level of reading and association of facts than is often assumed by those who do not perform these tasks on a daily basis for analytical objectives. Further, users who typically perform detection seem more able to use scarcer resources of associations among a small number of key words, even when the noise of poor MT output interferes.

While these findings should be regarded as preliminary pending further validation, they indicate an apparently consistent, single-thread hierarchy of tolerance on which MT systems can be plotted. With this scale, we can facilitate the prediction of whether a particular MT system is suitable for a particular downstream language handling task.

4. Future Research

The MT task tolerance scale can be of immediate predictive use in decisions about MT approaches, system selection, or integration architecture. If the single-thread, deterministic order of the tasks is validated, then one need only establish the least tolerant task for which a particular system is useful. One can then predict that the system is also useful for all of the tasks on the scale that are more tolerant and none of the tasks that are less tolerant.

There are two ways in which the determination of the least-tolerant-acceptable point can be established for an MT system. One way is to replicate the process described here for the discovery of the scale itself, i.e., engaging task-specialist users to perform the snap-judgment and task-specific exercises. With appropriate controls (such as including other translations, ideally from the same language pair) this should identify the area on the scale that the tested system's output is still judged to be of use.

A second way to plot an MT system on the scale involves the distillation of classes of translation phenomena which cause a translation to be less useful for a particular task than it might otherwise be. These translation errors may be linguistic in nature, but may also include formatting problems, representations of dates/numbers, mishandling of character sets, etc. Using pedagogical classifications of Japanese-English error types (Connor-Linton 1995), we have developed a preliminary classification of pair-specific error types into which the errors that actually occur in the test corpus can be categorized (Taylor and White op.cit., Doyon et al. op.cit.). Meanwhile, it is possible to determine which errors make a difference with which task by noting the errors in the acceptable texts at each tolerance level. From these diagnostic errors, categorized into translation error types, it should be possible to develop small test sets of translation patterns which cover the diagnostic errors for each potential downstream task. Any MT system in the same language pair can then use these test sets to determine instantly its suitability for a range of downstream tasks. Future work will focus on the development and validation of this test set and on making it generally available.

5. Conclusion

The MT Functional Proficiency Scale project has served to contribute rigor to one of the remaining intuitive areas of machine translation research, the notion of what sort of language handling task might be tolerant of poor MT output. It is now possible to transcend vaguely expressed goals of MT quality/usefulness with specific characterizations of the hierarchy that expresses the tolerance different tasks have for MT output.

Clearly, the general application of these findings will depend on the ability to extend them to other language pairs and ultimately on the ability to express the diagnostic patterns for any pair in a easily executable form. Ultimately, this method could prove to be a standard for MT evaluation in the modern context of task-oriented, fully integrated automatic processes.

6. References

Chinchor, Nancy, and Gary Dungca. (1995). "Four Scorers and Seven Years Ago: The Scoring Method for MUC-6." Proceedings of Sixth Message Understanding Conference (MUC-6). Columbia, MD.

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

Doyon, Jennifer, Kathryn B. Taylor, and John S. White. (1999). "Task-Based Evaluation for Machine Translation." *Proceedings of Machine Translation Summit VII '99*. Singapore.

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.

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.