[*First International Conference on Language Resources & Evaluation*, Granada, Spain, 28-30 May 1998]

A Task-Oriented Evaluation Metric for Machine Translation

John S. White and Kathryn B. Taylor
PRC Inc.
and the
Federal Intelligent Document Understanding Laboratory (FIDUL)

McLean, VA USA
[{white_john | taylor_kathi}@prc.com]

Abstract

Evaluation remains an open and fundamental issue for machine translation (MT). The inherent subjectivity of any judgment about the quality of translation, whether human or machine, and the diversity of end uses and users of translated material, contribute to the difficulty of establishing relevant and efficient evaluation methods. The US Federal Intelligent Document Understanding Laboratory (FIDUL) is developing a new, task-oriented evaluation metric and methodology to measure MT systems in light of the tasks for which their output may be used.

This paper describes the development of this methodology for Japanese-to-English MT. It includes a sample inventory of the tasks for which translated material is used, (e.g., filtering, detection, extraction) and describes exercises in which users perform each task with MT output. The methodology correlates the recorded subjective judgments of the raters in the DARPA MT Evaluation with users' performances on the task-based exercises. Analysis of the errors in scored texts determines whether the presence of certain error types in MT affects specific tasks and not others. Source language patterns that produced errors become a test set that can be easily and efficiently scored to evaluate the performance of any new Japanese-to-English MT system in terms of the task inventory.

Introduction

In recent years the context envisioned for machine translation has changed dramatically, as it has for all text-handling technologies. MT has become part of a larger information-handling process, rather than a standalone activity. The contemporary context requires that MT be integrated into end-to-end processes, which are largely or mostly automated, and for which the evolutionary trend is toward less and less human intervention between the process stages. There is a growing expectation that documents of all forms, including hardcopy, can be merged automatically into a corpus of on-line information, in a homogeneous form and language. The requirements for MT have changed accordingly: the presumption of less human intervention requires a more precise judgment of the capabilities of an MT system to produce output suitable for the next step in the text-handling process.

In these emerging environments, monolingual analysts or other information consumers will perform one or more text-handling tasks using translated material. Each type of text-handling task (e.g., filtering, detection, extraction, summarization, publication) requires translated text input of a certain quality. Some operational tasks may tolerate a wider range in the accuracy and completeness of MT output, while others require near-human accuracy and fluency.

A measure of an MT system's ability to produce suitable output for "downstream" text handling components will address perhaps the most salient question in the contemporary context. The FIDUL MT Proficiency Scale project, currently underway, is developing a reusable, efficient and meaningful predictor of the text-handling tasks that an MT system's output will support.

Evaluation of MT has always been one of the fundamental issues in the field. (Church and Hovy, 1991; Dostert, 1973; Pierce, 1966; Van Slype, 1979). MT evaluation is difficult because there is no one "right" translation, and therefore great difficulty in creating a useful, extensible ground truth for evaluation. Also, MT evaluation has different requirements for the different MT stakeholders (translators, information consumers, managers, researchers, etc.), as well as for different theoretical approaches and language pairs.

The Defense Advanced Research Projects Agency (DARPA) MT initiative within the Human Language Technology (HLT) Program faced these challenges with a series of evaluations in the mid-1990's (White and O'Connell, 1994; 1996). The goal of this effort was to evaluate the core translation algorithms of the sponsored systems, attempting to factor out the wide diversity of theoretical approach, language pair and end-use presumptions. The measures developed in this effort give meaningful results for the precise question of the potential of the core technology of particular translation approaches. The cost, however, is high: to counter the inherent subjectivity of judgments about translation, a large number of translations, controls, raters and decision points must be maintained.

Any new MT evaluation method should be developed to be readily reusable, with a minimum of preparation and participation of raters or subjects. One way to accomplish this is to take advantage of the corpus in which these judgments have already been made, namely, the rated translations of the DARPA series. This corpus consists of several translations each of approximately 400 newspaper articles, originally in French, Spanish, or Japanese. Most of these articles have two professionally translated versions and up to five machine translated versions generated either by sponsored research systems or by commercial MT systems whose developers volunteered to participate. Raters (native English speakers who have no special knowledge of the source languages) made judgments of the translations along three measures:

- Adequacy: the degree to which the tested translation contains information present in a professional translation, measured over sub-sentence syntactic units;
- Fluency: the degree to which the translation meets the expectations of ordinary English intelligibility and flow, measured by sentence;
- Informativeness: the degree to which specific, necessary items of information can be located in the translation, measured in a multiple choice format.

For the purposes of the new metric described in this paper, the DARPA series provides a corpus of multiple variants, whose characteristics have already been measured, usable as a set of controlled samples for user judgments.

The purpose of the MT Proficiency Scale project is to collect user judgments of the suitability of these translations for the tasks they perform, and then to analyze translation errors in light of those judgments, ultimately developing a simple test set of diagnostic patterns.

Development of the MT Proficiency Metric

The development of this measure involves four principal steps:

- identifying the text-handling tasks that users perform with translated material as input
- discovering the order of text-handling task *tolerance,* i.e., how good a translation must be in order for it to be useful for a particular task;
- analyzing the translation problems (both linguistic and non-linguistic) in the corpus used in determining task tolerance;
- developing a set of source language patterns which correspond to diagnostic target phenomena.

The result is a series of patterns which are diagnostic of the difference in tolerance level. This series is then incorporated in a simple test set, which, when applied, will predict for what text-handling tasks a system's output is suitable.

Identification and Ordering of Text-Handling Tasks

Certain analytical text-handling tasks require more accurate and fluent material than others. It should be possible to rank these tasks on a scale from least tolerant (high end) to most-tolerant (low end) of translation errors and omissions. Table 1 shows a hypothetical ranking of tasks from least tolerant (Publishing) to most tolerant (Filtering).

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 Text-Handling Tasks

A ranked order of text-handling tasks such as Table 1 implies that an MT system whose output can facilitate tasks on a particular point on the scale will likely be able to also facilitate tasks lower on the scale, and is unlikely to facilitate tasks higher on the scale. According to Table 1, an MT system that produces a fluent translation (publication quality output) will also support the capture of specified key information from the same translation (extraction). An MT system whose output allows users to recognize that a document is of interest (detection) may not be suitable for capturing all specified key information (extraction) but will perform acceptably for filtering (rapid disposal of irrelevant documents).

In order to develop this metric, it is necessary to identify which text handling tasks can be done with relatively poor quality text, and which require high quality text. Using a variety of interview techniques with U.S. government analysts, who perform one or more text handling tasks as an ordinary part of their jobs, we are establishing an order for these tasks with respect to the quality of text required.

Questionnaires have been designed to capture user observations about the role of translation-supported tasks in their day-to-day work, and the quality of translation they perceive as necessary to accomplish those tasks. In the interviews, users are asked which text-handling tasks (manual or automated) they typically perform, and whether there are other tasks that should be added to the list.

The users are then asked to perform a variety of activities associated with their specific text-handling tasks, using translations from the DARPA corpus.

The intended effect of the interviews with users is to determine, within the context of their individual ordinary activities, whether they could use a particular translation to do their jobs. There are several issues associated with eliciting this judgment. Many such users have a functional mission (e.g., find information about grain production in Europe) rather than the performance of a text handling task *per se.* Since the samples presented are from general news articles, they will not usually be relevant to the domain of their functional task. Secondly, human factors effects can bias the judgment in a variety of ways. For example, the user may be more inclined to respond affirmatively to the question “can you do your work with this document” even when he/she actually cannot (M. Vanni, personal communication).

To address these issues and a variety of others, the interview is broken up into three distinct exercises:

1. *Sorting translated documents by suitability for the text-handling task.* Users are asked to choose from a set of documents those which might be of good enough quality to enable them to do their jobs.
2. *Performing a directed task on a set of documents.* Users are given a task that is similar to, but distinct from, their ordinary text-handling task, and asked to perform this modified task on a set of translation documents. For example, a user who typically performs information extraction is asked to fill in the answers to several labeled slots (“date of action,” “date of report”, “perpetrator”, etc.). A user who performs filtering is told to set aside the documents that are definitely not relevant to a given topic. The directed task validates the judgments of the first, abstracting away from the domain issue (because the specific task is rather different from their actual tasks) and from several human factors issues.
3. *Helping to identify the translation problems which cause a document to be less useful than it otherwise might be.* This exercise is aimed at identifying the translation phenomena that should be included in the MT Proficiency Scale test set. This exercise presents to different persons performing the same text-handling tasks several versions of translations, in which certain portions have been "fixed" (i.e., expert translations have been substituted for those portions). The portions selected for correction are those in which known translation problems can be readily isolated. Sufficiently controlled for pre- and post-test effects, this exercise will help in the identification of problem categories which may align along the order of tasks.

Of course, no one group of users will generally require or have expertise in every one of the text-handling tasks that uses translated material as input. Thus the ranked list will be a merged set over a variety of user groups.

The result of these exercises will be a characterization of the relevant text-handling tasks, ordered by their tolerance to the quality of MT output. It remains to be established whether to expect a single ordering of the text-handling tasks (e.g., document detection is always more tolerant than extraction), or a non-deterministic order (detection is sometimes less tolerant than extraction in different subject domains, extraction requirements vary, etc.). It appears, however that even a multi-path ordering, once described, will suffice for the MT Proficiency scale, as long as there is convergence at either end (a reasonable assumption - topic filtering must always be more tolerant than technical editing).

Correlation of MT Output Properties to Task Scale

An ordering of text-handling tasks of the sort described above will make feasible characterizations of MT systems by the tasks which their output facilitates. If it is possible to predict the least tolerant text-handling tasks that a system's output can facilitate, then we will also know that the output is sufficient for all of the more tolerant tasks. The same texts from which the ordering can be inferred also provide evidence of translation problems which indicate the boundary between acceptability for one text-handling task and another.

Developing the diagnostic test that will make that prediction involves identifying the correlation of corpus texts to the task hierarchy, distilling translation problems that appear to be “diagnostic” (i.e., appear to mark the difference between a text being at one level rather than a higher one), and then characterizing those translation phenomena in a compact pattern for the ultimate diagnostic test.

The process of eliciting from users the effects of certain translation phenomena, discussed above, is the focal part of the identification of diagnostic translation errors; however, a complete and careful analysis of errors in the corpus used in the task-based exercises will also be necessary.

The phenomena encountered are categorized in accordance with established contrastive principles of Japanese and English. These contrasts are described in pedagogy ((Connor-Linton, 1995) is an excellent example). Use of the pedagogical models has the advantages of exhaustiveness and descriptive adequacy apart from issues in MT theory. However, there are many other phenomena that are not described in those treatments, since they are unlikely to occur in human translation; trivial but ubiquitous examples are character conversion errors and punctuation in numerals.

Characterization of the translation problems depends on the parallelism of the DARPA corpus (especially the fact that there are two expert translations of every text), and also on the fact that the texts have already been rated in the original DARPA evaluations, at the sentence and even sub-sentence level. These ratings provide a clue to the location of the errors, providing places for focusing analysis.

Note that we analyzing surface effects, rather than causes (for example, an unknown word can cause a missed parse, which causes many surface problems to occur). We are assessing only surface problems without trying to reconstruct their original cause, for two important reasons: (1) the actual cause of errors is unknown to those who are doing the classification; and, (2) from a user's perspective, the causes are generally irrelevant to whether a system can produce output adequate for a particular task (assuming, as this test will, that user-adaptable features of a system have been optimized).

Sample diagnostic examples drawn from the DARPA MT evaluations are provided in the tables below. They are representative of problems in MT which might affect an extraction task: problems with numerical expressions, time expressions, choice of prepositions, pronouns and word sense selection.

Expert	Score	Sample MT Output
[There were approximately 30 students in the class, of which 70% were Japanese.]	1	Auditing raw a little more than 30 with the person, the inside seven tenths was the Japanese person.

Table 2 Problems with Numerical Expressions

Expert	Score	Sample MT Output
[She has been in Indonesia as an exchange student since August 1990.]	2	From the year 199 X year August while studying abroad, seeing during the life experience of their locale, it has
[reporting on what she has seen and heard through her daily life experiences in that country.]	3	reporting the fact that that etc you thought to Indonesia. Every week it serialization as principle.

Table 3 Problems with Time Expressions

Expert	Score	Sample MT Output
[What is fascinating with Lesley Glaister is the pleasure she takes]	2	What fascinates by Lesley Glaister, is the fun as it takes
[All of this in the sparest, most concrete, style.]	3	All this in the quietest and most pragmatic style,
[that which suits best domestic tragedy.]	2	someone who is the best in the domestic incidents.

Table 4 Problems with Choice of Preposition, Pronoun, and Word Sense

Preparation of Diagnostic Test Suite

The identification of diagnostic phenomena proceeds roughly along these lines: the distribution of texts by their acceptability for particular text-handling tasks, as judged by users who routinely perform those tasks, will also have clusters of translation problems of particular types. Many problems of many types will be in the set found to be acceptable for the most tolerant tasks (e.g., filtering). Fewer problems of each type, and, more significantly, fewer types of problems, will appear in texts rated acceptable for the less tolerant tasks (e.g., extraction). The direct identification by users of potentially diagnostic phenomena provides keys to the identification and weighting scheme of translation problems. Ideally, the point on the text-handling task order where problems of a particular type disappear from the acceptable texts means that errors of that type are diagnostic at that juncture. Patterns representing such problem types are therefore represented in the ultimate MT Proficiency Scale test set.

Having classified translation problems into categories, and having determined which problem categories appear to be diagnostic at which level on the text-handling task order, the actual MT Proficiency Scale test can be developed. This process involves creating Japanese patterns which correspond to diagnostic types (or simply using exemplars from the corpus, where these display one problem category rather than many, as is often the case). Many example patterns of many categories are developed into a simple text file, along with the lexicon for the patterns, so that systems with lexical development can train optimally for the MT Proficiency Scale test.

When run against any Japanese-English MT system, the output of the translation of this simple file will be scored by comparing the output of the patterns against expert translations of them. Clearly, such scoring is a subjective assessment, since, as noted above, there is no one “right” translation. Minimizing the subjectivity of these judgments remains an issue. However, the patterns will be designed so that a judgment about the fidelity and intelligibility of the translation of a pattern should be straightforward.

The scoring should reveal a cutoff between a pattern series for which the trained MT system did reasonably well (e.g., translated most of the patterns acceptably) and a series where it failed more often. This will essentially indicate the level of proficiency, and can be related immediately to the highest point on the text handling tasks hierarchy for which the system output is useful.

The diagnostic test set must be validated for the accuracy of the predictions it makes about the tasks an MT system can support. One method of validation might be to use an automated text-handling system (e.g., an extraction system) to verify whether an MT system tested suitable for extraction was in fact suitable. However, the state of the art in automated text-handling systems is such that any system, no matter how robust by current standards, might add indeterminacy to the outcome of a validation. Consequently, the validation, like the development of the metric, will be heavily oriented toward the users. In effect, the process consists of putting new MT output through the same user-exercise and analysis steps that the pre-scored corpus was, and then comparing the judgments about this new output against the score produced by running MT Proficiency Scale test on the same output. An additional advantage of this approach is that gathering the validation input from users can actually be done at the same time as the original interview sequence.

Conclusion

The MT Proficiency Scale is an attempt to capture both the practical realities of translation in the context of users, and the subjective measures of translation fidelity and intelligibility (through reuse of the DARPA HLT scored corpus). It will provide a quick, inexpensive and portable diagnostic set to predict the suitability of an MT system's output for real uses in modem text handling and analysis.

There are several critical issues to be confronted in the course of this development, in particular the merging of *user* judgments over many user groups and missions into a single hierarchical scale of text handling tasks, and the accurate representation of translation problems in the final diagnostic test set. These and other issues will require careful analysis of user judgments and translation phenomena.

Subsequent study will determine to what extent the processes, and perhaps even data, of the current project can be extended to other language pairs. Ultimately, the MT Proficiency Scale will serve as a reusable, potentially self-testing means of determining the performance of an MT system for the actual purposes to which MT will be applied now and in the future.

References

Church, Kenneth, and Eduard Hovy. (1991). Good Applications for Crummy Machine Translation. In Jeannette G. Neal and Sharon M. Walter (eds.) *Proceedings of the 1991 Natural Language Processing Systems Evaluation Workshop.* Rome Laboratory Final Technical Report RL-TR-91-362.

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

Dostert, B. (1973). User’s Evaluation of Machine Translation, Georgetown MT System, 1963-1973. *Rome Air Development Center Report AD-768 451.* Texas A&M University.

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.

Van Slype, G. (1979). *Critical Methods for Evaluating the Quality of Machine Translation.* Prepared for the European Commission Directorate General Scientific and Technical Information and Information Management. Report BR 19142. Bureau Marcel van Dijk.

White, J., and T. OConnell. (1994). The ARPA MT evaluation methodologies: evolution, lessons, and future approaches. *Proceedings of the 1994 Conference, Association for Machine Translation in the Americas.*

White, John S., and Theresa A. O'Connell. (1996). Adaptation of the DARPA machine translation evaluation paradigm to end-to-end systems. *Proceedings of AMTA-96.*