Evaluation of Machine Translation Output for an Unknown Source Language: Report of an ISLE-Based Investigation

Keith J. Miller, The MITRE Corporation, Washington D.C., USA keith@mitre.org
Donna M. Gates, Carnegie Mellon University, Pittsburgh, USA dmg@cs.cmu.edu
Nancy Underwood, CST, Denmark nancy@cst.ku.dk
Josemina Magdalen, The Hebrew University of Jerusalem, Josemina@banter.com

Abstract

It is often assumed that knowledge of both the source and target languages is necessary in order to evaluate the output of a machine translation (MT) system. This paper reports on an experimental evaluation of Chinese-English MT and Spanish-English MT from output specifically designed for evaluators who do not read or speak Chinese or Spanish. An outline of the characteristics measured and evaluation follows.

Keywords

Machine Translation, Gisting, Comprehensibility, Fidelity, Evaluation

Introduction

The evaluation of the performance of machine translation systems is a topic that has instigated much discussion. This is partly because in addition to all of the usual metrics used in software evaluation, the linguistic quality of the output must be considered. Even when the evaluation of machine translation is reduced to an evaluation of the output quality, there are no universally accepted measurements. This is primarily due to the fact that in the absence of an absolute gold standard, it is difficult to determine what constitutes a good translation. The problem is compounded when it is necessary to evaluate the output of a system without any knowledge of the language in which the source documents are written. This paper reports on work aimed at addressing this situation. The work was primarily performed at a workshop on MT evaluation held at the University of Geneva in the spring of 2001. The workshop was sponsored as part of the MT evaluation effort of the International Standards for Language Engineering (ISLE) research program. Information on ISLE, and in particular information concerning the workshop and the ISLE taxonomy for machine translation evaluation drawn upon in this report can be found at
http://issco-www.unige.ch/projects/isle/taxonomy2/

Approach to Evaluation

The approach to evaluation is based on earlier EAGLES work in MT evaluation beginning with the EAGLES 7-step recipe[1] According to the 7-Step Recipe, in order to carry out a meaningful evaluation, the evaluators must know the end user's purpose in seeking translation of documents originally in another language. That is, what task is the end user trying to accomplish? Following this suggestion, we spent not an inconsiderable amount of time defining our hypothetical user and the task for which that user had chosen to consider the use of machine translation. This time was not wasted – it helped us to concretize many of the other decisions we had to make during the evaluation design. The next section describes our hypothetical user and task, and the following sections discuss the ISLE characteristics that we decided would be relevant for that user's task and the metrics we developed to measure those characteristics. Finally, we discuss our findings with respect to these metrics and their potential for development in future evaluations.

User and Task Description

The user of the MT system[2] we are evaluating is an English-speaking librarian who has to classify and gist documents written in other languages. He has to be able to retrieve them based on the gist in English produced by an MT system. The librarian most likely does not speak or understand the source language(s) so he must rely on the output of the MT system. We evaluated the chosen MT system with these needs in mind. The analysis of the librarian's task had a serious impact on the characteristics we have chosen to measure as part of the evaluation described in this paper. For example, we decided that faithfulness of translation (fidelity) is of major importance since the librarian must rely on the MT system output for the tasks of classification and gisting without any possibility of crosschecking with the source text. For very similar reasons the correct coverage of terminology is also important. We also believe that comprehensibility is of great importance both for the efficiency and for the accuracy of the librarian's work. Other ISLE measures for MT systems, like the "syntax of the target language

[1] available at http://issco-www.unige.ch/projects/eagles/ewg99/7steps.html,

[2] We do not wish to disclose the name of the system used in the testing and we shall subsequently refer to it simply as "the MT system."

sentence" or the "coverage of cross-language phenomena", while very important in a general evaluation, become, in the context of our librarian's task, less relevant. Thus, they were not chosen as measures in this exercise.

Characteristics to be measured

Based on the user and task described in the preceding section, the following six ISLE characteristics were identified as the most relevant to evaluate:

Comprehensibility, Readability, Fidelity, Coverage, Terminology and Utility of output

Due to time and resource constraints, it was not possible to measure the utility of the output, as this would have entailed a full-blown task-based evaluation of the type described in (White, et al. 2000). For this reason, only the first five characteristics were measured in this experiment. We further assumed that none of the evaluators had any knowledge of the source language of the documents, which proved to be true in the case of Chinese. We also compared the use of the metrics on the unknown source language (i.e., Chinese) documents to documents in which some of the evaluators had knowledge of the source language (i.e., Spanish). This had a distinct impact on the different fidelity and readability scores in cases where untranslated words appeared from the known source language (Spanish). Further comments on the distinction between the two cases are included in the discussion section of this report. Finally, most of the tests described in this report can be carried out in the absence of any knowledge of the source language text or its correct translation. However, it was assumed that for the purposes of testing the fidelity of the machine translation output as well as for testing the correct translation of domain-specific terminology, at least one (and in our case, only one) human translation of the evaluated texts was available to the evaluators. It is important to note that the human translation could only be referenced at those points in the evaluation that it was specifically called for, and that those evaluation steps came after those measures that could potentially be affected by having seen a 'correct' translation (e.g., the comprehensibility and readability measures).

Metrics for the characteristics

Comprehensibility

The comprehensibility measure seeks to address the question: "Is the text understandable?" Two metrics were decided upon for purposes of measuring comprehensibility of the output text. The first of these was a forced-choice subjective judgment of intelligibility on a sentence-by-sentence basis. Raters were asked to assign each sentence a 1 or a 0, depending on whether a sentence was comprehensible[3] or not. The final comprehensibility score for a text is then the total number of comprehensible sentences divided by the total number of sentences. The second method for measuring comprehensibility was an adaptation of the cloze test (Taylor, 1953). This test is based on whether, when presented with a new output text with every Nth word deleted, the test-taker is able to successfully recover the missing words. In our system, scores are based on the percentage of deleted words exactly recovered by the test takers.

Readability

In contrast to comprehensibility, readability addresses the ease with which the output text can be read. It was decided that comprehensibility and readability are different from one another because a text may be comprehensible after several close readings even though it is difficult to read. Furthermore, a more easily readable text may save the user time because he doesn't have to reread the text before understanding it. However, the metric that we decided upon for testing readability, that is measurement of the time required to read an output text aloud, was beyond the resources available for this project. Thus, although a readability metric was designed, it was not carried out, and no separate readability score will be reported.

Fidelity

Fidelity was designated the most important characteristic for the hypothetical librarian's task. Note that fidelity is a measure of the information successfully conveyed from the source language text to the target language output. Since by design none of the evaluators had any knowledge of the source language (Chinese), it was necessary to perform this test by judging the fidelity of the translation with respect to an available human translation rather than the original text. Fidelity scores were computed in the following manner: Each sentence was assigned a value from a subjective 4-point scale. These individual sentence values were then averaged over the whole text. This is essentially the test proposed in Van Slype (1979). The scoring is performed with values ranging from 0 to 3, based on the amount of information in the human translated sentence, which is also in the test sentence, according to the following guidelines:

0 = no information
1 = less than 50% of the information
2 = more than 50% of the information
3 = all the information

[3] Evaluators were allowed to re-read the text several times to determine whether or not it was understandable. Immediacy of comprehensability would fall under the scope of the readability measure.

Coverage

Given that it is not possible to deal with cross-language phenomena since by definition of our task these are unknown, our metric for coverage is quite simple. It is calculated as the percentage of translated words. This test was carried out automatically by counting the untranslated words in the output text, and dividing by the total number of words in the output text.

Terminology

For the terminology test, we identified potential domain-identifying terms in the human translated text and determined whether they occurred in the machine-translated texts. As with coverage, the metric is the percentage of correctly translated domain terms. In the event that the evaluators had knowledge of the source language, an alternative methodology would be to identify domain terms in the source language document, define acceptable translations for each domain term, and then determine the percentage of these domain terms that were correctly represented in the machine translation output.

Testing the Data

We performed tests with a commercially available MT system. Translations were obtained for 8 segments of a Chinese chemical weapons treaty, each segment consisting of approximately 150 lines. We additionally performed some of the tests on 4 segments of 150 lines from the Spanish version of the treaty.

Comprehensibility

The 0/1 Test

We tested the comprehensibility of the translation of 8 texts produced by the Chinese-To-English system and 4 texts produced by the Spanish-To-English system. In general, two persons did the test and the results below are an average of the individual results:

Chinese-To-English:
Text1: 37.60%
Text2: 45.10%
Text3: 58.33%
Text4: 57.62%
Text5: 52.00%
Text6: 60.00%
Text7: 64.00%
Text8: 60.00%

Spanish-To-English:
Text1: 76.60%
Text2: 88.00%
Text3: 73.60%
Text4: 47.60%

This measure may be somewhat biased toward short translations because the shorter the text is, the easier it is to "read into" the meaning, or to infer it from the context. On the other hand, if there are serious MT problems such as many un-translated words remaining in the text, then the meaning of very short chunks/sentences cannot be deduced. So texts with short sentences can also be problematic to evaluate. We tested a treaty segment, which turned out to be the "Table of Contents". Some of it was easy to understand because of the simple nature of text, but the section titles with very few terms translated were nearly impossible to understand. However, even here not everyone agreed. One evaluator found the opposite to be true.

It was felt that the 0/1 Comprehensibility measure does not have enough granularity. Being forced to assign either 0 or 1 resulted sometimes in a score of 1 even when some of meaning was lost, or a score of 0 in spite of the fact that some of the meaning got through. The 0/1 comprehensibility test can also be biased by the lack of knowledge of the specific domain the text is related to. Sometimes, if one is familiar with the domain one can infer more from a worse translation. Our results, based on this test, have shown that for Chinese-To-English comprehensibility is generally low. We could identify a few general phenomena, like a non-standard structure of the sentence in English related to terms (probably noun phrases) translated into finite verbs. Later on when we performed the next text we could see that the terms were translated into nouns in the human translation (HT). The lack of plurals and of properly translated numerals also reduced comprehensibility. After performing the test vis-à-vis the HT, we could see that these phenomena also lead to a decrease in the translation fidelity. All these sorts of phenomena made comprehensibility very difficult. As mentioned before, we evaluated both a Chinese-To-English and a Spanish-To-English system. It was noticed that even for those of us did not know Spanish, the comprehensibility of the MT output of the Spanish-To-English was considerably greater, due partially to the fact that the there are many cognates between English and Spanish. In addition, we know that there is greater similarity in the syntax of the sentence between English and Spanish than between Chinese and English. This fact probably determines a higher quality, where comprehensibility is concerned, for the Spanish-To-English text rather than for the Chinese-To-English one.

Thus there were a number of open issues arising from our experience with this test.

Issues for 0/1 Comprehensibility test

a. One main issue is how to control in a subjective evaluation for the evaluator's knowledge of the source language or about the source language? This is not supposed to be an issue, but there are various levels of knowing a language (or about a language) in reality.

b. Another interesting test we would have liked to do was to compare more thoroughly the comprehensibility of translations from languages we know versus languages we

do not know. In such a case, one could decide to choose languages with the same degree of similarity to the target language, for example French and Spanish to English, in order to avoid the bias created by the very different cross-languages differences like Chinese versus Spanish to English.

c. Is it possible to control for the evaluator's specialized linguistic knowledge of the domain to which the texts are related? Here it may be relevant to devise a cloze test on a human translation for comparison.

d. What should the minimal unit of translation be that we are scoring? Sentences in this domain are rather long and complex. It is difficult to decide at what point an evaluator should decide that a sentence is badly translated because the first half is bad or that he should decide that it is good because he can at least understand half of it. This is relevant for fidelity as well. Some researchers have proposed a smaller unit based on speech-acts or phrases containing a semantic nucleus (Lavie et al, 1996) or even chunks of phrases.

The Cloze Test

Initially, we had planned to perform a cloze test on both the MT and the HT data. We thought that this would provide an objective comparison of the comprehensibility of the MT and neutralize the influence of the users (lack of) expert knowledge of the sub-domain. Unfortunately, we discovered that the lack of expert knowledge in the domain significantly hindered our ability to perform a Cloze test on either set of data. This occurred both with the Chinese-to-English translations as well as with the Spanish-To-English translations.

In general, we experienced a high level of frustration in performing the cloze test and most of us gave up rather quickly The cloze test, in its current incarnation, was found to be useless even with highly motivated evaluators (e.g., the authors of this paper). It may be possible to reformulate the test for trials in future evaluations. As a case in point, Miller (2000) successfully uses the cloze test as a fine-grained MT evaluation measure. However, in that work, items deleted in the cloze texts were carefully controlled, and did not principally include domain-specific content words. Furthermore, the texts used in Miller's study were not from a very narrow, highly technical domain.

Fidelity

We have tested the fidelity of the translation of 3 (of the 8 we used for comprehensibility) texts produced by the Chinese-To-English system and 4 texts produced by the Spanish-To-English one. The numbers of the texts correspond to those in the comprehensibility test, first we present the number of assignments of each score (0,1,2 or 3), the number with a score more than 0 and then the final weighted rating for each text:

Chinese-To-English:

Text2: 0 – 36/61; 1 – 11/61; 2 – 12/61; 3 – 2/61;
25/61 non-zero; final weighted grade 41/61;
Text5: 0 – 38/54; 1 – 10/54; 2 – 3/54; 3 – 3/54;
16/54 non-zero; final weighted grade 25/54;
Text7: 0 – 1/27; 1 – 17/27; 2 – 9/27; 3 – 0/27;
26/27 non-zero; final weighted grade 35/27;

Spanish-To-English:

Text1: 0 – 2/26; 1 – 14/26; 2 – 10/26; 3 – 0/26;
24/26 non-zero; final weighted grade 34/26;
Text2: 0 – 7/42; 1 – 14/42; 2 – 20/42; 3 – 1/42;
35/42 non-zero; final weighted grade 57/42;
Text3: 0 – 26/43; 1 – 8/43; 2 – 6/43; 3 – 3/43;
17/43 non-zero; final weighted grade 29/43;
Text4: 0 – 23/44; 1 – 10/44; 2 – 10/44; 3 – 1/44;
21/44 non-zero; final weighted grade 33/44;

Our results showed a low rate of fidelity with respect to the source text for the machine translated text as compared to the human translations. Even though our Spanish-To-English MT translation had a lower Fidelity than the Chinese-To-English translation in terms of the high number of untranslated terms, we believe that the librarian's task of gisting would be easier given the output of the Spanish-To-English system. Due to the fact that Spanish and English had a significantly greater number of cognates in our tests than did Chinese and English, the higher comprehensibility due to understandable cognates would be very helpful to a librarian with the job of determining the gist of the documents. We also found indications that a fidelity score would be higher if untranslated cognate words, which could still be understood, were considered acceptable translations to some extent. Thus, we feel that Fidelity should not be considered in isolation from the other measures. The final evaluation is a summary of all of these measures. A lower fidelity does not necessarily mean that the librarian's task is less doable. We feel that it is necessary to take into account the way in which each metric is interpreted. A very low fidelity does indicate something very important about the usability of the system. If it is not a faithful translation of the source text, it is not useful. If it has some fidelity problems related to untranslated words then the problem may be compounded by problems reflected by the comprehensibility score, as evidenced by the differences between our Spanish and Chinese Evaluation scores.

Comprehensibility versus Fidelity

A comparison of comprehensibility and fidelity is warranted here. For comprehensibility, it was often the case that we assigned a "1" due to the subjective feeling that we understood what was being conveyed. Only after looking at the gold standard (HT) did we find that we were mistaken (e.g. one of us took the term "*meter*" as used in the Spanish translation to refer to *measures/measurements* (i.e. not an absolutely terrible

translation) but it turned out that the correct word was "*perimeter*" and thus it was completely wrong)[4].

Another interesting finding of this test was the treatment of negation. Sometimes the negation was translated incorrectly (or simply not translated), resulting in an affirmative, rather than a negative sentence. Under the comprehensibility test this type of sentence was scored as understandable in most cases. However, when performing the fidelity test with a comparison between MT and HT, the incorrectness of the translation was revealed and then assigned a fidelity score of "0"[5]. An additional problem that we found was that by not preserving the formatting of the document, it is possible to introduce more confusion, which can, in turn, influence both the comprehensibility and fidelity scores. When the paragraph, section, and subsection numbers appeared in the middle of a target sentence, or if a title was mixed up with the subsequent sentence, it may have resulted in lower scores for the sentence simply because the evaluator was annoyed or confused.

Again there are a number of open issues arising from the fidelity tests

Issues for the Fidelity test

a. Cognates in related languages? Theoretically, as the design of the fidelity measure would indicate, any source language word that appears in the target language should result in a lower score for that sentence. In practice, we found that this was sometimes difficult to enforce in the case of cognates, which often appear to be so close to a target language word that the correct meaning could still be conveyed. This was found to be the case with Spanish-English but not with Chinese-English. In such cases, the results may have been biased and/or inconsistent.

b. What is the minimal unit of translation to be evaluated? For our purposes, we only considered the entire sentence or a sentence fragment when no complete sentence was present (see discussion of Comprehensibility).

c. The metric we used for scoring fidelity is based on a 4-point scale. We have found this metric acceptable for this evaluation. However, we need to pay close attention to the evaluation data. Errors that dramatically change the sense of the sentence like shifts in polarity (i.e., negative becoming positive) should be seriously punished. Some of us thought that a 3–point scale would be a better idea. We also believe that these grading scales should be tested for intercoder agreement in order to have a concordance in the results.

[4] This may also be due to the fact that the accent marks were incorrectly formatted so that "peri'metro" was read by the MT system as "per metro".

[5] It is just this sort of error that illustrates why a comprehensibility measure alone would not be an adecquate measure of the overall quality of an MT system.

Coverage[6]

In general, we noticed that overall coverage with the MT system was high while the quality of the translation seemed intuitively to be quite low. What seems more relevant is the coverage of the terminology. Since a word may be translated with a very different meaning than was intended, we believe it would be necessary to take into account these incorrectly translated words. For future evaluations of this type, we believe that it would be worth investigating the inverse of this coverage score (i.e., 1-coverage).

The table below shows the coverage scores for each of the segments that were randomly selected from the Chinese corpus.

Coverage for 4 Chinese segments

Segment	Coverage
14	0.957
26	0.981
29	0.980
36	0.985

Terminology

In the translation from Chinese, the terminology coverage is medium (better than the Spanish one). However, fidelity is low because many terms, though translated, were incorrect senses of the source terms given the context in which they appeared.

The vast majority of the special domain terms were not translated (e.g., Phosphorus *oxychloride, Phosphorus pentachloride , etc*[7]). Another group off terms were not correctly translated (e.g., Annex *on implementation and verification (HT) = About carries out appendix which and investigates (MT), Verification annex (HT) = investigates appendix (MT), State Party (HT) = signatory state (MT), Host State (HT) = country territory (MT)Conduct of inspections (HT) = inspection carries on (MT), Pre- (HT) = Before (MT)Standing arrangements (HT) = Rule arrangement (MT)*).

We guessed (while scoring comprehensibility) that a few of the non-translated words from Spanish might be

[6] The quality of the Spanish MT may have been affected by issues of data quality regarding the input texts. We noticed a lot of non-translated words, especially for words with diacritics. This included a space appearing where the accented "i" should appear in Spanish words (e.g., "arti'culo" => " art culo"), much to our amusement. It is possible that the encoding of the Spanish texts was different from the encoding expected by the MT system and that this may be the source of the problem. Problems with the encoding of diacritics and special symbols are a natural source of problems in MT and special attention must be given to them when using any MT system.

[7] For readibility's sake, the English translation is provided here in place of the actual Chinese source.

specialized names for objects, so we did not count against these. To our surprise, the fidelity tests then revealed that these terms were in reality not translated and SHOULD have been. We believe this is more likely to happen between languages that share a common alphabet. For Spanish-English[8] texts, very few expressions were correctly translated vis-à-vis the Human Translation (HT). The vast majority of the terms were not translated leaving more than 50% of the terms in the target text to still be in Spanish). A few examples: *parraf, combinación, incluir, transportar, relación, etc.* Another group of Terms were not correctly translated, for example, *perimeter (HT) = meter (MT)*[9].

Known versus Unknown Language

We found the translation from Spanish to be easier to understand due partly to the fact that there were so many cognates between English and Spanish. We tried to filter this out as much as possible, but it was difficult not to miss some of the un-interpretable lines because we could still understand the meaning of the sentences. We think we can all agree that the Spanish translations are "better" than the Chinese ones, but there are probably a number of reasons, which affect this judgment, like the greater similarity between English and Spanish than between Chinese and English, at the level of alphabet and terms as well as syntax and historical relatedness. The terms (especially Latin derivatives) in Spanish are really very similar to those in English for these texts, and this makes the task of the librarian who needs to do the gist easier, even though many of the terms were not translated. Therefore, we surmise that the gist task can be performed quite satisfactorily even with a poor translation. The Chinese texts are far less comprehensible/readable than the Spanish ones, even though the Spanish translation looks worse, at least in terms of terminology coverage. The quality of the translation in terms of the metrics we have set is quite low, very low terminology coverage[10] and medium fidelity for the translated parts. The comprehensibility is quite good in Spanish due to the similarity in terms, and not to the quality of the translation. With respect to measuring the fidelity for Spanish, due to the fact that so many of the untranslated Spanish words resemble English so closely, as mentioned earlier, this affected our scoring in that we would count the information content as the same in both the English and Spanish. It is unclear whether this is reasonable since only some of the evaluators know Spanish, and so other English speakers may also be expected to be able to understand the Spanish cognate words.

[8] Translations from Spanish may not be valid since the accent marks were formatted with atypical ASCII codes, which may not have been proper system input.

[9] This was clearly due to the accented character being deleted.

[10] Unless the MT system used incorrect character encoding and/or dictionary.

General comments

We found it much easier to grade the files in the ASCII format, as we did in our geographically dispersed, post-workshop team, rather than on paper as we did at the workshop in April. We regraded the segments from then as well, and found we gave better scores this time around, just because of the format. We believe the format was a factor in the original scores we gave; this serves as a reminder that human factors are important in this type of evaluation experiment. We need to remember this when we design evaluation procedures. At least one scorer mentioned the influence of a training effect. While going through the texts, after already knowing what they were about, the scorers began to "learn" some Spanish terms and thus comprehensibility became artificially improved on later texts. This was surprising since we very carefully designed into the evaluation procedure that the scoring be performed in a very specific order to avoid such bias. We determined that the fidelity, coverage and terminology testing should be performed only after the evaluators performed the comprehensibility and readability tests since knowledge of correct text could influence these measures. We found that during the fidelity test, we also experienced some interference from learning some Spanish during the comprehensibility test.

Conclusions

The outcome of this experiment is that, based on features suggested in the ISLE framework for MT evaluation, our team was able to evaluate the output of a Chinese-English and Spanish-English MT system without knowledge of the source languages. Some of the tests proved more discriminatory than others. In particular, the test of raw coverage, or the percentage of translated words, seemed to provide little value, in that all of the scores were clustered in the 90% range. A much better high-level discriminator was the terminology test, which took into account the correct translation of domain-identifying terms. The same conclusion can be drawn from comprehensibility vs. fidelity. Even at a lower fidelity rate, we noticed that the texts were more comprehensible, and, hence would make it easier for our imagined librarian to perform the task devised for this experiment (i.e., gisting a document for later retrieval). Nonetheless, basing a gist on a translation with very low Fidelity (e.g. in which negative sentences are translated as affirmative sentences) would lead to undesirable task performance. We also conclude that similarity between languages, even when texts are poorly translated, might allow such a task to be completed easily despite the poor translation quality.

Bibliography

Lavie, et al., "Dialogue Processing in a Conversational Speech Translation System", *Proceedings of the ICSLP 96*, Philadelphia, USA, October 1996

Miller, K. 2000. The Machine Translation of Prepositional Phrases. Unpublished PhD Dissertation. Georgetown University. Washington, DC.

Taylor, W.L. 1953. Cloze Procedure: A New Tool for Measuring Readability. *Journalism Quarterly* 30:415-33.

Van Slype, G. 1979. Critical Study of 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

White, J., Doyon, J. & Talbott, S. 2000. Task Tolerance of MT Output in Integrated Text Processes. In Van Ess-Dykema, C., Voss, C., & Reeder, F., eds. Proceedings of the Workshop of Embedded Machine Translation Systems, ANLP/NAACL-2000. Association for Computational Linguistics, Seattle, Washington.