SPANISH-TO-ENGLISH TRANSLATIONS: ACCURACY VERSUS UNDERSTANDING

Milam Aiken, School of Business Administration, The University of Mississippi, University, MS 38677, 662-915-5777, aiken@bus.olemiss.edu
Zachary Wong, School of Business & Economics, Sonoma State University, Rohnert Park, CA 94928, 707-664-2377, zachary.wong@sonoma.edu

ABSTRACT

Machine translation (MT) systems continue to improve in accuracy, and a few studies have attempted to measure how well these programs translate one natural language (e.g., Spanish) to another (e.g., English) with specific formulas or surveys of how well a native speaker understands the machine-generated text. This study is perhaps the first to compare an accuracy formula with subject understanding. A test using *SYSTRAN* to automatically translate 12 samples of Spanish text to English showed that there were few significant correlations among understandability, reading ease, and accuracy. However, as expected, the understandability of a translation decreased with more difficult text (i.e., longer sentences and more syllables per word).

INTRODUCTION

As the need for natural language translation (e.g., Spanish to English) continues to grow, research on the use of computers to automate the process has resulted in several academic and commercial machine translation systems. These programs are typically very fast (generating a translation in a small fraction of the time a human fluent in both languages could provide) but are still less accurate than a good interpreter. Because human translation services are in short supply and expensive, however, many people are utilizing free, Web-based translation services such as *http://babelfish.altavista.com/* powered by *SYSTRAN* (*http://www.systransoft.com/index.html*) language translation software to provide a quick, if inexact translation of text found on the Web or in electronic mail. Once the user has a general idea about the content of the text, he or she can decide if a more accurate translation provided by a profession interpreter is necessary.

Several commercial MT systems have been evaluated for accuracy using a variety of measures such as word counts, self-assessed understanding, etc., but to our knowledge, no study has compared source text difficulty, a numeric accuracy formula, and an end-user quiz of understanding to determine if, for example, the accuracy measure is correlated with understanding, or if text difficulty makes translation easier or harder.

First, we describe a few earlier studies of language translation, focusing solely on Spanish and English, for simplicity. Next, we describe ways to measure translation accuracy, followed by a study of subject understanding.

PRIOR STUDIES OF ENGLISH AND SPANISH TRANSLATION ACCURACY

Although studies of Spanish and English translation are not directly comparable due to differences in the number of subjects evaluating the text, the language skill of the subjects, the difficulty and nature of the source text, the MT software used, etc., the research below gives an indication of the accuracy can be achieved by such systems.

Study 1 (Spanish and English Translation) (Aiken, et al., 1994a)

Spanish-speaking students reported on average that 6.6 samples of translated text (using *Spanish Assistant 5* from MicroTac Software) were grammatically incorrect (26 % of the comments translated from English) and one sample was misunderstood (4% of the text passages translated from English). English-speaking students reported on average that 10.7 samples of text were grammatically incorrect (55% of the text passages translated from Spanish) and 3.5 comments were misunderstood (15% of the comments translated from Spanish). However, many of the errors in translation occurred because the text samples had misspelled words. That is, the text was "free form," not taken from textbooks, but rather, generated as one would write a letter to another person.

In a separate analysis, objective, independent reviewers were asked to evaluate the grammatical accuracy and understandability of 100 text samples translated from Spanish to English and 100 passages translated from English to Spanish. The English reviewers rated the Spanish-to-English grammatical accuracy at 46% and the understandability at 95%. The Spanish reviewers rated the English-to-Spanish grammatical accuracy at 75% and the understandability at 98%. The increase in accuracy was due in part to all of the source text being spelled correctly.

Study 2 (Spanish and English Translation) (Aiken, et al., 1994b)

In one study using *Spanish Assistant 5*, 23 passages of text were written in English "free form" and 13 written in Spanish. Translations of Spanish and English were then evaluated by independent observers. A total of 7 of the 13 Spanish comments (54%) had some kind of error, while only 7 of the 23 English comments (30%) had some kind of error.

Study 3 (Spanish and English Translation) (Aiken, et al., 1998)

An analysis of translations of free-form text using *Spanish Assistant 5* showed that 24% of the Spanish comments had grammatical errors, and 29% of the English comments had errors. The Spanish-speaking evaluators understood 81% of the Spanish comments, and the English-speaking evaluators understood 91% of their comments.

Study 4 (Spanish-to-Translation) (Bezhanova, et al., 2005)

In this study, 17 English sentences were translated into Spanish using *LogoMedia*, *SYSTRAN*, and *PROMT*. Results are shown in Table 1. The authors concluded that all three of the MT systems produced usable translations, and that none has an obvious advantage. However, the *SYSTRAN* translations were generally the worst. In addition, the authors found that vided short sentences were translated very well, but many longer sentence translations were very difficult to understand (an indication that source text difficulty could play a vital role).

Study 5 (English-to-Spanish Translation) (Aiken, et al., 2004)

Three expert Spanish speakers evaluated two sets English text samples translated by *SYSTRAN* into Spanish and were asked to evaluate the overall understandability and to count the numbers of major and minor errors. Results are shown in Table 2. At best, only 85% to 89% of the translated text could be understood. The results also show the wide variability of self-reported understanding that can occur.

MEASURING TRANSLATION ACCURACY

There are no universally accepted and reliable measures of machine translation accuracy (Balkan, et al. 1994; Falkedal, 1994; White & O'Connell, 1994), and as shown in study #5 above, a sample of text given to three different professional translators can yield three different results. Some studies focus on the percentage of sentences with minor or major errors, some focus on the percentage of text that is understood by subjective evaluators, and others compute a measure automatically using N-grams (NIST, 2002; Papineni, et al., 2002).

One objective measure of natural language translation accuracy is to count the number of errors per hundred words. However, word insertions and deletions can have an effect on the intelligibility of a translation, and these errors are not reflected in simple word error rates. In an attempt to be more systematic, in this study, we utilize the formula below:

$$\text{Translation Accuracy} = 1 -(d+s+i)/\max(r,c))$$

Where, d = the number of words missing from the correct translation. E.g. in the sentence "Apple red," the word "is" is missing

s = the number of wrong word choices. E.g., in the sentence "Apple was red," the word "was" is used while the correct translation might have used the word "is."

i = the number of words incorrectly added. E.g., in the sentence "Apple is the red," the word "the" is added incorrectly

r = the number of words in the translation

c = the number of words in the correct translation.

However, this formula does not take into consideration word order errors. For example, in the sentence “Apple red is,” there are no missing, added, or incorrect words. Although the words are out of order, the sentence can still probably be understood.

In addition to measuring translations accurately, it is not clear how well an objective measure such as the formula above correlates with subject human understanding of a translation. One study (Culy & Richemann, 2003) using the automatic BLEU (bilingual evaluation understudy) technique showed a strong correlation between the measure and human judgments of translation quality. For example, a test of BLEU with Spanish and English had a correlation coefficient of 0.975 for adequacy, 0.972 for fluency, and 0.943 for informativeness. However, another study of Automatic Evaluation (Turian, et al., 2003) showed that the correlation between human judges and automatic measures of MT quality was low. The most important finding in this research was that, even though human evaluation of MT is itself inconsistent and not very reliable, automatic MT evaluation measures are even less reliable and are far from being able to replace human judgment.

Finally, the role of source text difficulty in language translation has not been explored to any great extent (Hale & Campbell, 2002). It seems like common sense that a translation of a short sentence of mono-syllabic words (such as those found in a 1st-grade text book) would be fairly accurate, but one study (Aiken & Wong, 2001) of a computer speech recognition system (one could argue, a similar technology to that of natural language translation) showed that added textual complexity (longer sentences) added context, making speech-to-text generation more accurate.

TRANSLATION STUDY

Translation accuracy is highly dependent on the quality of the source. Therefore, we obtained random Spanish sentences from two introductory Spanish textbooks, and two Web sites (believed to have no spelling or grammatical errors). Along with the source text (shown in Appendix 1), we added measurements of Flesch reading ease (0=very difficult, 100=very easy) and Flesch-Kincaid grade level (1=very easy, 12=very difficult). Source text difficulty ranged from a low of reading ease = 75.8 and a grade level = 3.6 for text sample number 4, to a high of reading ease = 0 and a grade level = 12 for text sample numbers 10 and 11."

A Spanish-to-English translation study was chosen because of the shortage of subjects fluent in Spanish. A sample of 123 undergraduate students (49 females) from two universities were asked to evaluate the understandability of the *SYSTRAN*-generated translations shown in Appendix 1, using a scale of 1 = ”I have no idea what this means” to 7 = “I am sure what this means.” After a short break, they then were asked to choose the correct meaning of the translations in the quiz shown in Appendix 2 (an attempt to obtain a more objective measure of their understanding). In addition, the students recorded their Spanish- and English-speaking abilities using a scale of 1 = “none” to 7 = “excellent.”

RESULTS

Subjects reported not being able to speak Spanish well (Mean = 2.7, Std. Dev=1.5), but they were able to speak English fairly fluently (Mean = 6.4, Std. Dev = 0.9).

Translation number 4 was understood the best, while number 3 was understood the least (Table 3). A difference-of-means T test showed that only translations 3 and 9 were significantly below the median on the understandability scale (not understood), although translations 8 and 11 were not significantly different from the median. In general, subjects were able to choose the correct translation in the second quiz, with the exception of translations 8 and 11. Only translations 7 and 10 had no grammatical or word choice errors, although their phrasing is still awkward. Nevertheless, the students reported not being able to understand only two of the 12 translations.

Translation accuracies were calculated using the formula above by an independent evaluator, and the results (along with reading ease scores) are shown in Table 4.

There were weak and insignificant correlations between the translation accuracy and reading ease scores (R = 0.34, p=0.27) and between translation accuracy and the percentage of subjects getting the correct answer on the quiz (R = 0.19, p=0.56). However, there was a moderate, negative, significant correlation between the percentage of subjects getting the correct answer on the quiz and reading ease (R = -0.62, p=0.03).

Because items measured on a 7-point Likert scale yield nominal data, this violates the requirements of Pearson's Correlation test (Brace, et al., 2000, p. 111). Instead, we used Spearman's Rank Order Correlation non-parametric test (also called Spearman's Rho) for other comparisons. Results showed that there was little or no significant correlation between language ability and subjects' quiz answer understandability (see Table 5), between language ability and self-reported understandability (see Table

6), or between self-reported understandability and subjects' correct quiz answers (see Table 7).

In general, there was hardly any difference between males (70%) and females (71%) in terms of the percentage of correct answers on the quiz. However, there was a significant difference between males and females for translation 11 where females reported not being able to understand it (mean = 3.6) while males reported they did (mean = 4.5). In addition, there were significant differences in understandability between those who reported knowing Spanish and those who didn't for translations 3, 9, 10, and 11.

There could be several possible causes for the insignificant and/or low correlations. First, the accuracy formula might not be adequate to measure the understandability of a translation. Alternatively, the evaluator could have made errors in applying the formula. For example, he counted word order errors as insertion or deletion errors. Perhaps, word order errors should not have been counted. Second, subjects might have said they understood a comment, when they really did not. Third, the quiz answer choices were somewhat arbitrary, and some choices are very similar. Thus, an answer could have been marked as incorrect, even though its meaning was nearly identical to the correct translation. One of the few significant results (the negative correlation between the percentage of subjects getting the correct answer on the quiz and reading ease) shows that added complexity decreased the understandability of a sentence. That is, the longer the sentence and the more syllables per word, the harder it was for the average reader to understand.

CONCLUSION

There are several limitations to the study. First, college students were used, and thus, results might be generalizable to other population groups (e.g., business managers). Second, other translation software could give more accurate translations. Finally, as note above, changes in the quiz answer choices or a different accuracy formula could affect the results.

It is difficult to measure the understandability of a translation, and even fluent interpreters can translate text differently. A translated sentence can be understandable with one or more word errors, or can be completely misunderstood with only one word error. Future research will investigate improved accuracy formulas to better model what is actually understood by a typical reader or listener.

REFERENCES

Available Upon Request

TABLES

Available Upon Request

APPENDICES

Available Upon Request