Spanish-to-English Translation Using the Web

Milam W. Aiken

The University of Mississippi
School of Business Administration; University, MS 38677
662-915-5777
maiken@bus.olemiss.edu

Zachary Wong

Sonoma State University
School of Business and Economics;, Rohnert Park, CA 94928
707-664-2377
zachary.wong@sonoma.edu

ABSTRACT

Several free translation services are now being offered via the World Wide Web in a variety of languages. While none of the services is as accurate as a fluent human interpreter, few if any studies have sought to determine their relative performance. A sample of 20 Spanish phrases from an introductory textbook were translated into English using four online services (SYSTRAN, SDL, WorldLingo, and InterTran). Results show that SYSTRAN and WorldLingo were the most accurate, followed by SDL and InterTran.

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 professional human interpreter is necessary.

Few studies have been conducted on the relative accuracy of Web-based machine translation (MT) systems. This study seeks to compare four such systems (SYSTRAN, SDL, WorldLingo, and InterTran) using an objective, formulaic measure of accuracy, a self-assessed judgment of understanding, and an evaluation of perceived meaning versus actual meaning. While results varied, in general translations provided by SYSTRAN and WorldLingo were superior to those provided by SDL and InterTran. However, accuracy can be improved further if translations from SYSTRAN or WorldLingo are provided with those from SDL.

BACKGROUND

Many free, Web-based translation services are provided on the Web, e.g.:

1. Systran – Altavista Babelfish (SYSTRAN): http://www.systransoft.com/index.html
2. Enterprise Translation Server (SDL, Inc): http://www.sdl.com/
3. WorldLingo Translator (WorldLingo): http://www.worldlingo.com/index.html
4. Translation Experts – InterTran Translator (InterTran): http://www.tranexp.com/InterTran/FreeTranslation.html

Some sites use other companies' software. For example, Alta Vista (http://babelfish.altavista.com/) uses SYSTRAN. Other sites contain links to several services. For example, Mezzofanti (http://www.mezzofanti.org/translation/) contains links to the above four sites and judged their accuracy on a 1-to-5 scale as (SYSTRAN = 3.5, SDL = 3, WorldLingo = 3, and InterTran = 2).

In another study of MT systems (Bezhanova, et al., 2005), 17 English sentences were translated into Spanish using *LogoMedia, SYSTRAN,* and *PROMT* and 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 short sentences were translated very well, but many longer sentence translations were very difficult to understand.

In a third study (Aiken, et al., 2004) of SYSTRAN, two sets of English text were converted to Spanish. Three expert Spanish speakers then evaluated the overall understandability and counted the numbers of major errors (those affecting understandability) and minor errors (those not affecting understandability). As seen in Table 2, only 85% to 89% of the translated text could be understood.

Additional Measures of Translation Accuracy

There are no universally accepted and reliable measures of machine translation accuracy (Balkan, et al. 1994; Falkedal, 1994; White & O'Connell, 1994). 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. A more complex formula (described below) can account for word changes, insertions, and deletions.

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

Where:

d = the number of words missing from the correct translation. E.g. in the sentence "Name is John," the word "My" is missing.
s = the number of wrong word choices. E.g., in the sentence “My name was John," the word "was" is used while the correct translation should "is."
i = the number of words incorrectly added. E.g., in the sentence "My name is called John," the word "called" 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 “John my name is,” there are no missing, added, or incorrect words. Although the words are out of order, the sentence can still probably be understood.

Also, 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.

Another test can be added to help judge the performance of an MT system. Few studies have sought to compare the “correct” translation with a sentence written by the human evaluator. For example, the correct translation of the Spanish sentence “¿Se permite tomar fotografías?” is “Is it permitted to take photographs?” However, the MT system might generate “It is permitted to take photographs?” and the human evaluator might understand this as “Is it able to take photographs?” As seen in this example, even fairly good translations can produce misunderstandings. In addition, a sample of text given to three different professional translators can yield three different results.

TRANSLATION STUDY

A Spanish-to-English translation study was chosen because of the shortage of human evaluators fluent in Spanish. Because translation accuracy depends on the quality of the source text, a random sample of 20 sentences were obtained from an introductory Spanish textbook (Pei & Vaquero, 1957).

Each of the four MT systems listed above were used to provide translations for the 20 Spanish sentences listed in Appendix 1. An objective human judge fluent in English was asked to write in English what he believed the SYSTRAN translation meant. SYSTRAN only was selected because the Mezzofanti Web site rated it the most accurate. In the next phase, he was asked to evaluate the understandability of all four of the machine-generated translations, using a scale of 1 = ”no understanding” to 7 = “complete understanding.” A second human evaluator compared the four systems' and the evaluator's translation with the actual meaning provided by the textbook. Finally, objective accuracy measures were calculated using the formula described above.

RESULTS

In cases where the program was unable to translate a word from Spanish to English, the source word was simply repeated (e.g., "Déme a cigarette package."). SYSTRAN, WorldLingo, and InterTran all were not able to translate at least one word in Spanish, but SDL provided translations for all words.

Also, all translations were not necessarily spelled correctly. For example, SYSTRAN and WorldLingo did not spell "possessor" correctly in the following translation: "This possesor is not clean." Another example of incorrect spelling is "It is allowed to take photographies?"

Translations provided by SYSTRAN and WorldLingo were the same, and these were understood by the evaluator better than those provided by SDL. The InterTran translations were considered the least understood.

Table 3 shows self-assessed ratings of understanding for the four systems using a scale of 1=no understanding to 7=complete understanding. Because SYSTRAN and WorldLingo produced the same translations, the scores were the same. In addition, an additional column containing the maximum of the SYSTRAN and SDL scores shows that greater understanding might be achieved by being presented translations from both of theses systems. For example, SDL gave superior translations in some cases, while SYSTRAN was superior in others.

Using Student's T-test, significant differences were found at $\alpha = 0.05$ between understanding ratings for SYSTRAN and InterTran, SDL and InterTran, and WorldLingo and InterTran. In addition, there was a significant difference between the ratings for SYSTRAN and the maximum ratings for SYSTRAN and SDL, indicating that both translations together are superior to just one of the two.

Table 4 shows the results of comparisons between the actual and translated sentences. SDL obtained two verbatim translations (i.e., no different than the correct translation), and SYSTRAN obtained one. SDL was obtained seven equivalent translations (different words from the correct translation, but the same meaning), and the SYSTRAN obtained four. Finally, both SDL and SYSTRAN obtained six translations with the same meaning, but grammatically incorrect. Thus, SDL was superior with 75% accuracy, while SYSTRAN and WorldLingo obtained 55%, and InterTran achieved only 10%.

The human evaluator was not able to understand (got the wrong meaning from) nine of SYSTRAN's translations. Although he rated SYSTRAN and WorldLingo superior in terms of understanding, SDL actually was more accurate it terms of what was actually meant. InterTran was clearly the worst with 18 wrong translations, primarily caused by replication of the original Spanish text in the translation.

Translation accuracies were calculated using the formula described above by an independent evaluator, and the results are shown in Table 5. There were no significant differences among the first three systems in terms of this objective measure of accuracy, but InterTran's translations were significantly poorer.

CONCLUSION

Although some studies of Web-based translation systems have been conducted before, this study is perhaps the first to provide a detailed analysis of four such system with three separate accuracy measures (self-perceived understanding, word changes, and evaluation of correct versus perceived meanings). Self-perceived measures of understanding were superior for SYSTRAN and WorldLingo, but SDL provided more accurate translations. Nevertheless, when combined with the Mezzofanti study, we consider SYSTRAN to be more accurate.

There are several limitations to the study. First, the study focused only on Spanish-to-English translation. All four Web-based services provide translations in other language pair combinations (e.g., English-to-German), and some MT systems could be better than others with different languages. Second, only 20 randomly selected samples of Spanish text were used. Other samples of text (either easier or more difficult) could provide different results. Finally, only one human evaluator was used to judge the understandability of the translated text, and only one other human evaluator was used to compare the first evaluator's understanding (as reflected in a written sentence) with the correct translation. Each human has different levels of language skill and understanding, and much subjectivity was involved. Further studies with a larger number of human evaluators is needed.

REFERENCES

Aiken, M., Ablanedo, J., Vanjani, M. (2004). An analysis of electronic meeting comment translation. *Communications of the International Information Management Association*, 4(4), 13-24.

Balkan, L., Netter, K., Arnold, D., and Meijer, S. (1994). Test suites for natural language processing. *Translating and the Computer*, London: Aslib, 51-58.

Bezhanova, O., Byezhanova, M., and Landry, O. (2005). *Comparative analysis of the translation quality produced by three MT systems*. McGill University, Montreal, Canada.

Culy, C. and Riehemann, S. (2003). The limits of N-gram translation evaluation metrics. *Proceedings of the Machine Translation Summit IX*, New Orleans, USA, September.

Falkedal, K. (1991). *Proceedings of the Evaluators' Forum*, April 21-24, Les Rasses, Vaud, Switzerland. Geneva: ISSCO.

NIST (2002). Automatic evaluation of machine translation quality using N-gram co-occurrence statistics. http://www.nist.gov/speech/tests/mt/doc/ngram-study.pdf

Papineni, K., Roukos, S., Ward, T., and Zhu, W. (2002). BLEU: A method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, July, 311-318.

Pei, M. and Vaquero, E. (1957). *Getting Along in Spanish.* Bantam, New York.

Turian, J., Shen, L., and Melamed I. (2003). Evolution of machine translation and its evaluation. *Proceedings of the Machine Translation Summit IX*, New Orleans, USA, September, 386-393.

White, J. and O'Connell, T. (1994). The ARPA MT evaluation methodologies: Evolution, lessons, and future approaches. In: *Technology Partnerships for Crossing the Language Barrier. Proceedings of the First Conference of the Association for Machine Translation in the Americas*, (5-8) October, Columbia, Maryland.