Published by: www.translationindustry.ir

An Analysis of Google Translate Accuracy

Milam Aiken, a Professor and Chair of Management Information Systems in the School of Business Administration at the University of Mississippi, U.S.A.
maiken@bus olemiss edu

Shilpa Balan, Ph.D., student in Management Information Systems at the University of Mississippi
sbalan@bus olemiss edu

Introduction

Although not appropriate for all situations, machine translation (MT) is now being used by many translators to aid their work. Many others use MT to get a quick grasp of foreign text from email, Web pages, or other computer-based material which they would not otherwise understand. Free, Web-based MT services are available to assist with this task, but relatively few studies have analyzed their accuracies. In particular, to our knowledge, there has been no comprehensive analysis of how well *Google Translate* (GT) performs, perhaps the most used system. Here, we investigate the translation accuracy of 2,550 language-pair combinations provided by this software. Results show that the majority of these combinations provide adequate comprehension, but translations among Western languages are generally best and those among Asian languages are often poor.

Background

The use of machine translation for preparation of a rough draft is a common practice among many professional translators (Champollion, 2003; Lagoudaki, 2008; O'Hagan & Ashworth, 2002), and many others use the technology to obtain the gist of foreign text because of its availability and relatively low cost (Altay, 2002). For example, it would be difficult to find a person quickly to translate a Web page in Finnish to Hindi, and the reader might only want to find out the basic content. Some professional translators might charge US $0.05 per word, and thus, a human translation of only 520 words would cost $26, far more than the reader might be willing to spend on questionable material.

Although Google Translate provides translations among a large number of languages, the accuracies vary greatly.

Article Source: translationjournal.net/journal/56google.htm

Even if a human translator is available, results can be obtained from MT much quicker. One study (Ablanedo, et al., 2007) found that a free Web-based MT system was 195 times faster than humans. Further, MT and human translation are not mutually exclusive. Once the reader has skimmed the results from the software, he or she might pay a professional if a more accurate translation is required.

Several free Web-based MT systems are available, including:

- *Applied Language*
- *Google Translate*
- *SDL Automated Translation Solutions*
- *Windows Live Translator*
- *Yahoo! Babel Fish SYSTRAN*

However, few studies have comprehensively evaluated their translation accuracy. One study (Bezhanova, et al., 2005) compared three systems and found *LogoMedia* to be best, followed by *PROMT*, and *SYSTRAN*. Another study (Aiken, et al., 2009a) compared four systems and found that *Google Translate* was best, followed by Yahoo, X10, and Applied Language. Finally, an NIST comparison of 22 MT systems in 2005 (many not free or Web-based) found that *GT* was often first and never lower than third in the rankings using text translated from Arabic to English and from Chinese to English. More detailed studies have been made of individual systems, e.g. Yahoo-SYSTRAN (Aiken, et al., 2006), but we believe *Google Translate* is used more frequently, provides more language-pair combinations, and is probably more accurate overall, and therefore, we will be focusing on it (Aiken & Ghosh, 2009; Och, 2009).

Automatic Evaluation of GT

A few studies have attempted to assess *GT*'s accuracy (e.g., Aiken, et al., 2009b), but to our knowledge, there has been no published, comprehensive evaluation of *Google Translate*'s accuracy, i.e., an analysis of all language pairs. Because it is impractical to obtain human translators for all 51 languages to analyze 50 passages of text each, automatic evaluation is necessary.

Although several techniques exist, BLEU (Bilingual Evaluation Understudy) is perhaps the most common (Papineni, et al., 2002), and some studies have shown that it achieves a high correlation with human judgments of quality (Coughlin, 2003; Culy & Richemann, 2003). Using this technique, scores ranging from 0 to 100 are calculated for a sample of translated text by comparing it to a reference translation, and the method takes into account the number of words that are the same in the two passages as well as the word order.

To help judge the appropriateness of using BLEU to analyze 2,550 (51 x 50) language pairs, we had two independent evaluators assess the comprehensibility of 50 non-English text samples translated to English with *GT*. (Note: At the time of the evaluation, GT supported only 51 languages. Armenian, Azerbaijani, Basque, Georgian, Haitian Creole, Latin, and Urdu have since been added to make a total of 58 languages supported.) The equivalent text for each of the following sentences (Flesch Reading Ease score of 81.6 on a scale of 0=hard to100=easy, Flesch-Kincaid grade level of 3.6 on a scale of 1 to 14) was obtained for all of the 50 non-English languages from Omniglot:

1. Pleased to meet you.
2. My hovercraft is full of eels.
3. One language is never enough.
4. I don't understand.
5. I love you.
6. All human beings are born free and equal in dignity and rights. They are endowed with reason and conscience and should act towards one another in a spirit of brotherhood.

The scores from Evaluator 1 and Evaluator 2 were significantly correlated (R = 0.527, p < 0.001), and the scores from each evaluator were also significantly correlated with the BLEU scores calculated using Asia Online's *Language Studio Lite* software (Evaluator 1: R = 0.789, p < 0.001; Evaluator 2: R = 0.506, p < 0.001). In addition, BLEU scores were significantly correlated (R = 0.499, p = 0.003) with comprehension measures from Aiken & Vanjani (2009) (R = 0.447, p = 0.010). Thus, we believe BLEU scores give a good indication of how well humans would understand the translated text.

We used *Language Studio Lite* to calculate BLEU scores for each of 2,550 translations obtained with *GT* for the text above. For example, the equivalent text for the six items was retrieved for French, translated to German with *GT*, and then, the BLEU score was calculated to see how well the translation matched the equivalent German phrases. The table of 2,550 BLEU scores can be found HERE. In this table, the source languages appear as row headings, and the destination languages appear as column headings. For example, the text in Arabic (source) was translated to Afrikaans (destination), and this translation was compared to the equivalent text in Afrikaans, resulting in a BLEU score of 46.

Next, we averaged the translations for each language pair to obtain an overall measure of how understandable translations between two languages would be. For example, the BLEU score for Icelandic to Bulgarian is 42, and the score for Bulgarian to Icelandic is 49, giving a final average of 45.5. The sorted list of 1,275 combined language pairs can be found HERE. (Note: The language order in this list is random. That is, Japanese-Malay is the same as Malay-Japanese.)

Comprehension Sufficiency

The language pair ranking gives an indication of the relative accuracy of translations among various language pairs, but it does not provide information on how adequate they are. For example, would translations between a pair of languages with an average BLEU score of 50.0 be sufficiently understandable, or is a minimum score of 70.0 required? Obviously, a higher standard should be placed on translations of legal, financial, medical, or other critical information, and a lower standard is probably alright for gathering the gist of informal, relatively unimportant material.

One standard is the reading comprehension score from the Test of English as a Foreign Language (TOEFL) required by many universities in the United States for students whose primary language is not English. For example, UCLA's graduate program requires a minimum score of 21 out of 30, while Auburn's MBA program requires a minimum of 16. In one study (Aiken, et al., 2011), 75 American students whose primary language was English took reading comprehension tests comprised of TOEFL passages in Chinese, German, Hindi, Korean, Malay, and Spanish translated to English using *Google Translate*. Results showed an average reading score of 21.90, just above the 21 minimum required by UCLA's graduate program, indicating that the comprehension of these translations was, on average, sufficient for material that a graduate student might encounter during the course of studies. The corresponding average BLEU score for these six tests was 19.67.

However, the material in our analysis was easier (Flesch Reading Ease = 81.6: Grade = 3.6 versus the TOEFL tests' Reading Ease = 63.5 and Grade Level = 8.3), and therefore, the average BLEU score of the six items above translated to English from Chinese, German, Hindi, Korean, Malay, and Spanish was much higher (58.83). If we assume a linear relationship between the BLEU and TOEFL reading comprehension scores, the corresponding TOEFL score for the six items translated to English from the six languages would be 24.5 out of 30. The adjusted minimum TOEFL reading scores for this easier material for UCLA would be 26.2 and for Auburn would be 20.0. Using 26.2 as the standard, 737 of the 1,275 language combinations would be sufficient for comprehension of graduate college material at UCLA, and 865 would be sufficient using Auburn's MBA standard.

Conclusion

Although *Google Translate* provides translations among a large number of languages, the accuracies vary greatly. This study gives for the first time an estimate of how good a potential translation might be using the software. Our analysis shows that translations between European languages are usually good, while those involving Asian languages are often relatively poor. Further, the vast majority of language combinations probably provide sufficient accuracy for reading comprehension in college.

There are several limitations to the study, however. First, a very limited text sample was used due to the difficulty of acquiring equivalent text for 50 different languages. Other, more complicated text samples are likely to result in lower BLEU scores. On the other hand, only one reference text was used in the calculations, again, due to the problem of obtaining similar passages. That is, each translation was compared to only one "correct" result. Other acceptable translations using alternative wording and synonyms would result in higher BLEU scores. Finally, human judgments of comprehension are usually preferable to automatic evaluation, but in this case, it was impractical due to the many language combinations that had to be assessed.

Finally, this evaluation is not static. *Google Translate* continually adds new languages, and the existing language translation algorithm is constantly improved as the software is trained with additional text and volunteers correct mistranslations. Although its performance is never likely to reach the level of an expert human's, it can provide quick, cheap translations for unusual language pairs.

References

1. Ablanedo, J., Aiken, M., and Vanjani, M. (2007). Efficacy of English to Spanish automatic translation. *International Journal of Information and Operations Management Education*, 2(2), 194-210.
2. Aiken, M. and Ghosh, K. (2009). Automatic translation in multilingual business meetings. *Industrial Management & Data Systems*, 109(7), 916-925.
3. Aiken, M., Ghosh, K., Wee, J., and Vanjani, M. (2010a). Aiken, M., Ghosh, K., Wee, J., and Vanjani, M. (2009a). An evaluation of the accuracy of online translation systems. Communications of the IIMA, 9(4), 67-84. *Communications of the IIMA*, 9(4), 67-84, in press. http://findarticles.com/p/articles/mi_7099/is_4_9/ai_n56337599/
4. Aiken, M., Park, M., Simmons, L., and Lindblom, T. (2009b). Automatic translation in multilingual electronic meetings. *Translation Journal*, 13(9), July.
5. Aiken, M., Vanjani, M., and Wong, Z. (2006). Measuring the accuracy of Spanish-to-English translations. *Issues in Information Systems*, 7(2), 125-128.
6. Aiken, M. and Vanjani, M. (2009). Polyglot: A multilingual group support system. *Issues in Information Systems*, 10(2), 101-106.
7. Aiken, M., Wang, J., Wu, L., & Paolillo, J. (2010b). An exploratory study of multilingual electronic communication, *International Journal of e-Collaboration (IJeC)*, 7(1), 17-29.
8. Altay, D. (2002). Difficulties encountered in the translation of legal texts: The case of Turkey. *Translation Journal*, 6(4).
9. Bezhanova, O., Byezhanova, M., and Landry, O. (2005). *Comparative analysis of the translation quality produced by three MT systems*. McGill University, Montreal, Canada.
10. Champollion, Y. (2003). Convergence in CAT: Blending MT, TM, OCR & SR to boost productivity. *Proceedings of the International Conference Translating and the Computer 25*, 20-21 November 2003, London. London: Aslib.
11. Coughlin, D. (2003) "Correlating Automated and Human Assessments of Machine Translation Quality" in *MT Summit IX, New Orleans, USA*, 23-27.
12. 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.
13. Lagoudaki, E. (2008). The value of machine translation for the professional translator. *AMTA-2008. MT at work: Proceedings of the Eighth Conference of the Association for Machine Translation in the Americas*, Waikiki, Hawaii, 21-25 October, 262-269.
14. Och, F. (2009). 51 Languages in Google Translate. Google Research Blog, August 31.
15. O'Hagan, M. and Ashworth, D. (2002). Translation-mediated communication in a digital world: Facing the challenges of globalization and localization. *Topics in Translation*. London: Multilingual Matters.
16. Papineni, K., Roukos, S., Ward, T., and Zhu, W. J. (2002). BLEU: A method for automatic evaluation of machine translation in *ACL-2002: 40th Annual meeting of the Association for Computational Linguistics*, 311-318.