Studies in Linguistics and Literature
ISSN 2573-6434 (Print) ISSN 2573-6426 (Online)
Vol. 3, No. 3, 2019
www.scholink.org/ojs/index.php/sll

Original Paper

An Updated Evaluation of Google Translate Accuracy

Milam Aiken[1*]

[1] School of Business Administration, University of Mississippi, Oxford, MS, USA

[*] Milam Aiken, School of Business Administration, University of Mississippi, Oxford, MS, USA

Received: June 27, 2019 Accepted: July 5, 2019 Online Published: July 17, 2019

doi:10.22158/sll.v3n3p253 URL: http://dx.doi.org/10.22158/sll.v3n3p253

Abstract

In 2011, a comprehensive evaluation of accuracy using 51 languages with Google Translate showed that many European languages had good results, but several Asian languages performed poorly. The online service has improved its accuracy over the intervening eight years, and a reevaluation using the same text as the original study shows a 34% improvement based upon BLEU scores. This new study shows that translations between English and German, Afrikaans, Portuguese, Spanish, Danish, Greek, Polish, Hungarian, Finnish, and Chinese tend to be the most accurate.

Keywords

Google Translate, machine translation, accuracy, BLEU, computational linguistics

1. Introduction

Google Translate (https://translate.google.com/) is perhaps the most used online translation service, and as of 2016, over 500 million people were translating 100 billion words with it each day in 103 languages (Kroulek, 2016). About 92% of its users are outside the United States (Brazil has the highest number), and the most requested translations are from English to Arabic, Russian, Portuguese, and Indonesian.

Google Translate (GT) was introduced in 2006 employing a statistical, Phrase-Based Machine Translation (PBMT) model, and the service was updated in 2016 with a Neural Machine Translation (NMT) model (McGuire, 2018). With the change, Google CEO Sundar Pichai stated that its machine translation improved from a score of 3.694 (out of 6) to 4.263, nearing human-level quality at 4.636 (Turner, 2016; Vincent, 2016). In addition, researchers at Google translated 500 randomly sampled sentences from Wikipedia and news websites using four languages and reported the following improvements in accuracy: English to Spanish (87%), English to French (64%), English to Chinese (58%), Spanish to English (63%), French to English (83%), and Chinese to English (60%), for an average improvement of 69% for all pairs (Wu et al., 2016). However, this test evaluated only a small portion of the languages provided.

This paper reports on a more comprehensive evaluation of how GT has improved using a broader range of languages. These results show that accuracy has probably improved by about 34% using the 51 languages that were available eight years ago. Using the same language pairs in the Google study but with different text, results show about 70% improvement which is consistent with the 69% reported.

2. Prior Evaluation Studies

Several small studies of GT have been conducted using a few languages. For example, one study (Chen et al., 2016) translated a pamphlet with Google from English to Spanish and Chinese. Results showed that there was a more accurate translation from English to Spanish than English to Chinese, and a Spanish human translator did not provide a significantly better translation than GT. The study also noted that the likelihood of incorrect translation increased with text reading difficulty.

In another study (Patil & Davies, 2014), researchers translated 10 medical phrases with GT using 26 languages. Results showed that 57.7% of the translations were correct with two African languages scoring the worst (45% correct) and eight Western European languages scoring the best (74%). Swahili scored lowest with only 10% correct, while Portuguese scored highest at 90%.

In perhaps the first comprehensive study of GT (Aiken & Balan, 2011), equivalent non-English text was obtained from Omniglot (http://www.omniglot.com/) for the following:

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.

Next, BLEU scores were calculated for all 2,550 pair combinations of the 51 languages then supported. Because human experts are often not available to evaluate the accuracy of translations, various scoring techniques have been devised, and BLEU (Bilingual Evaluation Understudy) is perhaps the most frequently used. With this method a translation is compared with one or more acceptable translations and it looks at the presence or absence of particular words, as well as the ordering and the degree of distortion (Pan, 2016). The result is a score from 0 to 1, 0 to 10, or 0 to 100 with the higher number representing a better translation. That is, a score of 100 means the reference text and translation are identical.

BLEU has several limitations, however. One is that there may be several different acceptable translations using synonyms that are not included in the reference text. Also, there is a possibility that an ungrammatical sequence of words in the wrong order can achieve a high score. A translation might have a low BLEU score and still be correct, or it might have a high BLEU score and be wrong. The score just gives an indication of accuracy, and just because a language pair has achieved high scores for other text does not guarantee a new translation with other text will be good.

Results of the Aiken & Balan study showed that that translations between European languages (e.g., German to Spanish) were good, while those involving Asian languages were usually poor. However, the vast majority of language combinations provided sufficient accuracy for reading comprehension with a college admission test.

Another study was conducted using the newer version of Google Translate with neural machine translation (Benjamin, 2018) using 20 English phrases: fly out of London, like a bat out of hell, out cold, out of bounds, out of breath, out of curiosity, out of focus, out of his mind, out of milk, out of order, out of pocket, out of steam, out of style, out of the closet, out of the game, out of the office, out of this world, out of time, out of wedlock, and out on the town. These phrases were translated to the 102 non-English languages supported by GT at the time, and fluent speakers of these languages evaluated the translations using three scores (Bard-percentage of the text free of grammatical and word-choice errors, with 100 representing human quality, Tarzan-the percentage of text that can be understood, and Fail-the percentage of text that is completely wrong). Data for this study are recorded at http://bit.ly/gt_scores. Results show that 50 languages achieved a Bard score of 50 or above, and 41 languages received a Tarzan score of 50 or above. Bengali, Haitian Creole, and Tajik failed 100% of the time, and these languages failed 80% or more of the time: Kurdish, Nepali, Latin, Malaysian, Urdu, Maori, Cebuano, Georgian, Persian, Punjabi, and Uzbek.

3. Update Study

The purpose of this research is to determine how much Google Translate has improved in translation accuracy over the past eight years. First, as a comparison, a subset of the 102 Bard, Tarzan, and Fail scores from the Benjamin study were recorded for the 50 non-English languages from the Aiken & Balan study and recorded in Table 1 below. As expected, there was a significant, positive correlation between the Bard and Tarzan scores (R=0.935, p<0.001) and a significant, negative correlation between the Tarzan and Fail scores (R=-0.935, p<0.001).

In Table 1, BLEU1 shows the BLEU scores from the Aiken & Balan study, using the older, PBMT version of GT. That is, for example, the translation from English to Afrikaans gave a score of 49. Because the software used to calculate BLEU scores in the earlier study is no longer available, new BLEU scores were calculated using Tilde Custom Machine Translation's Interactive BLEU score evaluator (https://www.letsmt.eu/Bleu.aspx), and to keep the evaluation consistent with the prior study, the 1-gram method was selected. BLEU2 shows the BLEU scores calculated with the same English text translated to the foreign language with the newer, NMT version of GT. For example, the English text translated to Dutch gave a BLEU score of 71. There was a significant, positive correlation between the old GT scores and the new GT scores using the text from the Aiken & Balan study (R=0.619, p<0.001), and the mean BLEU score was 34.4% higher with the new system. Further, there was a significant difference between the two sets of BLEU scores (T=-5.16, p<0.001). Several languages (e.g., Arabic, Hindi, and Serbian) showed large increases in scores, but a few (e.g., Chinese, Japanese, and Swahili) saw decreases.

Table 1. Google Translate Study Scores

	Language	Bard	Tarzan	Fail	BLEU1	BLEU2	Tarzan2	BLEU3
1	Afrikaans	67.5	87.5	13	49	71	90	83
2	Albanian	26.25	40	60	56	70	90	80
3	Arabic	32.5	40	60	0	60	85	76
4	Belarusian	40	55	45	29	57	87	83
5	Bulgarian	40	60	40	73	79	98	80
6	Catalan	37.5	60	40	56	75	90	70
7	Chinese	55	65	35	17	8	95	78
8	Croatian	55	65	35	5	83	85	77
9	Czech	43.75	55	45	55	64	100	86
10	Danish	40	70	30	84	84	100	82
11	Dutch	52.5	65	35	82	71	95	84
12	Estonian	27.5	45	55	62	58	98	70
13	Filipino	25	35	65	52	65	85	70
14	Finnish	45	65	35	73	82	98	77
15	French	45	60	40	91	89	95	88
16	Galician	52.5	60	40	49	36	77	73
17	German	60	82.5	18	77	72	99	81
18	Greek	52.5	70	30	68	67	99	75
19	Hebrew	45	45	55	35	54	77	67
20	Hindi	22.5	30	70	0	42	80	55
21	Hungarian	47.5	65	35	38	58	100	70
22	Icelandic	20	30	70	48	66	90	72
23	Indonesian	47.5	60	40	94	81	90	82
24	Irish	25	45	55	21	56	85	69
25	Italian	50	60	40	87	100	99	90
26	Japanese	42.5	55	45	26	8	83	80
27	Korean	42.5	55	45	27	56	95	75
28	Latvian	50	60	40	12	73	90	77

29	Lithuanian	20	30	70	43	63	75	61
30	Macedonian	35	60	40	11	50	82	61
31	Malaysian	5	10	90	10	72	95	76
32	Maltese	35	60	40	57	63	87	84
33	Norwegian	27.5	45	55	66	83	96	75
34	Persian	10	20	80	16	39	98	66
35	Polish	56.25	72.5	28	66	79	93	84
36	Portuguese	55	75	25	64	75	100	91
37	Romanian	42.5	55	45	68	53	80	84
38	Russian	40	50	50	68	74	92	84
39	Serbian	47.5	55	45	0	52	95	75
40	Slovak	27.5	45	55	38	68	93	83
41	Slovenian	40	60	40	85	68	82	72
42	Spanish	56.2	75	25	67	78	98	80
43	Swahili	15	25	75	76	61	62	70
44	Swedish	45	60	40	87	85	99	86
45	Thai	15	25	75	0	25	83	68
46	Turkish	40	50	50	53	74	96	75
47	Ukrainian	30	40	60	67	80	92	67
48	Vietnamese	20	35	65	23	71	96	72
49	Welsh	43.75	62.5	38	40	59	90	75
50	Yiddish	27.5	50	50	0	30	55	60

As discussed earlier, GT is used most often to translate English to Arabic, Russian, Portuguese, and Indonesian. Results show a large increase in the BLEU score for Arabic and moderate increases for Russian and Portuguese. Indonesian decreased 14.6%, but overall, there was a 35.6% increase in BLEU scores for these languages.

BLEU3 shows the BLEU scores with the foreign text translated to English with the newer version of GT. For example, the text in Hungarian translated to English gave a BLEU score of 70. As reported earlier, Google has reported a 69% improvement in accuracy using six language pairs. With the same six language pairs using different text, results show a 70.2% improvement, very close to Google's estimate.

Tarzan 2 shows the human comprehension scores for the non-English Aiken & Balan text translated to English. The German to English translation resulted in the following:

1) Nice to meet you.

2) My hovercraft is full of eels.

3) A language is never enough.

4) I do not understand.

5) I love you.

6) All human beings are free and born equal in dignity and rights. They are gifted with reason and conscience and are to meet each other in the spirit of brotherhood.

Items 3 and 6 were not identical and led to a little ambiguity, resulting in a BLEU score of 81 and a Tarzan score of 99.

The translation from Swahili to English gave:

1) I am happy to see you.

2) My crash car is full of caterpillars.

3) One luga is not enough.

4) I don’t know.

5) I love you.

6) They are all born free and equal in respect and justice. They are endowed with intelligence and conscience, and must act toward one another in the spirit of brotherhood.

Only item 5 was identical to the reference text, and there were several grammatical and word-choice errors, resulting in a Tarzan score of 62.

Many of the translations to different languages did not have an equivalent word for “hovercraft” and many had grammatical problems, especially with item 6. Several translations had variations of “nice to meet you” that while correct and understandable (e.g., “pleasure”), were not identical and thus resulted in lower BLEU scores.

There was a moderate, significant correlation between the Tarzan scores in the Benjamin study (English->foreign: Benjamin text) and (foreign->English: Aiken & Balan text) ($R=0.329$, $p<0.20$). There was also a significant correlation between the Tarzan and BLEU3 scores using the Aiken & Balan text ($R=0.530$, $p<0.001$). Table 2 shows descriptive statistics for these variables.

Averaging the Tarzan scores between the two studies, the top 10 languages for translation accuracy with English are (from best to worst): German, Afrikaans, Portuguese, Spanish, Danish, Greek, Polish, Hungarian, Finnish, and Chinese. The bottom 10 (from worst to best) are: Swahili, Yiddish, Malaysian, Lithuanian, Thai, Hindi, Persian, Icelandic, Filipino, and Hebrew.

Table 2. Descriptive Statistics

	Minimum	Maximum	Mean	Std. Deviation
Bard	5.00	67.5	38.47	13.98
Tarzan	10.0	87.5	52.90	16.26
Fail	12.5	90.0	47.10	16.26
BLEU1	0	94	47.42	28.41
BLEU2	8	100	63.74	18.94
Tarzan2	55	100	89.88	9.52
BLEU3	55	91	75.98	8.04

4. Conclusion

Although there are several limitations with this study (e.g., use of BLEU scores as a surrogate for comprehension, limited text sample size, etc.) (Hofstadter, 2018), results show that Google Translate accuracy has increased about 34% using 51 of its current 103 languages. With the same six language pairs used in the Google study, results with different text show remarkably similar results, bolstering confidence in these conclusions. Further studies should be conducted to see if the improvement in accuracy continues.

References

Aiken, M., & Balan, S. (2011). An analysis of Google Translate accuracy. *Translation Journal*, *16*(2), April. Retrieved July 8, 2019, from https://translationjournal.net/journal/56google.htm

Benjamin, M. (2018). Which languages are Google Translate best at translating? *Kamusi Project International*. Retrieved July 8, 2019, from https://www.quora.com/Which-languages-are-Google-Translate-best-at-translating

Chen, X., Acosta, S., & Barry, A. (2016). Evaluating the accuracy of Google Translate for diabetes education material. *JMIR Diabetes*, *1*, e3. https://doi.org/10.2196/diabetes.5848

Hofstadter, D. (2018). The shallowness of Google Translate. *The Atlantic*. Retrieved July 8, 2019, from https://www.theatlantic.com/technology/archive/2018/01/the-shallowness-of-google-translate/551570/

Kroulek, A. (2016). *11 Google Translate Facts you should know*. Retrieved July 8, 2019, from https://www.k-international.com/blog/google-translate-facts/

McGuire, N. (2018). How Accurate Is Google Translate in 2018? *Argo Translation*. Retrieved July 8, 2019, from https://www.argotrans.com/blog/accurate-google-translate-2018/

Pan, H. (2016). How BLEU measures translation and why it matters. *Slator*. Retrieved July 8, 2019, from https://slator.com/technology/how-bleu-measures-translation-and-why-it-matters/

Patil, S., & Davies, P. (2014). Use of Google Translate in medical communication: Evaluation of accuracy. *BMJ*, *349*, 1-3. https://doi.org/10.1136/bmj.g7392

Turner, K. (2016). Google Translate is getting really, really accurate. *Washington Post*, *October 3*. Retrieved July 8, 2019, from https://www.washingtonpost.com/news/innovations/wp/2016/10/03/google-translate-is-getting-really-really-accurate/

Vincent, J. (2016). Apple boasts about sales; Google boasts about how good its AI is. *The Verge*. Retrieved July 8, 2019, from https://www.theverge.com/2016/10/4/13122406/google-phone-event-stats

Wu, Y., Schuster, M., Chen, Z., Le, Q., Norouzi, M., Macherey, W., & Dean, J. (2016). *Google's neural machine translation system: Bridging the gap between human and machine translation*. Retrieved July 8, 2019, from https://arxiv.org/pdf/1609.08144v1.pdf