Journal of Management Science and Business Intelligence, 2020, 5–1
July. 2020, pages 26-31
doi: 10.5281/zenodo.3960835
http://www.ibii-us.org/Journals/JMSBI/
ISBN 2472-9264 (Online), 2472-9256 (Print)

A Comparison of Free Online Machine Language Translators

Mahesh Vanjani

Jesse H. Jones School of Business, Texas Southern University, 3100 Cleburne Street, Houston, TX 77004

Email: Mahesh.Vanjani@tsu.edu

Received on 5/15/2020; revised on 7/25/2020; published on 7/26/2020

Abstract

Automatic Language Translators also referred to as machine translation software automate the process of language translation without the intervention of humans While several automated language translators are available online at no cost there are large variations in their capabilities. This article reviews prior tests of some of these systems, and, provides a new and current comprehensive evaluation of the following eight: Google Translate, Bing Translator, Systran, PROMT, Babylon, WorldLingo, Yandex, and Reverso. This research could be helpful for users attempting to explore and decide which automated language translator best suits their needs.

Keywords: Automated Language Translator, Multilingual, Communication, Machine Translation

1 Introduction

Automatic Language Translators also referred to as machine translation software automate the process of language translation without the intervention of humans. Text from the source language is translated to text in the target language. The most basic automatic language translators strictly rely on word-for-word substitution. Some may include rule-based or statistically-modeled translation for more accurate translations (G2 Crowd, 2019). However, modern automatic language translators have come a long way with vastly improved translation capability.

Automatic Language Translators generally organize content around thematic subjects to make it feasible to access or learn commonly used expressions or phrases that obey proper rules of syntax and grammar (Harris, 2012). . For example, automatic language translators catering to international travelers would organize content based on airports, hotels, restaurants and key phrases in the local language. Electronic automated language translators can be viewed as digital phrase books. They store lots of ready-made phrases in the device's memory, enabling users to query the database and return results based on the search parameters. Most such language translators are "intelligent" in that they learn from use and are able to expand their knowledgebase to improve translation capability and quality.

While several free automated language translators are available online there are large variations in their capabilities. For this article we reviewed prior tests of some of these systems. Based on our research of published results of prior tests and experiments we present a new and current comprehensive evaluation of the following eight electronic automatic language translators: Google Translate, Bing Translator, Systran, PROMT, Babylon, WorldLingo, Yandex, and Reverso. This research could be helpful for users attempting to explore and decide which automated language translator best suits their needs.

2 Prior Comparisons using Human Review of Text

Several studies of online translation systems have been conducted using humans to review the quality of results. Some of these are summarized chronologically below:

- The National Institute of Standards and Technology compared 22 machine translation systems in 2005 (many were not free or Web-based) and found that Google Translate was often first and never lower than third in the rankings using text translated from Arabic to English and from Chinese to English (NIST, 2005).

- In a study comparing free, online systems, 17 English sentences were translated into Spanish using LogoMedia, Systran, and PROMT (Bezhanova, et al., 2005). All three produced usable translations, but Systran translations were generally the worst.

- A study using English and Spanish with one human evaluator found that Systran, SDL, and WorldLingo provided roughly equal results, and InterTran was substantially worse (Aiken & Wong, 2006).

- A sample of 10 German-to-English and 10 Spanish-to-English translations from four systems were reviewed by two evaluators in another study (Aiken, et al., 2009), and results showed that Google was best, followed by Systran and X10 (tie) and then Applied Language.

- Hampshire and Salvia (2010) evaluated 10 systems using English and Spanish and found that Google Translate was best, followed by Babylon, Reverso, Bing, Babelfish, Systran, PROMT, WorldLingo, InterTran, and Webtrance.

- In a survey of people using three translation systems (Shen, 2010), results showed that Google Translate was preferred when translating long passages, but Microsoft Bing Translator and Yahoo Babelfish often produced better translations for phrases below 140 characters. Babelfish performed well with East Asian Languages such as Chinese and Korean and Bing Translator performed well with Spanish, German, and Italian.

- Oliveira & Anastasiou (2011) compared Google Translate and Systran using English and Portuguese with a point system and found that Google was substantially better (107 points to 46).

- Papula (2014) also used a point system with English and Spanish and found that Microsoft Bing was best with a value of 68.4, followed by Systran (67.6), Promt (67.5), and Google Translate (67.3). Using English and Portuguese, the rankings were: Google Translate (60.9), Microsoft Bing (60.6), PROMT (58.4), and Systran (58.1).

- Four human translators compared three popular machine translation programs(Google Translate, Systran's translation program, and the Papago app from Naver) using English and Korean in another study (Brooks, 2017), and the results showed that Google Translate was best, followed by the Papago App, and Systran.

- In another study (G2 Crowd, 2019), users reported liking Google Translate best with a score of 4.6 out of 6, followed by Systran (4.5), Bing (3.7), Yandex (4.0), and Babylon (3.0).

- In a final study (Himmelein, 2019), English and German were used to compare Babylon, DeepL*, Google, Bing, PROMT, Systran, and WordLingo. Results showed that Babylon, Systran and WordLingo were the worst, and Google Translate was more accurate than Bing.

3 Automatic Evaluation of Text

Because humans fluent in many languages are often not available to evaluate translations from systems, the studies above used only a few languages and evaluators, resulting in poor statistical reliability. To address this problem, automatic evaluation techniques such as BLEU (Bilingual Evaluation Understudy) are sometimes used. With this technique, a translation is compared with one or more acceptable translations and it looks for the presence or absence of particular words, as well as the ordering (Pan, 2016).

The BLEU score was proposed by Kishore Papineni, et al. in their 2002 paper "BLEU: a Method for Automatic Evaluation of Machine Translation". As the authors noted, human evaluations of machine translation are extensive but very expensive. Depending on the availability of qualified translators human evaluations can take months to finish at the expense of human labor that cannot be reused or duplicated. The authors proposed a method of automatic machine translation evaluation that is quick, inexpensive, and, language-independent. The metric referred to as the Bilingual Evaluation Understudy Score, or BLEU for short, and, correlates highly with human evaluation, and that has little marginal cost per run.

BLEU scores can be used for evaluating an automatic translator generated sentence in the target language to the sentence in the source language. A perfect match results in a score of 1.0, whereas a perfect mismatch results in a score of 0.0. It correlates highly with human evaluation. While not perfect, it has been widely adopted due to the following advantages:

- It is quick and inexpensive to calculate.
- It is easy to understand.
- It is language independent.
- It correlates highly with human evaluation.
- It has been widely adopted.

BLEU has limitations, however. For example, different, acceptable translations might be omitted as reference text. In addition, the score reveals little about a particular passage. That is, an incomprehensible sequence of words could achieve a high score, but a perfectly understandable translation might obtain a low score. The score just gives an indication of accuracy, not an absolute measure. That is, there is no guarantee that an increase in BLEU score is an indicator of improved translation quality (Callison-Burch, et al., 2006).

Some critics have argued that BLEU scores are inaccurate and perhaps even worthless (e.g. Scarton & Specia, 2016). One study (Turian, et al., 2003) showed that the correlation between human judges and automatic measures of translation quality was low. However, other studies (e.g., Coughlin, 2003,. Culy & Richemann, 2003; and Papineni, et al., 2002) have found high correlations.

While researchers must use caution in applying the results, BLEU can be relevant within certain constrained conditions, e.g. comparisons of systems with a certain, specified sample of text. For example, Savoy and Dolamic (2009) evaluated three free, online systems using 117,452 documents translated from French to English and found that that Google Translate was most accurate followed by Babel Fish and then PROMT.

4 An Updated Automatic Evaluation of Systems

Some prior studies used online translation programs that are no longer available, and several online services are extremely limited in the number of languages supported. For example, DeepL Translator (https://www.deepl.com/en/translator) supports only nine languages, and Linguee (https://www.linguee.com/) supports only eight. However, the eight systems in Table 1 are currently available and provide support for at least a dozen languages.

Table 1: A sample of free, online translation systems

Translator	URL	Languages
Google Translate	https://translate.google.com/	103
Yandex	https://translate.yandex.com/	81
Bing Translator	https://www.bing.com/translator	61
Systran	https://translate.systran.net/translationTools/text	41
Babylon	https://translation.babylon-software.com/	23
PROMT	https://www.online-translator.com/	20
WorldLingo	http://www.worldlingo.com/	14
Reverso	http://www.reverso.net/text_translation.aspx	14

In addition, previous tests have frequently been conducted with humans reviewing the quality. While this might provide more accurate results, it necessarily limits the number of language comparisons as fluent speakers are difficult to obtain.

Some automatic evaluations have used the following text from www.omniglot.com:

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.

In one study (Aiken & Balan, 2011), two humans evaluated the equivalent text from 50 non-English languages translated to English, and results showed there was a significant, positive correlation between comprehension and BLEU scores (Evaluator 1: $R = 0.789$, $p < 0.001$; Evaluator 2: $R = 0.506$, $p < 0.001$). In another study (Aiken, 2019), there was a significant, positive correlation ($R=0.530$, $p < 0.001$) between one reviewer's comprehension of the translated text sample and BLEU scores. Therefore, for this limited text, the use of BLEU scores appears to be appropriate for a comparison of systems.

This new study compared the eight translation systems listed in Table 1 using the seven most-spoken languages in the world (Chinese, English, Hindi, Spanish, Arabic, Malay, and Russian). BLEU scores were calculated using Tilde Custom Machine Translation's Interactive BLEU score evaluator (https://www.letsmt.eu/Bleu.aspx). Table 2 shows an example with BLEU scores of Chinese-to-English translations.

Table 2: Translations from Chinese to English for "My hovercraft is full of eels."

System	Translation to English	BLEU Score
Yandex	My hovercraft is full of eels.	100
Google	My hovercraft is full of trout.	87
Bing	My hovercraft is full of mackerel.	87
Systran	My hovercraft is full of eel.	87
PROMT	My hovercraft, is filled with Eels.	66
Babylon	My air pad fitted boat full of 鱔 fish.	45
WorldLingo	My hovercraft has packed the finless eel.	44
Reverso	My air cushion ship was filled with eels.	40

Tables 3 -10 show average BLEU scores calculated from source languages (row headings) and targets (column headings) as follows:

Table 3: Google Translate BLEU scores
Table 4: Bing Translator BLEU scores
Table 5: Systran BLEU scores
Table 6: PROMT Online BLEU scores
Table 7: WorldLingo BLEU scores
Table 8: Reverso BLEU scores
Table 9: Babylon BLEU scores
Table 10: Yandex BLEU scores

There was no translation to Chinese because the software was not able to calculate scores for that character set. Overall average scores are shown in the cell at the bottom right of each table.

Only Google, Bing, and Yandex translated all seven source languages, with resulting overall average scores of 58.9, 57.1, and 56.6, respectively, and there was no significant difference between Google and Bing ($p = 0.65$) or between Google and Yandex ($p = 0.60$).

Babylon and PROMT did not support Malay, and WorldLingo and Reverso did not support Malay and Hindi. Systran supported only English as a source language consistently.

Eliminating Malay and Hindi from the analysis (and disregarding Systran) gave the following overall averages: Yandex – 68.5, Google – 67.6, Bing - 64.3, Babylon – 47.0, PROMT - 46.3, Reverso – 44.5, and WordLingo – 39.9. Again, there were no significant differences among the top three systems: Yandex-Google ($p = 0.68$) and Yandex-Bing ($p = 0.36$).

Table 3: Google Translate BLEU scores

Source/ Target	English	Hindi	Spanish	Arabic	Malay	Russian	Mean
Chinese	78	38	74	39	67	71	61.2
English		42	71	56	72	74	63
Hindi	55		53	30	62	31	46.2
Spanish	80	35		43	69	68	59
Arabic	76	40	67		62	83	65.6
Malay	76	40	69	49		48	56.4
Russian	84	39	65	51	65		60.8
Mean	74.8	39.0	66.5	44.7	66.2	62.5	58.9

Table 4: Bing Translator BLEU scores

Source/ Target	English	Hindi	Spanish	Arabic	Malay	Russian	Mean
Chinese	82	44	70	38	65	66	60.8
English		43	83	45	66	70	61.4
Hindi	64		53	24	58	43	48.4
Spanish	86	43		34	59	62	56.8
Arabic	68	39	57		56	63	56.6
Malay	73	38	67	32		56	53.2
Russian	90	38	77	42	65		62.4
Mean	77.2	40.8	67.8	35.8	61.5	60	57.1

Table 5: Systran BLEU scores (X = not supported)

Source/ Target	English	Hindi	Spanish	Arabic	Malay	Russian	Mean
Chinese	79	X	74	X	X	X	76.5
English		42	76	49	71	59	59.4
Hindi	78		58	X	X	X	68
Spanish	81	30		X	61	X	57.3
Arabic	80	X	X		X	X	80
Malay	74	X	60	X	X	X	67
Russian	87	X	X	X	X		87
Mean	79.8	36	67	49	66	59	65.1

Table 6: PROMT Online BLEU scores (X = not supported)

Source/ Target	English	Hindi	Spanish	Arabic	Malay	Russian	Mean
Chinese	71	31	59	38	X	11	42
English		42	61	55	X	13	42.8
Hindi	50		41	29	X	10	32.5
Spanish	64	36		33	X	11	36
Arabic	80	35	53		X	14	45.5
Malay	X	X	X	X	X	X	X
Russian	79	38	40	58	X		53.8
Mean	68.8	36.4	50.8	42.6	X	11.8	42.1

Table 7: WorldLingo BLEU scores (X = not supported)

Source/ Target	English	Hindi	Spanish	Arabic	Malay	Russian	Mean
Chinese	54	X	38	24	X	35	37.8
English		X	60	26	X	55	47
Hindi	X	X	X	X	X	X	X
Spanish	70	X		21	X	42	44.3
Arabic	42	X	27		X	30	33
Malay	X	X	X	X	X	X	X
Russian	56	X	30	27	X		37.7
Mean	55.5	X	38.8	24.5	X	40.5	39.9

Table 8: Reverso BLEU scores (X = not supported)

Source/ Target	English	Hindi	Spanish	Arabic	Malay	Russian	Mean
Chinese	67	X	53	15	X	28	40.8
English		X	67	20	X	59	48.7
Hindi	X	X	X	X	X	X	X
Spanish	77	X		7	X	32	38.7
Arabic	69	X	45		X	55	56.3
Malay	X	X	X	X	X	X	X
Russian	63	X	39	14	X		38.7
Mean	69	X	51	14	X	43.5	44.5

Table 9: Babylon BLEU scores (X = not supported)

Source/ Target	English	Hindi	Spanish	Arabic	Malay	Russian	Mean
Chinese	32	2	24	18	X	54	26
English		2	77	12	X	81	43
Hindi	3		3	4	X	31	10.3
Spanish	72	3		2	X	51	32
Arabic	54	1	26		X	65	36.5
Malay	X	X	X	X	X	X	X
Russian	83	2	78	16	X		44.8
Mean	48.8	2	41.6	10.4	X	56.4	40.5

Table 10: Yandex BLEU scores

Source/ Target	English	Hindi	Spanish	Arabic	Malay	Russian	Mean
Chinese	90	39	77	51	60	65	63.7
English		40	85	51	63	65	60.8
Hindi	52		47	14	10	11	26.8
Spanish	85	43		43	60	69	60
Arabic	82	41	67		56	73	63.8
Malay	84	42	66	48		68	61.6
Russian	85	38	61	48	62		58.8
Mean	79.7	40.5	67.2	42.5	51.8	58.5	56.6

5 Conclusion

This study is perhaps the first to compare eight popular, free, online translation systems with seven languages in all combinations (except Chinese as the target). Results showed that, similar to other studies, Google Translate was more accurate overall as compared to the other seven automatic language translators used for this study. In general, as expected, Google Translate is more accurate when the source language and target language are similar languages or dialects. For example translation from English to Spanish, or, Malaysian Malay to Indonesian will generate better translation than translation from German to Hindi. So while Google Translate had the best results and works well it might not be the most ideal option for some specific language pairs. In addition, as compared to other automatic language translators, Google Translate provides support for far more languages than competitors do.

6 Limitations and Directions for Future Research

As with similar studies this study suffers from some limitations including the use of automatic evaluation rather than human review of translations, a limited sample of text, and a limitation of only six languages. Future research can potentially include other, and perhaps more complex, text samples, different languages and language pairs, and, alternate evaluation and analysis techniques.

References

Aiken, M. (2019). An updated evaluation of Google Translate accuracy. Studies in Linguistics and Literature, 3(3), 253-260. https://doi.org/10.22158/sll.v3n3p253

Aiken, M. and Balan, S. (2011). An analysis of Google Translate accuracy. Translation Journal, 16(2), April, https://translationjournal.net/journal/56google.htm

Aiken, M., Ghosh, K., Wee, J., and Vanjani, M. (2009). An evaluation of online Spanish and German translation accuracy. Communications of the IIMA, 9(4), 67-84.

Aiken, M. and Wong, Z. (2006). Spanish-to-English translation using the Web. Proceedings of the Southwestern Decision Sciences Institute, March 9 – March 13, Oklahoma City, Oklahoma.

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

Brooks, R. (2017). A translation showdown: Man vs machine translation. The Language blog. https://k-international.com/blog/human-translation-vs-machine-translation-contest/

Brownlee, J. (2019). A Gentle Introduction to Calculating the BLEU Score for Text in Python. Retrieved March 17, 2020, from https://machinelearningmastery.com/calculate-bleu-score-for-text-python/

Callison-Burch, C., Osborne, M. and Koehn, P. (2006). Re-evaluating the role of BLEU in machine translation research. 11th Conference of the European Chapter of the Association for Computational Linguistics: EACL, 249-256.

Coughlin, D. (2003). Correlating automated and human assessments of machine translation quality. Proceedings of the Machine Translation Summit IX, New Orleans, USA, September.

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.

G2 Crowd (2019). Best machine translation software. G2 Crowd, https://www.g2.com/categories/machine-translation

Hampshire, S. and Salvia, C. (2010). Translation and the Internet: Evaluating the quality of free online machine translators. Quaderns. Rev. trad. 17, 197-209.

Harris, W. (2012). How Electronic Language Translators Work. Retrieved March 22, 2020, from https://electronics.howstuffworks.com/gadgets/travel/electronic-language-translators1.htm

Himmelein, G. (2019). DeepL: The new gold standard in online translation? Softmaker, https://www.softmaker.com/en/blog/bytes-and-beyond/deepl-the-new-gold-standard

NIST (2005). NIST Multimodal Information Group. NIST 2005 Open Machine Translation (OpenMT) Evaluation LDC2010T14. Philadelphia: Linguistic Data Consortium, 2010.

Oliveira, R. and Anastasiou, D. (2011). Comparison of Systran and Google Translate for English→Portuguese. Traducció i software,

lliure Revista Tradumàtica, Tecnologies de la traducció , 118-136. DOI: 10.5565/rev/tradumatica.14

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

Papineni, K., Roukos, S., Ward, T., and Zhu, W. J. (2002). BLEU: A method for automatic evaluation of machine translation. ACL-2002: 40th Annual meeting of the Association for Computational Linguistics, 311–318., https://www.aclweb.org/anthology/P02-1040.pdf

Papula, N. (2014). MT quality comparison: Google Translate vs. Microsoft Translator. Multilizer. http://translation-blog.multilizer.com/mt-quality-comparison-google-translate-vs-microsoft-translator/

Savoy, J. and Dolamic, L. (2009). How effective is Google's translation service in search? Communications of the ACM, 10(52), 139-143.

Scarton, C. and Specia, L. (2016). A reading comprehension corpus for machine translation evaluation. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia. https://www.aclweb.org/anthology/L16-1579

Shen, E. (2010). Comparison of online machine translation tools. tcworld, http://www.tcworld.info/e-magazine/translation-and-localization/article/comparison-of-online-machine-translation-tools/

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.