Issues in the machine translation of one language to another language - A case study of Google Translator

Tarah Ferrell

Abstract: *Online translations are in extensive use in various areas and disciplines. They have the advantage of reducing the time, effort, and cost. They are also a learning tool for non-native speakers of the language. However, the accuracy and quality of translations is not up to the standard that their academic use could be recommended. This study aims to analyze the issues in the machine translation of one language into another language. The researcher selected Google Translate as the case study. Machine translation systems are broadly classified as rule-based and statistical machine translations. Google Translate uses statistical machine translation. Secondary data, collected from peer-reviewed journal articles, was analyzed for identifying the issues in machine translations. The findings of the study showed four major issues in online translations. These include the issues of linguistic accuracy, differences in the capabilities between languages, inability of handling ambiguous words, and translated plagiarism. Future studies may expand on the work of this study to analyze other online translation software as well.*

Key words: Machine translation, online translation, Google Translate, rule-based systems, statistical machine translation.

1. INTRODUCTION

Machine translation refers to the translation of the text of the source language into target language through the use of computer. The process takes place automatically, and there is no human intervention. Machine translation is also termed as automated translation or instant translation. The systems of automated translation are usually classified into two types. These are rules-based systems and statistical systems. Rules-based systems rely on grammar rules, vocabulary, and dictionary. There are also specialist dictionaries to translate technical terms of a certain discipline. The systems may produce less accurate translations than statistical systems; however the translation text is usually consistent due to the applications of rules. Statistical systems do not rely on language rules. In fact, they are not aware of the linguistic rules. They build their 'expertise' by learning from the data set. They can translate the texts through the analysis of large amounts of data of language pairs. They produce more accurate translations than rules-based systems. However, the translations are less consistent due to the continuous process of learning from the data sets [Callison-Burch et al., 2011]. Another way, of looking at the machine translation, is to differentiate between the word for word translations and phrase translations. The advantage of phrase-based translations is that it is possible to analyze the context and semantics in phrase pairs. Bilingual word embeddings have proved very useful in this respect [Zou et al., 2013]. On the contrary, word for word translation does not take into account the context and semantics of the language.

2. PROBLEM STATEMENT

Machine translation has been in the process of development since 1940. Since then, many approaches and techniques have been used. With its growing use, there is also confusion and misinformation about its capabilities, purpose, and use. This paper aims to highlight issues in the machine translation of one language to another language. The researcher employs a case study method for the research. Google Translate is one of the most widely used online translation services of the world. Due to its growing use, the researcher selected this online translator for analyzing the issues in online translations. The findings of the study will not only highlight the issues of Google Translate, but also provide an overall picture of the issues in machine translation from theoretical, linguistic, and technical perspectives.

3. GOOGLE TRANSLATE

Google Translate is a product of Google that provides multilingual service. It is used for the translation of the text from one language to another language. The service was based on a software engine SYSTRAN before October 2007. Since October 2007, Google Translate is using in-house, proprietary technology that uses the concepts of statistical machine translation. The service offers a web interface [Sfetcu, 2014]. The online translation is gaining increased importance due to the desire of the people to complete the tasks in the shortest possible time. Online translations save the people from paying a professional translator and leafing through a dictionary. Despite the popularity of Google Translate, scholars argue if the technology can be an alternative to the services of a trained professional. One of the issues with Google Translate is that it never confesses to not having an answer. Also, the translation is not context sensitive. The issue of context becomes particularly evident for full sentences. There are syntactic differences between languages. Online translations often attempt to translate content word for word. The effect can have considerable implications for an entire document. If a file contains specialized or technical subject matter, it may be hard to read if translated from one language to another using Google Translate. It is particularly noticed if no post-translation improvements are made. The writing style of online translations is also inappropriate for formal project submissions [Sheppard, 2011].

A writer, who is an expert in academic English, uses a number of levels in his writings. The writer crafts the sentence with well-formed words. The sentences are linked together into clear, cohesive, and coherent paragraphs. He also aligns the manuscript with the generic expectations of the target audience. The alignment is also made for the stylistic conventions related to specific genres. Technological developments have endeavored to assist academic writers with many of the issues. For example, the wide availability of published papers has enabled the writers to study the structure and style of vast swatches of academic writing. Also, the concordance tools assist in the examination of own interlanguage through different perspectives. The literature of English for Academic Purposes (EAP) has discussed various advancements that can become a part of the toolkit of all academic writers. For example, the spell checker of word-processing programs allows the non-native speakers to write English that has fewer errors comparatively. The autocorrect facility also serves the same purpose. Many word processors also provide grammar checkers. Microsoft Word has now also introduced Format Consistency Checker that is indicated by blue underlining [Groves & Mundt, 2015]. The success of online translations may render all these assistive tools insignificant. However, given the level of success of these translations, it seems a distant reality.

There are also issues of confidentiality in online translations. Google implies the right of using the typed information for improving the service of translation. It is also worth noting that professional translations are expensive, and the costs of translations may be higher than the costs of proofreading. Hence, Google Translate is also used for generating an initial draft that may be given to a proofreader for improvement. It is essential to know the limitations of Google Translate for using it as efficiently as possible. There are several significant limitations of Google Translate. The first is its inability to generate the equivalent text according to context. The second is its tendency to translate the word for word. The lack of sensitivity of online translations to the syntax and idiomatic expressions results in flawed translations. The third major issue is the issue of confidentiality [Sheppard, 2011].

4. LIMITATIONS OF ONLINE TRANSLATIONS

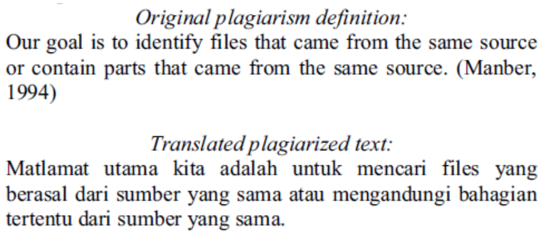
Computer-assisted translations are still unsophisticated in the context of human-computer interface. In statistical machine translation, the translation is first computed and then output is shown to the reader as a fait accompli. The translations made by Google Translate can substantially impede semantic interpretation. A primary cause of the misalignment is when statistical machine translation shows a false equivalence for the original text version and translated text version. It is an indicator that the translator could not assimilate adequate context. Machine intelligence typically proceeds with the formulation of an algorithm. The algorithm emulates aspects of human cognitive and perpetual activity. The algorithms process the same digital texts and come up with same or better results. The primary objective is to achieve objective algorithmic optimization. It is achieved without explicit consideration to semantic context. It is hoped that the context will emerge implicitly in the form of correlations that are inherent in the algorithm computation. It may become counterproductive particularly in the conditions when there is a need for human intervention later. The task of finding and correcting the mistaken results may exceed the time of translating the text manually [Chessa & Brelstaff, 2011].

In addition to the assistance tools for the writers, online translation technology is one that could replace and overtake these features. This technology is commonly known as web-based machine translation (MT). Google Translate is an example of a web-based machine translation. Google Translate is regarded as a statistics-based translation tool. The statistics-based tools calculate probabilities of different translations of a phrase. The probability is calculated for a phrase being correct. The translation with the highest probability is displayed to the viewer. It is unlike traditional method of translation that provides word for word translation. Google Translate also provides interactivity to the user. The user has the facility of correcting the original translation. The updated translation is absorbed into the database. The history of machine translation dates back to 1940s, when punch card systems were being used. Since then, the methodology has experienced significant advances and several setbacks. Despite the use of developments in artificial intelligence, the translations are still far from perfect. However, online translations still have widespread uses. These include the use of machine translation by non-native speakers of the language and screening news reports by intelligence agencies of the governments [Groves & Mundt, 2015].

5. TRANSLATION PLAGIARISM

Google Translate also raises the issue of translation plagiarism. It is a translation of a sentence in the source language to a target language. Suppose a sentence is available in Google in Russian language. If the sentence is translated from Russian language to English language, it is hard to be detected by the plagiarism software, as the translated version is not published. The translation is a complex fuzzy process, and plagiarism software is not trained to translate the work of other languages to check for plagiarism in the academic work. Also, modern online translations use statistical machine translation. In these translations, system learns from the training set and produces the translations based on the highest probability. Hence, two online translation software may not produce the exact output due to different training sets. Also, one translator itself may produce different outputs at various points in time. The inconsistency in translations is due to the performance improvements in statistical methods [Kent & Salim, 2010].

Most of the plagiarism detection software use fingerprint-based approach for plagiarism detection. The basic idea behind this approach is to divide the whole suspected document into small parts. The parts are word-based, statement-based, and line-based. A comparison of the parts is made with the source documents for the detection of similar part. However, the approach is not much strong as even slight modification in text influences the fingerprint of the document. According to the definition of plagiarism, a sentence is plagiarized even if its structure is changed, but the idea or thought remains same, and there are no citations or credits to the original author. Academic plagiarism suffers from the techniques of paraphrasing. The technique either replaces the original words with synonyms or modifies the structure of the sentence. Plagiarism tools are not robust against synonyms. Translated plagiarism is also a growing problem besides common plagiarism. The tools are unable to detect translated plagiarism. For example, if a sentence is translated from English Language into Malaysian Language, no tool will be able to detect plagiarism in the translated version. Kent & Salim [2010] provides an example, in which the English text has been translated into Bahasa Melayu version. The translated text is not detectable by the plagiarism tools. It is because the fingerprinting approach fails due to the differences of the fingerprint between the translated version and the original version [Kent & Salim, 2010].



Original plagiarism definition:

Our goal is to identify files that came from the same source or contain parts that came from the same source. (Manber, 1994)

Translated plagiarized text:

Matlamat utama kita adalah untuk mencari files yang berasal dari sumber yang sama atau mengandungi bahagian tertentu dari sumber yang sama.

Figure 1: An example of Translated Plagiarism [Kent & Salim, 2010].

6. DATA ANALYSIS

For EAP community and writers, it is significant to ascertain if Google Translate can accurately render a source text into English language. The academic writers cannot use the tool if it is not guaranteed to produce accurate and quality translation. Also, it needs to be seen that how much work needs to be done in case of proofreading. If the time in proofreading exceeds the time of manual translation, then the online translations lose their significance in academic writing. Also, as will be illustrated in the examples that follow, there are certain errors that are made consistently by Google Translate. Using the tool may reveal the reader of the manuscript that the writer has used Google Translate instead of writing on his own. It may create a negative impression on the reader as online translations have not been accepted as yet as the authentic sources of translation.

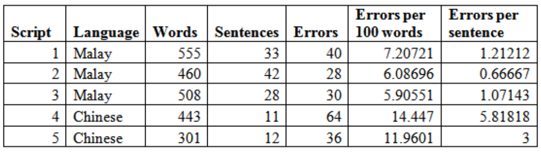
There are two main approaches used by the scholars in this respect. One approach has been proposed by Colina, in which the issue is examined from the point of view of translation quality. Another approach is to evaluate the translation competence of Google Translate. Groves & Mundt [2015] studied the linguistic accuracy of the translation of Google Translate. They used taxonomy of error types. The details of error analysis enabled the assessment of the linguistic accuracy of the translation. The study produced the following error list for the sample translations:

Code	**Title**	**Examples from translated texts**
WF	Word form	A student test detects only the ability to say yes or **memory**
ART	Article	same way to learn **the** memorizing
VT	Verb tense	one that **measured** the level of ability in several ways
VF	Verb phrase	the individual who failed the exam **ignored** or looked down upon by society
PL	Plural	**Examination**, especially in Malaysia plays an important role
AGR	Agreement	Activitie**s** such as off-site **is** very dominant
PREP	Preposition	Abuse and misunderstanding among students **on** examinations should be eliminated
WC	Word choice	Examination is considered something very **high**
COM	Comma	Learning aspects such as, music and art, **can not** be measured
SP	Spelling	How can the ideological principles **Specifically** implement them?
WO	Word order	students will focus on **such topics only**
WW	Wrong word	**Support** parents and teachers are required so that they can be overcome

Code	Title	Examples from translated texts
AP	Apostrophe	Third, teachers and students too expect **students** exam results
SS	Sentence structure	This result is that parents do not ignore and **less affection on them**
MW	Missing word	First, the examination has been highly beneficial to students but **also students** to study a topic that will be tested only on the exam
REF	Pronoun reference unclear	and students will focus on **such** topics
PRO	Pronoun incorrect	Teachers will also place high expectations on **him**
RO	Run on	I believe that in order to test the ability of the method to detect the candidates more harm than good, in other words, the examination system is not a good way to test students' abilities
FRAG	Fragment	In addition, people who have a bias to the students who got poor marks from students who get higher scores.
UNCLEAR	Unclear	College entrance examination system for screening system, especially in the eyes of their talents

Table 1: Error List of Sample Translations [Groves & Mundt, 2015]

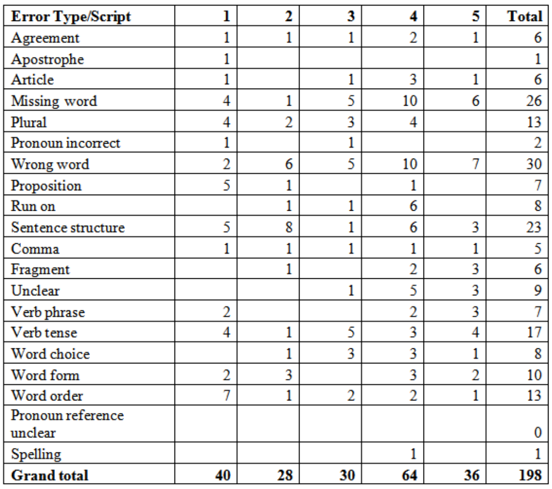
According to another finding of the study, Google Translate showed more accuracy for European languages than Asian languages. There were variances found in the translation quality between languages. Table below shows the errors per script for the languages of Malay and Chinese. Google Translate was used to translate Malay and Chinese languages into English language.



Script	Language	Words	Sentences	Errors	Errors per 100 words	Errors per sentence
1	Malay	555	33	40	7.20721	1.21212
2	Malay	460	42	28	6.08696	0.66667
3	Malay	508	28	30	5.90551	1.07143
4	Chinese	443	11	64	14.447	5.81818
5	Chinese	301	12	36	11.9601	3

Table 2: Errors per script [Groves & Mundt, 2015]

The results show that translation was more accurate for Malay than Chinese. There can be two main reasons for the difference in accuracy. The first is the relative strength of English in Malaysia. Also, there are numerous online documents in English language as well as Malay language. It provides a larger dataset to Google engine for the task of translation. Since the Google engine produces statistical machine translation, hence learning from the training dataset influences the quality and accuracy of the translations. Groves & Mundt [2015] also produced the classifications of errors for the sample five scripts studied as shown in the table below:



Error Type/Script	1	2	3	4	5	Total
Agreement	1	1	1	2	1	6
Apostrophe	1					1
Article	1		1	3	1	6
Missing word	4	1	5	10	6	26
Plural	4	2	3	4		13
Pronoun incorrect	1		1			2
Wrong word	2	6	5	10	7	30
Proposition	5	1		1		7
Run on		1	1	6		8
Sentence structure	5	8	1	6	3	23
Comma	1	1	1	1	1	5
Fragment		1		2	3	6
Unclear			1	5	3	9
Verb phrase	2			2	3	7
Verb tense	4	1	5	3	4	17
Word choice		1	3	3	1	8
Word form	2	3		3	2	10
Word order	7	1	2	2	1	13
Pronoun reference unclear						0
Spelling				1		1
Grand total	**40**	**28**	**30**	**64**	**36**	**198**

Table 3: Error Classification [Groves & Mundt, 2015]

From the above table, it is evident that the most common errors are related to word choice. The next significant errors were related to sentence structure and missing words. It confirms that Google Translate produces errors with the subtle differences of meaning between words in the source language and the words in the target language. It also highlights that the parser fails to parse certain structures effectively and may turn to word for word translation. It eventually results in the lack of clarity in the output text.

Another finding of the study was that the Google Translate produced certain translated sentences with complete grammatical accuracy. Not only were the sentences accurate syntactically, but also the translation has a convincing academic style. It shows that the translate engine can produce good translations for long sentences. However, it is the consistency that is an issue in the case of Google Translate.

7. CONCLUSION

Machine translations are classified as rule-based and statistical machine translations. Rules-based systems focus on grammar rules, vocabulary, and dictionary. Statistical methods apply the statistical techniques to find the probabilities of all possible translation pairs and produce the output that has the highest probability of being correct. Google Translate is a type of online translator that uses statistical machine translation. The community of English for Academic Purposes (EAP) has always been interested in finding technology tools that could assist, improve, and expedite the process of academic writing. The word processing tools such as the spell checker, auto-correct, and format consistency checker have proved useful in producing accurate and good quality writing. Online translation tools have the potential to supersede all these features as it is better to translate the text in its entirety in one go rather than translating it and then correcting errors in the word processor. However, it is easier said than done. Machine translation of one language into another language has its set of issues and problems. Despite using the advanced concepts of artificial intelligence, the machine translation has yet to achieve the level of sophistication. It is because of the involvement of various stakeholders, complexities of the languages, and the differences in the treatment of words in the source and the target language. Specifically, in the case of Google Translate, this study found several issues in translation.

The first issue is the weak linguistic accuracy of the translation. The translation engine produces the output having the highest probability. However, it is still based on heuristics, and the best possible result may be far from being the accurate translation of the word. The translation produces errors in verb tense, word form, article, plural, agreement, word order, word choice, spelling, punctuation, run on, and fragment. Also, it appears when the statistical approach fails to produce the translation; the engine adopts the word for word translation technique. The second issue is the difference in the capabilities of translator between languages. Since the translator is based on statistical machine translation, therefore, larger the training set, the better the translation will be. The third issue is the inability of handling ambiguous words. The translator does not produce quality output if the word of the source language does not exist in the target language, or there are multiple meanings of the word. The findings of the past studies have shown the most common errors in Google Translate are related to word choice, sentence structure, and missing words. Google Translate has also created the issues of translated plagiarism. Most of the plagiarism tools are based on fingerprint approach. When the text of another language is translated into English or any other language, the fingerprints are lost. The plagiarism software in such cases fails to detect the plagiarism.

In summary, online translations have widespread use in many areas of applications. They can prove useful for non-native speakers of the language. They are also helpful in reading texts of other languages for the purpose of intelligence and investigation. However, they are not suited for academic writing and professional submissions. It is because the quality of translation is not of the standard that can be trusted without any proofreading. Also, the errors produced in sentence structure, semantics, and contexts are so numerous that the proofreading is a tedious task. In the cost-benefit analysis, the cost of a professional translation may be higher than the cost of proofreading. However, in other circumstances, there is a long way to go when online translations will reach a point that their academic use would be beneficial.

8. REFERENCES

[1] Callison-Burch, C., Koehn, P., Monz, C., & Zaidan, O. F. Findings of the 2011 workshop on statistical machine translation. In Proceedings of the Sixth Workshop on Statistical Machine Translation. Association for Computational Linguistics, 22-64. 2011.

[2] Chessa, F., & Brelstaff, G. Going beyond Google Translate?. In Proceedings of the 9th ACM SIGCHI Italian Chapter International Conference on Computer-Human Interaction: Facing Complexity, 108-113. 2011. doi:10.1145/2037296.2037324

[3] Groves, M., & Mundt, K. Friend or foe? Google Translate in language for academic purposes. English for Specific Purposes, 37, 112-121. 2015. doi:10.1016/j.esp.2014.09.001

[4] Kent, C. K., & Salim, N. Web based cross language plagiarism detection. In Computational Intelligence, Modelling and Simulation (CIMSiM), Second International Conference, 199-204. 2010. doi:10.1109/CIMSiM.2010.10

[5] Sfetcu, N. Google Products, Services and Tools. Nicolae Sfetcu. 2014.

[6] Sheppard, F. Medical writing in English: the problem with Google Translate. La Presse Medicale, 40: 6, 565-566. 2011.

[7] Zou, W. Y., Socher, R., Cer, D., & Manning, C. D. Bilingual Word Embeddings for Phrase-Based Machine Translation. EMNLP. 2013.