An Evaluation on Online Translation Services with Japanese to English Translation *

Hei Man, Haley FONG

January 29, 2024

1 Introduction

1.1 Machine Translation (MT): A Brief Introduction

Machine translation (MT) refers to the use of computerised systems to produce translations from a natural language to another (Hutchins & Somers, 1992). The process can either be or be not interfered by humans. Depending on the extent of human interference, types of MT can be classified across a spectrum – from Human Translation, Computer-Aided (Human) Translation (CAT), Human-Aided Machine Translation (HAMT), and finally the ideal of Fully-Automated High-Quality Translation (FAHQT).

The history of MT could be traced back to the 1960s (Yang, Wang, & Chu, 2020). MT started off as a rule-based system, in which lexical items of a source language (SL) is mapped to its corresponding word in a target language (TL). Rules of how sentences are formed in both the SL and the TL are then listed respectively. Utilizing the mappings of the concerned languages, rule-based machine translation (RBMT) could be performed with first analysing the SL input then generating the TL output (Okpor, 2014). However, RBMT falled short in capturing contextual information, for example, idioms and ambiguous sentence structures. Such hurdle was overcame by corpus-based systems including statistical MT (SMT) and example-based MT (EMBT). SMT had been the SOTA approach since the 1980s (Brown et al., 1990), and has evolved robustly in the following few decades. The rationale behind these systems was to make statistical predictions on equivalent TL candidates, using bilingual SL and TL data with aligned examples of words, phrases or even sentences. As MT technology continued to grow, hybrid systems of RBMT and SMT came to light, maximiz-

*This essay is an extension of the coursework for 7LN0003 Machine Translation, under the Master of Arts in Computational Linguistics programme in the University of Wolverhampton. It was originally submitted in May 2020. Full results could be accessed here: `https://github.com/hmsquare/7LN0003-MachineTranslation`

ing the strength of respective types of MT. With recent breakthroughs in neural networks and large language models, MT technology enters a new era of advancement with neural MT (NMT) (Maučec & Donaj, 2019).

Not only did MT technologies grow rapidly throughout the years, so did its accessibility to the masses. From SYSTRAN to Google Translate, MT services on the Internet has undoubtedly gained a booming increase of users (Okpor, 2014). This essay thus looks into popular online translation services, evaluating how they perform in translating daily texts from Japanese to English. The investigation aimed to answer the following two questions: whether online translation services are good enough for day-to-day use by laymen, and whether the systems are reliable for translating languages of specific purposes (LSPs). The study was originally carried out in 2020 with MT services by Google and Weblio respectively, yielding intriguing observations where each system has its own pros and cons. Taking the exponential refinements with NMT into consideration, an extended investigation was carried out with the systems to see how has they and the previous conclusions aged through the past few years.

1.2 MT Evaluation

Along with the development of various types of MT systems, numerous approaches have been put forward to measure their performances. The performance of an MT system is important to a few stakeholders in MT – the developers, the translators and the audience of the translation (Hutchins and Somers, 1992). For instance, with MT evaluation, developers can optimize translation quality and efficiency of systems, reducing the workload of human translators (in either pre- or post-editing), and ultimately providing natural, high quality translations to message recipients.

MT evaluation can be performed either by humans or computer programmes. According to Doherty (2019), MT products could be evaluated in two main ways, the first through linguistic measures and the second through user experience. As human perception in subjective, a number of metrics are utlized to assess the quality of an MT output. Fluency, readability and accuracy are classic examples of linguistic measures. For user experience, data like eye activity when reading the MT output, as well as webpage brwosing activities (number of visits, scrolling frequencies etc.) could be collected and exploited for MT evaluation.

Alternatively, MT evaluation by machines involves the comparison of the MT output against a reference translation. Being named as 'string-based' approaches, smaller units of the input and output texts, such as characters, words, phrases or sentences would be compared to generate a score. Common scoring methods include BLEU (Bilingual Evaluation Understudy), METEOR (Metric for Evaluation of Translation with Explicit Word Ordering) and the Levenshtein (Edit) Distance.

1.3 Challenges in Japanese to English Translation

One major challenge in Japanese-English translation is to preserve sociolinguistic information (Furuhata-Turner, 2013). In Japanese, sociolinguistic information is encoded into either different terms having the same meaning, or in the suffixes of verbs phrases. On the other hand, orthography could give rise to obstacles in meaning disambiguation. In the Japanese writing system, *kanas* and *kanjis* are used. The former are Japanese-invented characters which are syllabic (Sampson, 1985). The latter are Chinese characters that are either imported from China or invented locally in Japan. The same combination of *kanas* can result in different *kanjis* and at the same time one kanji can have multiple pronunciations, meaning that it can be broken down into different combinations of *kanas*. Last but not least, ellipsis is common in Japanese discourse (Okamoto, 1985). Noun phrases and verb phrases are often left out as a result of the distinctive communication style of a high-context culture. Information like the referents of pronouns and the performers or recipients of actions are omitted, and it is expected that listeners or readers can pick up such information from the surrounding context. Table 1 gives a summary of these difficulties for the sentence *I am Hanako*.

Formality	Casual	*Watashi-wa / Hanako-**da***
	Formal	*Watashi-wa / Hanako-**desu***
	Honorific	***Watakushi**-wa / Hanako-**degozaimasu***
	Dialect	*Watashi-wa / Hanako-**ja***
Orthography	*Kana* (Possible *kanjis*)	わたし-は/ はなこ/ です (わたし→私 or 渡し / はなこ→華子 or 花湖)
	Kanji (Possible *kanas*)	私- は/ 花子/ です (私→ わたし or わたくしor し / 花子→ はなこ or かし)
Ellipsis	Omitting pronoun	*(Watashi-wa /) Hanako – desu*

Table 1: Examples of how *I am Hanako* could be expressed in Japanese

2 Methodology

2.1 Experimental Procedures

2.1.1 Text Collection and Hypothesis

Two pieces of texts were taken for the study – a children's fiction and a news article. The former is expected to be a piece of expressive text of simpler language, while it is likely for the latter to be written in more complex, adult-like language of the goal to be informative. Hence, apart from looking at how well the MT systems could identify and transform the information in the texts, whether the results retain the texts' expressive and informative natures would be studied. In other words, how emotional nuances are captured, as well as how concepts or events are being acknowledged would be focused when evaluating the translation results.

For the piece of children's fiction, it was retrieved from an online collection of bilingual stories[1]. The story retrieved was written in full Japanese characters, and each line of the story was then aligned with an English translation.

The news article is was retrieved from BBC. The article was originally written in English[2] and a translated version is available on the Japanese version of the website[3]. For a fair comparison, the texts are around 1500-1700 characters long. Hyperlinks to the texts are supplied at the bottom of the page.

2.1.2 Text Preprocessing and Alignment

A third piece of source text was created by re-writing the story from Japanese characters into Chinese characters where appropriate. This was to test whether the use of Chinese characters facilitate meaning disambiguation during the translation process. The resulting document consisted of around 1300 characters.

Moving onto sentence alignment, each line of the story was aligned with the respective translation into an excel file for comparison. For the news article, sentences that only exist in either the English or the Japanese version were removed so that a parallel comparison

[1]Yukionna (The Snow Woman), from Hukumusume Fairytale Collection `http://hukumusume.com/douwa/English/jap/01/22_j&E.html`

[2]Coronavirus: World Health Organization members agree response probe `https://www.bbc.com/news/world-52726017`

[3]WHO、新型ウィルス対応を検証へ　全加盟国が同意`https://www.bbc.com/japanese/52733876`

could be made. Only the sentences that were remained would be passed into the MT systems in the next stage.

2.1.3 Translation with Online MT Services

2.1.3.1 Weblio

Weblio is a Japanese company that has been providing online dictionary services since 2005. Their online translation services has been available since 2012, offering Japanese to/from Chinese, Korean and English translation. Weblio provides its users with support on studying foreign languages (mainly English) for overseas employment and travel (Weblio, 2020). Specifically, there is a sample-sentence-lookup function on their website, in which users can input a Japanese sentence and the system returns English translations under different contexts for users to reference. Examples are from their self-developed email corpus and the Wikipedia Japanese-English parallel corpus. The approach of MT is not specified on Weblio's website.

2.1.3.2 Google Translate

Google began providing online translation service since 2006 (Turovsky, 2016). It started off by using statistical MT, with training data from corpora of texts from the United Nations and the European Union. It also invites its users, the Translation Community to make contributions to improve their MT system. For instance, users are invited to provide their own translations to phrases provided by the system. Or, they can help rate possible translation so that the system learns which ones are better. In recent years, Google has upgraded their MT service by using deep learning (Turovsky, 2017).

2.1.3.3 DeepL

DeepL entered the online translation service scene in 2017 as one of the neural MT providers (Fitria, 2023). Despite being available in less languages, it is more accurate than Google Translate. In Takakusagi et al. (2021)'s study, it was found that DeepL could even translate text of a specialized field. According to DeepL's official webpage [4], it is their unique NMT algorithm, which is sensitive to "subtle meanings" and "slightest nuances", that sets them apart from other MT services. Although DeepL's development processes are not disclosed, it is known from the company's profile [5] that they began constructing their system using Linguee's bilingual translation data.

[4]Why DeepL?: Translation quality `https://www.deepl.com/en/whydeepl`
[5]DeepL's company profile: `https://www.deepl.com/files/press/companyProfile_EN.pdf`

2.1.4 Evaluation of Translation Output

The translation outputs from the online systems were evaluated both automatically and manually.

BLEU was used as the automatic evaluation method. BLEU stands for "BiLingual Evaluation Study", in which a numerical metric is utilized to measure the closeness between a MT output with a gold, corpus standard (Papineni, Roukos, Ward, & Zhu, 2002). The BLEU scores for each sentence in the output was calculated. The score of a piece of output would then be computed by averaging the sum of the scores of each sentence. The more the BLEU score is closer to 1, the higher the quality is the MT output. Copying the MT outputs into a .txt file, BLEU score for the outputs were generated using the Natural Language Toolkit in Python. Here are the source codes of the evaluation program (Brownlee, 2019).

As BLEU only shows the similarity of the MT output with a reference, it could only tell how accurate a result is with respect to the reference. Whether the translation is readable or appropriate in the target culture is yet to be judged (Hutchins & Somers, 1992). This is why manual evaluation is necessary, and hence the results would be evaluated by a human, who would comment on the outputs' accuracy and fluency.

2.2 Experimental Setup

Three trials of experiments were carried out to access the performance of the MT systems. The first using the piece of children's fiction written in full Japanese characters and the second with the one written with *kanjis*. These trials assessed the importance of meaning disambiguation using *kanjis*, as well as the ability of the systems to translate everyday expressions such as onomatopoeic words and dialectal expressions, which do not have direct English translations. Last but not least, the final trial was performed by back-translating the news article and see how well the systems performed with texts of informal, figurative and formal, informative natures.

3 Results

Table 2 shows the BLEU scores for the experiments. The BLEU scores for the 2020 results were re-computed in 2023 for a fair comparison across versions of the systems. It is speculated that the difference between the 2020 scores and 2023 scores is due to Python NLTK's version difference. In 2023, the scores were calculated using weights of `weights=(1.0, 0, 0, 0)` for a closer comparison. This section comments on the systems would be made along with examples extracted from the translation results.

	Fiction (*kana)*	Fiction (*kanji*)	News Article
Weblio 2020	0.407 (0.547)	0.470 (0.613)	0.589 (**0.478**)
Google 2020	**0.465 (0.587)**	**0.526 (0.616)**	**0.632** (0.434)
Weblio 2023	0.387	0.437	0.591
Google 2023	0.496	0.519	**0.654**
DeepL 2023	**0.523**	**0.566**	0.609

Table 2: BLEU scores for all experiments with the online translation services in 2020 and 2023 (alongside with BLEU scores computed in 2020)

In general, Weblio's output in 2020 is largely similar to that of 2023. Google's results have displayed slight improvement over the years, and the addition of DeepL proved that systems making use of deep learning techniques aced the task. Comparing scores of these systems, it is suggested that DeepL could be better at processing figurative language while informative texts favour Google.

3.1 Experiment 1: Children's Fiction in full Japanese characters (*kanas*)

Google greatly outperformed Weblio in this experiment. In 2020, Weblio's performance was in fact very poor, in which only 19 out of 52 lines of the story was fully translated into English. A large amount of instances remained in Japanese in the output text. Yet despite of Google's effort to translate the whole text into English, a number of instances were incorrectly translated due to ambiguity with the Japanese characters as illustrated in Table 3. Therefore, even if Google's results were readable, they did not make much sense. At the same time, from Table 4, it was noticed that the sentiment of characters' speech was incorrectly translated. However, among Weblio's few attempts to fully translate sentences into English, it did correctly capture such sentiment. Moving forward to 2023, Google successfully fixed its mistake, but the absence of *kanjis* posed a major obstacle in accuracy. Only DeepL was able to overcome the issue flagged in Table 3

Source	ある　おおあめ　の　ひ、おのきち　の　いえ　の　まえ　に ひとり　の　おんなのひと　が　たっていました。
Weblio 20, 23	あるおおあめのひ, おのきちのいえのまえにひとりのおんなのひとがたっていました.
Google 2020	There was one woman standing in front of a certain *candy* tree and in front of one another.
Google 2023	There was a woman standing in front of one of the great *candy* houses.
DeepL 2023	One *rainy* day, a woman was standing in front of Onokichi's house.
Standard	On a very *rainy* day, there was a young woman standing in front of Onokichi's house.

Table 3: Example of an incorrect translation due to *kana* ambiguity

Source	「う～、さむい！」
Weblio 20, 23	Ugh, ... is cold!
Google 2020	"Wow, it's cold!"
Google 2023	"Ugh, it's cold!"
DeepL 2023	It's so cold!
Standard	"Brrr.... It's cold!"

Table 4: Example demonstrating the systems' abilities in capturing sentiment

3.2 Experiment 2: Children's Fiction rewritten with Chinese characters (*kanjis*)

Comparing the example in Tables 3 and 5, it is obvious that *kanjis* played a crucial role in meaning disambiguation, as noted with *candy* and *rain* with あめ. Table 6 demonstrated how DeepL benefited from *kanjis* in the same manner, and that they are important for the translation systems to render as much information in the source text as possible. It is interesting to see Google and DeepL took turns to be challenged with *kana* ambiguity, and that DeepL actually missed 大雨 in its translation despite *kanjis* were supplied. For Weblio, although it had no problem in correctly translating nouns and using preferred candidates, the system still fell short in handling dialectic and onomatopoeic expressions, whereas Google and DeepL managed to ignore the former and accurately translated the latter.

Source	ある大雨の日、おのきちの家の前に一人の女の人が立っていました。
Weblio 20, 23	On the day of a certain heavy *rain*, one woman stood before a person of おのきちの.
Google 2020	One *rainy* day, a woman was standing in front of Onochichi's house.
Google 2023	One day, on a *rainy* day, a woman was standing in front of Onokichi's house.
DeepL 2023	One day, a woman was standing in front of Onokichi's house.
Standard	On a very *rainy* day, there was a young woman standing in front of Onokichi's house.

Table 5: Example of how *kanjis* improved translation quality

Source (*kana*)	「こんやは　ここで　とまる　より、しかたあるめえ」
Weblio 20, 23	こんやはここでとまるより, しかたあるめえ
Google 2020	"Koya is a way to stay, rather than *stop* here."
Google 2023	Well, there's no better way for him than to *stay* here.
DeepL 2023	There's nothing else to do but to *stop* here.
Source (*kanji*)	「今夜はここで泊まるより、仕方あるめえ」
Weblio 20, 23	Than "I *stay tonight* here way あるめえ"
Google 2020	There is no way to *stay* here *tonight*.
Google 2023	"I have no choice but to *stay* here *tonight*."
DeepL 2023	We have no choice but to *stay* here *tonight*.
Standard	"We have no choice but to *spend the night here*."

Table 6: Example where dialectal expressions remained untranslated by Weblio

3.3 Experiment 3: Backtranslation of a piece of News Article

In 2020, it was unexpected that Weblio outperformed Google from the BLEU scores, especially when it once again gave up translating when it came across a relatively new concept. For instance, Tables 7 and 8 show names which were translated in Google's output but not in Weblio's. This suggested that Google's neural MT system is capable of more frequent updates with help from its Translate Community, and that Weblio's fails to catch up with the fast-changing world. Years after the outbreak of the pandemic, with the increased availability on relevant texts online and the ability for the deep learning systems to update their databases regularly (Piergentili, Savoldi, Fucci, Negri, & Bentivogli, n.d.), translation outputs of Google and DeepL further managed to give the proper title of Tedros. Google also corrected its transliteration for *Zhao Lijian*

On the other hand, how the systems deal with gender bias during translation (Savoldi, Gaido, Bentivogli, Negri, & Turchi, 2021) caught attention. Failing to translate the title for Zhao Lijian, Weblio's result remained gender neutral as the Japanese source does. It is intriguing to see Google assigning the title of *spokeswoman* to the official in 2020, since it is biased for such position to be translated with a pinch of masculinity according to Cho, Kim, Kim, and Kim (2019). This is shown in DeepL's output and the original article where *spokesman* was used instead. In Google's 2023 output, the system chose to address the official using the gender neutral title *spokesperson*.

Source	趙立堅報道官は19日、北京で開かれた記者会見で、 アメリカは責任逃れをしようと中国を中傷しようとしていると述べた。
Weblio 20, 23	趙立堅報道官 stated that the United States was going to slander China at a press conference open in Beijing on 19th to do buck-passing.
Google 2020	At a press conference in Beijing yesterday, spokeswoman *Zhao Zheng* said that the United States is trying to slander China in an attempt to escape responsibility.
Google 2023	Spokesperson *Zhao Lijian* said at a press conference in Beijing on the 19th that the United States was trying to smear China in an attempt to avoid responsibility.
DeepL 2023	Spokesman *Zhao Lijian* said at a press conference held in Beijing on March 19 that the U.S. is trying to smear China in an attempt to evade responsibility.
Standard	Foreign ministry spokesman *Zhao Lijian* told a press briefing in Beijing on Tuesday that the US was trying to smear China in order to avoid its own responsibilities.

Table 7: Example demonstrating the systems' approaches in translating names and titles

Source	WHOのテドロス事務局長はすでに、 WHOのパンデミック対応の検証に同意している。 一方で、末端に至るまでの点検が必要だとする意見には反対している。
Weblio 20,23	The テドロス secretary general of WHO has already agreed to inspection of the pandemic correspondence of WHO. On the other hand, I object to an opinion to need check before reaching the end.
Google 2020	WHO Executive Secretary Tedros has already agreed to verify WHO's pandemic response. On the other hand, I disagree with the view that it is necessary to carry out inspections all the way to the end.
Google 2023	WHO Director-General Tedros Adhanom Ghebreyesus has already agreed to review the agency's pandemic response. However, he disagrees with his opinion that end-to-end inspection is necessary.
DeepL 2023	WHO Director-General Tedros has already agreed with the WHO's verification of the pandemic response. On the other hand, he disagrees with the need for end-to-end inspections.
Standard	Dr Tedros had already agreed to a review of the agency's handling of the pandemic, while dismissing suggestions it needed a far-reaching overhaul.

Table 8: Example showing the system's abilities to handle unseen names

Finally, comparing the neural MT systems and Weblio's outputs, the translation of the word "puppet" sparked an interesting observation as circled in purple. Literally translating the Japanese translation of "puppet", it would be "controlled doll" as in the left column. In back-translation, the word became "marionette", which is the French word borrowed into Japanese to refer to "controlled dolls". This prompted the idea that Weblio is more faithful to Japanese, and the claim can be supported by the retention of active voice in the Weblio outputs listed in Table 9 too.

Source	ドナルド・トランプ米大統領はWHOを中国の「操り人形」と呼び、
Standard	President Donald Trump has labelled the organisation a ”puppet” of China
Weblio 20, 23	The U.S. President Donald Trump called WHO a Chinese ”marionette”
Google 2020	US President Donald Trump called WHO a Chinese ”puppet”
Google 2023	US President Donald Trump has called the WHO a ”puppet” of China
DeepL 2023	U.S. President Donald Trump has called WHO a ”puppet” of China
Source	決議案は、欧州連合（EU）が100カ国を代表して提案した。
Standard	The EU presented the resolution on behalf of 100 nations.
Weblio 20,23	European Union (EU) suggested the resolution on behalf of 100 countries.
Google 20,23	The resolution was proposed by the European Union (EU) on behalf of 100 countries.
DeepL 2023	The resolution was proposed by the European Union (EU) on behalf of 100 countries.

Table 9: Examples showing Weblio's faithfulness towards Japanese

4 Discussion

Summarizing the observations presented in the previous section, it is clear that Weblio is less accurate than Google and DeepL, even though results could still be considered satisfactory. The abilities for the neural MT systems to update their translation memories are undoubtedly the key towards higher translation accuracy and fluency. For instance, Google's involvement of the user community and DeepL's database refreshes helped the system to conquer hurdles such as dialectal or onomatopoeic expressions, as well as unseen names and titles.

For Weblio, it is noted that translation outputs are more faithful to the source texts, in which candidates that are more frequent with Japanese English norms were preferred. To be specific, considering Weblio as a platform for Japanese users to learn English for education and work, materials exploited for their translation system would be of a more formal register – which was reflected in the more serious tone in the results generated. Contrasting with Google's system development process, the involvement of a user community helped smooth the translation results with a more natural and casual tone, eventually contributing to the more fluent English outputs. Meanwhile, DeepL's promise in providing human-like translation is well-kept. Even if MT systems are not designed to translate source texts in full *kanas*, DeepL effortlessly delivered better results when translating the children's story.

Besides, the NMT systems demonstrated a higher level of coherence as seen in the following example, which is the following line of the one in Table 5.

Source	「雨で、困っておいでじゃろう」
Weblio 20, 23	Because of rain 困っておいでじゃろう
Google 2020	*I'm* in trouble because of the rain."
Google 2023	"*You* must be in trouble because of the rain."
DeepL 2023	*She said,* "*You* must be in trouble because of the rain.
Standard	"*You* are soaking wet!."

Table 10: Translations of a character's speech after the narration in Table 5

As mentioned in Section 1.3, it is a common practice to omit subjects in Japanese discourse. This is a classic example where MT systems are challenged with the task to figure out what to put in place of the missing subject. With the source texts being fed into the systems as a whole [6], only the NMT systems are able to tell that the subject of the line should be a person. It seems that Google was still on its way to solve this problem in 2020, as it placed *I* into the subject position, which was a rather common solution when the missing subject problem is encountered by MT systems. In 2023, it eventually learned from aforementioned context that the attention of the narrator is currently on *Onokichi*, who is probably sheltered from the rain by being inside of the house. Hence, from the point of view of Onokichi, the line should be about the other character who got wet. As a result, *you*, rather than *I*, was picked as the subject of the line. For DeepL, an attempt to recover the speaker of the line was made as well, which unfortunately ended up to be an extra, problematic move.

[6]Except for DeepL where a word limit is posed. The texts were fed into the systems with parts divided at sentence boundaries.

On the other hand, the consistency of word choice used throughout the entire text is particularly important in storytelling. When handling characters' names, Weblio successfully rendered as *snow fairy* with its inclination towards Japanese-influenced English in the second experiment. Yet, it failed to remain consistent when it comes to translating お雪 (*O-yuki*), using a combination of candidates such as *snowy*, *snow* and *Oyuki*. The translation for おゆき in the first experiment, however, was kept consistent. Google in 2020 could be said to have the worst consistency among all trials. First, it did a poor job in consistently transliterating the names of *Shi-ge-sa-ku* and *O-no-ki-chi*, where outputs like *shikisaku*, *Shigeru Sakuto*, *Onochichi* and *Onochi* were seen throughout the results in both experiments on the story. Second, in spite of the fact that *Yuki-onna* was a consistent name for ゆきおんな in the first experiment, 雪女 was seen to be *Yuki-onna*, *snow woman* and even *Yukiko* in the second experiment. Thirdly, the system failed to identify お雪 as a named entity, outputting inconsistent translations with *snow* and *Oyuki*. This observation is similar to that of Weblio's where the use of *kanas* helped improve consistency instead. These issues were very much resolved in Google Translate 2023, with characters' names correctly transliterated and 雪女 being consistently *Yuki-onna*. DeepL topped the shortcoming Google's inability to choose a consistent translation for お雪 in 2023. Still, it is found that the system could face difficulties in transliterating names when there are no *kanjis* too. Nonetheless, considering the fact that NMT systems are not trained to work without *kanjis*, DeepL won the race of translating with near-human consistency.

5 Conclusion

All in all, Weblio Translate, Google Translate and DeepL are easy-to-use online MT systems which caters to the needs of commoners. In particular, the latter 2 NMT systems would be suggested for more accurate and more fluent translations with a stronger grasp on up-to-date concepts. Albeit Weblio being a useful tool for Japanese speakers to learn English at school or at work, for more native-like results, Google or DeepL should be preferred. Even if experimental results show that DeepL is fitter than Google when dealing with children fiction and vice versa for informative texts, DeepL could still be preferred for professional or field-specific purposes as it offers more post editing functions.

Regardless of the seemingly promising performances of the NMT systems, a number of improvements are to be made in order for current MT systems to reach human-like performance. From Google's failure to use consistent character names to the error from DeepL on subject insertion, it is hinted that one of the next steps to move NMT forward is related to context. For instance, current NMT systems are thought to work by first passing an encoded sentence of the source language into a neural network, then yielding the resulting sentence in the target language through decoding the calculations made by the neural network (Stahlberg, 2020). Since the computation is carried out at the sentence level, connections between sentences are often overlooked. Fernandes, Yin, Neubig, and Martins (2021) introduced an approach to take multiple sentences from both the source and target languages when training an MT system. In the particular context of this project, zero anaphora resolution would be a key area to look explore. On top of Rusli and Shishido (2022)'s suggestion in exploiting previous context, it is reminded by Wong, Maruf, and Haffari (2020) that future contexts could be helpful towards combating the issue as well. This could be echoed with the example highlighted in Tables 5 and 10, in which the next line in the story depicts the scene of Onokichi letting the woman into the house, proving that the woman is not the speaker of the line in Table 10.

At the end of the day, it should be appreciated how far NMT has made to break everyday language barriers. Meanwhile, such achievements are merely the first steps into better international communications. The existing puzzle of using language with empathy towards different cultures would continue to provide grounds for NMT research.

References

Brown, P. F., Cocke, J., Della Pietra, S. A., Della Pietra, V. J., Jelinek, F., Lafferty, J., ... Roossin, P. S. (1990). A statistical approach to machine translation. *Computational linguistics*, *16*(2), 79–85.

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

Cho, W. I., Kim, J. W., Kim, S. M., & Kim, N. S. (2019). On Measuring Gender Bias in Translation of Gender-neutral Pronouns. In (p. 173-181). Association for Computational Linguistics. doi: 10.18653/v1/W19-3824

Doherty, S. (2019). Translation technology evaluation research. In *The routledge handbook of translation and technology* (pp. 339–353). Routledge.

Fernandes, P., Yin, K., Neubig, G., & Martins, A. F. T. (2021, 5). Measuring and increasing context usage in context-aware machine translation.

Fitria, T. N. (2023). Performance of Google Translate, Microsoft Translator, and DeepL Translator: Error Analysis of Translation Result. *Al-Lisan: Jurnal Bahasa (e-Journal)*, *8*(2), 115–138.

Furuhata-Turner, H. (2013). Use of Comics Manga as a Learning Tool to Teach Translation of Japanese. *The Journal of Language Learning and Teaching*, *3*(2), 72–83.

Hutchins, W. J., & Somers, H. L. (1992). *An Introduction To Machine Translation*.

Maučec, M. S., & Donaj, G. (2019). Machine translation and the evaluation of its quality. *Recent trends in computational intelligence*, *143*.

Okamoto, S. (1985). *Ellipsis In Japanese Discourse*. University of California, Berkeley. Retrieved from `https://www.proquest.com/openview/367338822ab0118375d84faba24bfea1/1?pq-origsite=gscholar&cbl=18750&diss=y`

Okpor, M. D. (2014). Machine translation approaches: issues and challenges. *International Journal of Computer Science Issues (IJCSI)*, *11*(5), 159.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation . In *Proceedings of the 40th annual meeting of the association for computational linguistics* (pp. 311–318).

Piergentili, A., Savoldi, B., Fucci, D., Negri, M., & Bentivogli, L. (n.d.). *Hi guys or hi folks?* benchmarking gender-neutral machine translation with the gente corpus.

Rusli, A., & Shishido, M. (2022). Zero-pronoun annotation support tool for the evaluation

of machine translation on conversational texts. *Journal of Natural Language Processing*, *29*, 493-514. doi: 10.5715/jnlp.29.493

Sampson, G. (1985). Writing systems: Methods of Recording Language. *London, UK: Hutchinson*.

Savoldi, B., Gaido, M., Bentivogli, L., Negri, M., & Turchi, M. (2021, 8). Gender bias in machine translation. *Transactions of the Association for Computational Linguistics*, *9*, 845-874. doi: 10.1162/tacl_a_00401

Stahlberg, F. (2020). *Neural machine translation: A review* (Vol. 69). Retrieved from `https://scholar.google.com/scholar?q=%22neural+machine+`

Takakusagi, Y., Oike, T., Shirai, K., Sato, H., Kano, K., Shima, S., ... others (2021). Validation of the reliability of machine translation for a medical article from Japanese to English using DeepL translator. *Cureus*, *13*(9).

Turovsky, B. (2016). *Ten Years of Google Translate.* Google. Retrieved from `https://blog.google/products/translate/ten-years-of-google-translate/`

Turovsky, B. (2017). *Higher Quality Neural Translations For A Bunch More Languages.* Google. Retrieved from `https://blog.google/products/translate/higher-quality-neural-translations-bunch-more-languages/`

Wong, K., Maruf, S., & Haffari, G. (2020, 4). Contextual neural machine translation improves translation of cataphoric pronouns.

Yang, S., Wang, Y., & Chu, X. (2020). A survey of deep learning techniques for neural machine translation. *arXiv preprint arXiv:2002.07526*.