Student Writings with DeepL: Teacher Evaluations and Implications for Teaching

Brian J. Birdsell

Hirosaki University

Technological changes have the power to disrupt standard educational practices. One recent advancement is neural machine translation (NMT) systems such as Google Translate and DeepL which due to their widespread use have already impacted foreign language education. To explore the effect of NMTs on student essay writing and teachers' evaluation of it, a small-scale study was conducted in which students were divided into two groups, one group used the NMT DeepL and the other did not. English teachers assessed these essays by evaluating them using a standard rubric and then judging whether they believed NMT was used. Results from a Mann-Whitney *U* Test indicate that teachers tend to evaluate essays that used NMT higher than those that did not and they can accurately judge whether NMT was used. Implications of this study are discussed as well as possible ways to effectively use NMT in the writing classroom. As technology continues to improve, foreign language education also has to evolve with these changes.

テクノロジーの変化は、標準的な教育の実践を混乱させる力を持っている。最近の進歩としては、ニューラル機械翻訳（NMT）システムの普及が外国語教育にも影響を与えている。学生と教師双方へのNMTの影響を理解するために、2つのグループの学生を対象に、一方のグループはNMTを使用し、もう一方のグループは使用しないでエッセイを書くという研究を行った。英語教師はこれらのエッセイを標準的なルーブリックで評価し、NMTが使用されているかどうかを判断した。Mann-Whitney U Testの結果から、教師はNMTを使ったエッセイをそうでないエッセイよりも高く評価する傾向があり、NMTを使って書いたかどうかを正確に判断できることが示唆された。本研究の意義は、ライティングの授業でNMTを効果的に使用する方法を議論することにある。テクノロジーが進化し続ける中、外国語教育もその変化に合わせて進化していかなければならない。

Technological change has been accelerating at an exponential rate, reshaping the way we communicate, host academic conferences, and teach. Ray Kurzweil (2004) in his essay *The Law of Accelerating Returns* emphatically stated, "We won't experience 100 years of progress in the 21st century – it will be more like 20,000 years of progress (at today's rate)" (p. 381). Nowhere can this be more apparent than in the development of online neural machine translation systems (NMTs) where over the past 4 to 5 years, they have experienced remarkable growth. One company, DeepL, has been a driving force behind these changes and according to its website (www.deepl.com), many corporations and media outlets have voiced their praise of DeepL's online translator as being more nuanced and accurate than other NMTs. These sudden technological advancements like DeepL will lead to educational disturbances that are still quite uncertain. For instance, it is still hard to assess the effect NMTs are exerting on teaching, particularly on foreign language writing classes. Moreover, this rapid progress has raised many questions for teachers on this topic. "Should I be teaching students how to use NMTs in the class?" "Are students using NMTs to do their homework and how do I know if students are using it?" "Is it wrong for students to use NMTs to do their homework?" Such questions do not have simple answers. However, this paper aims to start a dialogue among teachers concerning NMTs and foreign language education. This paper is divided into three sections. In the first section, a short history of machine translation is reviewed along with a search of the literature from the field of foreign language studies during its development. In the second section, a small-scale study that investigates the impact DeepL has on student essays is reported. Finally, in the third section, based on results from the study, how to manage the use of NMTs in the language classroom is considered and explored.

A Short History of Machine Translation Systems and Their Impact in the Foreign Language Classroom

Stage 1 The Beginnings (Late 1990s – 2006)

Over 70 years ago, before the collective reality of the World Wide Web and personal computers, Georgetown University Institute of Language and Linguistics showed off the power of an IBM computer, as it translated sentences from Russian into intelligible English. This event demonstrated the awe-inspiring ability of computers, but more importantly showed the real possibility that language could be translated by a machine (See https://www.ibm.com/ibm/history/exhibits/701/701_translator.htm for more information on this historical event). Over the next several decades, work on developing machine translation progressed, albeit slowly, using rule-based machine translation (RBMT). In short, humans (as in, linguists and computer scientists) establish a set of rules for the grammatical structure, word order, and phraseology for the source and target languages and then map one to the other. Thus, it is human-labor intensive and prone to many errors because language has many rules, but also many exceptions to these rules; words change in meaning, and language, generally speaking, can be rather ambiguous. In addition, RBMT lacks contextual cues and pragmatic knowledge and thus fails to recognize the connotative meanings of words and phrases. In others words, RBMT acts like a machine; it follows a set of rules, looks for matches between the two languages, and has no comprehension of what it is translating.

Two things dramatically changed the field of machine translation in the mid-1990's. First was the introduction of statistical machine translation (SMT) and then the spreading popularity of the World Wide Web (see Hutchins, 2007). Statistical models work by analyzing large amounts of data—translations, corpora—looking for patterns between the two languages. It can then construct a hypothesis or a statistical probability in how a word or a whole phrase should be translated. In regards to the World Wide Web, Babel Fish (an online MT) was established as part of AltaVista Translation, which was later taken over by Yahoo! and allowed anyone with the Internet to easily translate text between multiple languages.

On the other hand, in this period, these early MTs were given low assessments for their overall performance, for example, one rater noted that, as "when compared to expert human translators, MT systems perform only about 65% as well on the average" (Anderson,1995, p. 68). However, educators like Cribb (2000) speculated that the rise and sophistication of MTs could dramatically affect the TESOL profession since they provide a viable alternative to language learning, as he states, "some view the pursuit of foreign language competence as an admirable expenditure of effort, others may see it as unnecessary if an effective alternative exists" (p. 566).

As the Internet became more commonplace, MTs too became widely available and accessible to anyone with a computer and internet connection. As a consequence, students took advantage of them to complete their homework assignments. Yet, the translations were far from perfect and instead often resulted in "nonsensical compositions" that left many teachers "frustrated and baffled" (Luton, 2003, p. 769).

Stage 2 High Expectations & Shortcomings (2006 – 2016)

In 2006 Google and Microsoft came onto the scene with their own MTs and progress continued, but still these MTs had many of the same shortcomings as their predecessors, including overly literal translations, difficulties with idioms (metaphors), unnatural writing, and inability to account for cultural references (Correa, 2014; Niño, 2009). Researchers also began to analyze the role of MTs in the language classroom. For example, Groves and Mundt (2015) questioned whether MTs should be a "friend or a foe" for language teachers. They analyzed linguistic accuracy of Google Translate (GT) (Malay and Chinese to English pairings) and noticed great improvements of the translations from earlier MT systems. However, they found that GT still struggled with word choice, sentence structure, and missing words. Despite these shortcomings, they concluded their article by asserting that MTs should be viewed as a "friend" to language teachers and should be integrated into the class, particularly for use in English for Academic Purposes classes. They argued that doing this would allow instructors to shift the emphasis away from the low-level mechanics of English and begin to focus on "deep literacy." Deep literacy is when the reader (or writer) engages with the text in a sustained and intensive way and reads or writes "for its own value, for pleasure, for altering consciousness" (Davies, 2007 p. 51). Groves and Mundt (2015) stated that "as long as we accept this technology and try to work with it, not against it, it has the potential to make the teaching of EAP a much more exploratory and critical activity" (p. 120). In another study, Bahri and Mahadi (2016) analyzed the attitudes of international students in Malaysia towards using GT as a supplementary tool for learning Malay. Participants showed a strong positive attitude to using MT for language learning. To summarize, between 2006 and 2016, the nonsensical compositions had been replaced with translated text that began to really demonstrate the growing capabilities of MTs, however, MTs still lacked that subtle and deeper knowledge of language.

Stage 3 Deep Learning (2016 – current day)

All of this changed in the years 2016 and 2017 with the arrival of neural machine translation (NMT). Modeled after the human brain, NMT is designed to imitate the structure of the brain in complexity and the ability to learn from big data. Thus, NMTs are more humanlike than their nonneural MT competitors. Bentivogli et al. (2016) writes "[NMTs represent] a further step in the evolution of rule-based approaches that explicitly manipulate knowledge, to the statistical/data-driven framework, still comprehensible in its inner workings, to a sub-symbolic framework in which the translation process is totally opaque to the analysis" (p. 1). Despite the mystery of the inner workings of NMTs, Bentivogli et al. (2016) found in their analysis (between English and German pairings) that NMT outperformed a phrase-based SMT on a number of translation features, such as requiring less postedit effort, better translation on lexically rich texts, fewer morphological and lexical errors, and substantially fewer word-order errors.

Google in 2016 released its own version of an NMT, which some suggest has been able to reduce translation errors by more than 55 to 85% (Le & Schuster, 2016). Around the same time, DeepL was launched. Both of these NMTs continue to push boundaries of online machine translation. For instance, over the past couple of years, improvements have been continuously made, such as the *Transformer*, which has an attention mechanism allowing the NMT to better understand larger contextual information (Vaswani et al., 2017). Accepting the reality that this technology will only continue to improve over the next decade, it is a crucial time for teachers and language researchers to assess its current and future impact on students and teachers, as they design their classes.

The Present Study

Many previous studies have used Google Translate as the MT in their research (e.g., Bahri & Mahadi, 2016; Groves & Mundt, 2015), and DeepL has been widely under researched. Therefore, this study aims to fill this gap in the literature by focusing on DeepL. Moreover, it is important to explore how teachers evaluate the quality of essays that have used NMT compared to those that have not, and secondly whether teachers can identify an essay that has used NMT and what strategies they use to do this. Therefore, in this study the following two research questions were investigated:

RQ1. Will experienced English teachers evaluate essays written with NMT differently than those written without it?

RQ2. Can English teachers distinguish between essays written with NMT and those written without it? And what strategies do they use to arrive at this conclusion?

Participants

First-year students (*N* = 17) at a national university in Japan were recruited for this study during the second half of the first semester of 2021 (June to July). They were all in advanced English classes, as determined by a university placement test at the start of the year (roughly 650-700 TOEIC score based on a conversion provided by the placement test maker). At that time, they were enrolled in reading classes [and were not directly learning the structure of an essay. All students gave informed consent, were provided information about the purpose of the study, could opt out at any time, and after the completion of the study received extra credit points for their participation. The university was informed and approved this research project.

Design

Participants were randomly divided into two groups: (1) a control group (N = 9) and (2) a test group (N = 8). To answer RQ1, a between-subject study was conducted using a nonparametric two-independent samples Mann-Whitney *U* Test. The independent variable or grouping variable was based on whether the student participants used or did not use NMT to complete an essay writing task. The dependent variables came from the teacher evaluators who assessed the essays using a standard rubric (see Del Vecchio, 2017). As for RQ2, the evaluators responded to a single Likert item asking them to judge whether the essay had used NMT and then an open-ended question that asked them how they came to this conclusion.

Material and Procedure

In both groups, the student participants were instructed to write an argumentative essay (roughly 500 words in length) for or against holding the Tokyo 2020 Olympics in 2020. These Olympics were postponed for a year due to the COVID-19 pandemic, so they took place in the summer of 2021. A word length was used to provide the participants some direction for the length of the essay and consistency in length between the essays. They submitted their essays using Google Forms and did this outside of the scheduled class time. There was no time limit to complete this essay. In the control group, they were informed not to use any kind of machine translation system, but they could use a dictionary to write the essay. In the test group, they were told to write the essay first in Japanese (the L1 of the participants) and then to cut and paste the essay into DeepL to translate it, and to then read through it and check for any problems with the translation.

It should be noted here that this was completed on their own and it cannot be confirmed whether they followed these instructions or not.

Procedure for Assessment

I collected all 17 essays, removed any information to reveal who the authors were and coded them into a control (non-NMT) group (n = 9) and a test (NMT) group (n = 9). In order not to overburden the evaluators, I grouped the papers into three sets. I randomly assigned NMT and non-NMT papers to each set. I adapted Del Vecchio's (2017) rubric which consists of a 10-point scale on 4 dimensions of: *content & organization*, *unity & coherence*, *grammar & vocabulary*, and *style* as the rubric for evaluating these essays. The evaluation process was as follows. First, each evaluator received links to Google Forms that contained the beforementioned three sets. The essays appeared in random order. Then, each evaluator read the essays in that set and used the above rubric to evaluate them. Next, the evaluators used a single item 6-point Likert scale ("1" definitely not to "6" definitely) to judge whether they thought the student used NMT or not. In addition, below this Likert scale item, the evaluators then wrote their reasons why they came to that conclusion. Finally, they submitted the Google Form. They did this for all three sets. In total five (male = 4, female - 1) experienced English teachers who had some knowledge of NMTs were recruited to assist in this evaluation process. Three of these evaluators had over 10 years of university teaching experience with English as their L1. The other two evaluators had over 2 years of university experience with German and Hungarian as their L1, but are fluent in English. Each evaluator judged all 17 essays and were aware that some of the essays had used DeepL.

Results

To assess the interrater reliability between the five evaluators, Cronbach's alpha was used and the internal consistency was acceptable (α = .80). The four scores based on the rubric from the five evaluators were aggregated into one overall score (maximum 40 points) for each essay, resulting in the following descriptive statistics: the control group (M = 26.02, SD = 3.36) and test group (M = 33.13, SD = 2.22). To evaluate the difference between essay scores for the control and test groups a Mann-Whitney *U* Test was conducted. The test revealed significant differences in the scores for the control (Mdn = 5.22, *n* = 9) and test (Mdn = 13.25, *n* = 8) groups, *U* = 2.00, z = -3.274, p = .001, r = 0.79. To answer RQ1, the test group or students who used DeepL to complete the essay received a significantly higher score on the essays with a large effect size compared to the control group, who did not use any NMT.

As for RQ2, the evaluators responses to the single item Likert scale ("*Do you think this essay used NMT*?", "1" definitely not to "6" definitely) were aggregated into one score for each essay, resulting in the following: the control group (M = 2.16, SD = .58) and test group (M = 4.1, SD = .62). This indicates that the evaluators had a strong sense for which students had used DeepL. Based on their responses to the open-ended question, the evaluators used a number of strategies, to inform their decision whether they thought the essay used or did not use an NMT. When they gave the essay a *high* probability (4, 5, 6 on the 6-point Likert scale) that it had used an NMT, they noted a few characteristics in the essay, namely repetition (DeepL has a tendency to repeat sentences), it was simply "too good," and uncommon word choices or grammar for Japanese university students. On the other hand, when the evaluators gave the essay a "low" probability that it had used an NMT, they stated that it had bad or inaccurate grammar, used problematic or unnatural word choice, and finally contained spelling mistakes.

Discussion

This study analyzed the effect of NMTs on student essays—first, by examining how experienced teachers scored the essays using a standard rubric and secondly whether these teachers could accurately identify essays that used NMT. Several relevant outcomes can be observed from the results. The teachers evaluated the essays that used NMT more favorably. In addition, they used their knowledge about students' writing styles and their knowledge about NMTs to assess whether the students used NMT to complete their essays or not. So, the time of "nonsensical compositions" (Luton, 2003, p. 769) by students who too quickly embraced this new technology in the 1990s and early 2000s has been replaced with essays that are considered "too good" by students who use NMTs as indicated by evaluators in this research. However, results from this study show students who do not embrace this technology may receive lower evaluations for their essays. Therefore, from a practical side, this study highlights the importance of finding ways to integrate teaching about NMT into the writing class. Yet, what often seems to happen is that teachers simply avoid the topic altogether since it brings up a number of critical questions teachers might ask themselves:

- *A pedagogical question for teachers:* Does using NMT constitute a form of cheating? This is especially important since one study (Jolley & Maimone, 2015) has shown that 97.66% of students reported some use of online translation tools and 74.8% view using these tools for writing assignments as somewhat or completely ethical. A similar finding has been confirmed in another more recent study (O'Neill,

2019) with 87.7% of students self-reporting having used online translation for completing graded assignment even if its use was prohibited by the institution.

- *An assessment question for teachers:* How can I truly know whether or not a student has used NMT to complete a writing assignment and am I evaluating NMT-assisted essays higher? Do I need to change my rubric (minimizing grammar and vocabulary and emphasizing cohesion, structure and depth of content)?
- *An existential question for teachers*: If students use NMT, then how does this effect the goals of this writing course?

To consider why it is important to integrate NMT into a writing course is to really reconsider the role of the teacher, as one who is "teaching students the habits that will facilitate independent, lifelong, language learning" (Ducar & Schocket, 2018, p. 792). NMT has definitely become a tool that will assist in this language learning pursuit. However, if students simply cut and paste their writings into the NMT and then use the generated output, as is, very little learning occurs. As a result, teaching students *how* to effectively use NMT should become an important part of a writing course. Below are practical examples that focus on developing learners' critical thinking skills while interacting with NMTs.

First, in order to teach learners, the importance of maintaining agency and control over their writing, the teacher needs to stress the dangers of simply cutting and pasting text into an NMT and submitting the output to complete a writing assignment. This is both unproductive in terms of learning the language and could be discerned as ethically fraudulent behavior. Two higher order thinking skills (HOTS) are crucial when teaching how to use NMTs; how to *analyze* and how to *evaluate* the generated output.

Although NMTs have improved immensely over the past couple years, problems still do occur. So, one approach is to look for text that breaks the system and then to consider why this happens. For instance, consider the following English text taken from a CBS news story (https://www.cbsnews.com/news/remembering-and-recapturing-the-real-janis-joplin/) and translated to Japanese using Google Translate.

Figure 1
Sample Translation using Google Translate

English	Japanese
Janis Joplin became one of the biggest stars of her time, the greatest white female blues singer of the rock era, with a voice that could both blow your **mind** and break your **heart**.	ジャニス・ジョプリンは、彼女の時代の最大のスターの1人、ロック時代の最大の白人女性ブルース歌手になりました。その声は、あなたの心を吹き飛ばし、あなたの心を壊す可能性があります。

In English, "mind" and "heart" are clearly distinguishable constructs, with the latter being connected to the emotional system and the former to the cognitive system. They are also standard collocates with the preceding verbs "blow" and "break," respectively, and not interchangeable (e.g., "blow your heart" and "break your mind"), however it is a little surprising that the NMT does not recognize these phrasal structures. For instance, in the translation, both of these are translated with「心」*"kokoro"* and thus the translation is unnatural in Japanese. Doing such activities with students can raise their awareness of the limitations of NMT and that it is not a perfect system and they should be skeptical of the output. Using this example, students could also learn about the metaphorical uses of "mind" and "heart" in English. That could extend into teaching various idiomatic phrases that use these words (e.g., "keep it in mind," "come to mind," "have a big heart," "have a heart to heart") and then translate these phrases using an NMT and analyze the accuracy of these translations. In addition, using newspaper text can also provide for an interesting analysis since they often use a variety of creative writing techniques such as personification. Consider the following line from a CNN article (https://edition.cnn.com/2017/06/29/asia/china-xi-jinping-protesters/index.html).

- Original: In the last few years…, there has been increased pressure on Hong Kong's **windpipe**. – *DeepL translation to Japanese*: ここ数年。。。香港の風穴を開けるための圧力が高まっています。– *DeepL Backtranslation* to English- In the last few years, there has been increasing pressure on Hong Kong to wind down.

In the original English version, Hong Kong is personified as a human with human anatomy like a windpipe. Therefore, we map our knowledge about the body onto this city (Hong Kong). Moreover, we see the abstract geopolitical tension between Hong Kong and China as two living entities physically interacting with each other and in the case of this

example, China is putting pressure on the windpipe of Hong Kong. DeepL translates the sentence accurately, however for "windpipe," it translates it as 「風穴を開ける」 "*kazaana wo akeru,*" which in its metaphorical sense means to bring fresh air into an organization that is in a state of stagnation. In the backtranslation above, we can see that the sentence loses its meaning. The metaphor in English has a dramatic effect on the mental imagery of the sentence and using the NMT to highlight the breakdown in the translation might help enhance the students' metaphorical competence with the language. Doing activities such as these with students can highlight the difficulty of translating idiomatic expressions between languages, the use of personification in English, and the importance of mental imagery for rich meaning construction of the language.

Moreover, teachers and students can use the NMT as a computer-assisted language learning (CALL) tool to *analyze* and *evaluate* different possible translations. For example, Lee (2020) had students first compose a paper in their L1 (Korean) about a TED video, then translate this into English (without the assistance of MT), next use MT to translate their paper into English, and finally edit their initial English translation by comparing the two translations. Results from this study show significant improvements in the writing, as measured by a decrease in vocabulary and grammatical errors, between the initial and final versions, to which the author suggests, MT can function similarly to peer-editing.

Conclusion

This study examined how differently five EFL teachers evaluated essays composed using NMT and those that did not use such technology and whether they could identify the essays that used NMT and the strategies to arrive at this conclusion. Findings from this study suggest that EFL teachers are able to accurately identify essays that used NMT and tend to score these essays higher than those that did not use it. Advancements of NMTs like DeepL are accelerating at a rapid pace and therefore, it is important as teachers to discuss a best approach to teaching NMTs with the goals of raising learners' awareness of how to most effectively use them not only to improve their writing skills for the class, but also to develop their lifelong language and critical thinking skills.

Bio Data

Brian J. Birdsell is an Associate Professor in the Center for Liberal Arts Development and Practices at Hirosaki University. He has a Ph.D. in Applied Linguistics from The University of Birmingham (UK). His main research interest involves the overlapping fields of embodied semantics and conceptual metaphor theory, and applying these to second language learning. He publishes on matters of creativity and enhancing learner creativity in the classroom and content and language integrated learning (CLIL). <Brian@hirosaki-u.ac.jp>

References

Anderson, D. D. (1995). Machine translation as a tool in second language learning. *CALICO Journal*, 68-97. https://www.jstor.org/stable/24147744

Bahri, H., & Mahadi, T. S. T. (2016). Google translate as a supplementary tool for learning Malay: A case study at Universiti Sains Malaysia. *Advances in Language and Literary Studies, 7*(3), 161-167.

Bentivogli, L., Bisazza, A., Cettolo, M., & Federico, M. (2016). Neural versus phrase-based machine translation quality: a case study. *arXiv* preprint arXiv:1608.04631.

Correa, M. (2014). Leaving the "peer" out of peer-editing: Online translators as a pedagogical tool in the Spanish as a second language classroom. *Latin American Journal of Content & Language Integrated Learning, 7*(1), 1-20. https://doi.org/10.5294/3568

Cribb, V. M. (2000). Machine translation: The alternative for the 21st century?. *TESOL Quarterly, 34*(3), 560-569. https://doi.org/10.2307/3587744

Davies, C. (2007). What can technology do for/to English? In A. Adams, & S. Brindley (Eds.), *Teaching secondary English with ICT* (pp. 50-66). Maidenhead: Open University Press.

Del Vecchio, A. (2017). Developing scoring rubrics for ESL writing assessment. *Journal of Bunkyo Gakuin University, Department of Foreign Languages*, (16), 1-11.

Ducar, C., & Schocket, D. H. (2018). Machine translation and the L2 classroom: Pedagogical solutions for making peace with Google translate. *Foreign Language Annals, 51*(4), 779-795. https://doi.org/10.1111/flan.12366

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

Hutchins, J. (2007). Machine translation: A concise history. *Computer aided translation: Theory and practice*, 13(29-70), 11.

Jolley, J. R., & Maimone, L. (2015). Free online machine translation: Use and perceptions by Spanish students and instructors. In A. J. Moeller (Ed.), *Learn languages, explore cultures, transform lives* (pp. 181-200). Minneapolis: 2015 Central States Conference on the Teaching of Foreign Languages.

Kurzweil, R. (2004). The law of accelerating returns. In Teuscher C. (Eds) *Alan Turing: Life and legacy of a great thinker* (pp. 381-416). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-662-05642-4_16

Le, Q. V., & Schuster, M. (2016) A neural network for machine translation, at production scale. *Google AI Blog*. Retrieved from https://ai.googleblog.com/2016/09/a-neural-network-for-machine.html

Lee, S. M. (2020). The impact of using machine translation on EFL students' writing. *Computer Assisted Language Learning, 33*(3), 157-175. https://doi.org/10.1080/09588221.2018.1553186

Luton, L. (2003). If the computer did my homework, how come I didn't get an" A"?. *The French Review*, 766-770. https://www.jstor.org/stable/3133085

Niño, A. (2009). Machine translation in foreign language learning: Language learners' and tutors' perceptions of its advantages and disadvantages. *ReCALL, 21*(2), 241-258. https://doi.org/10.1017/S0958344009000172

O'Neill, E. M. (2019). Online translator, dictionary, and search engine use among L2 students. *CALL-EJ*, 20(1), 154-177. Retrieved from http://callej.org/journal/20-1/O'Neill2019.pdf

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998-6008).