ENGR 401 Assignment 1: Code of Ethics for AI

Bryony Gatehouse

April 2021

1 Introduction

Machine Translation is the use of software to translate a text from one language to another. They are widely used in the internet to increase the amount of information available to people. However, ethical dilemmas have arisen due to the development of these systems.

This report will examine real-life ethical problems with the goal of designing and creating a Machine Translation Code of Ethics. The explored problems will include Machine Translation failures and consideration of the human language. In this study, four main area will be discussed, based off the designed principles. These are Privacy, Incorrect or Misleading Data, Biases, and Learning from Mistakes.

2 Background

2.1 Machine Translation

2.1.1 Statistical Machine Translation

Statistical Machine Translation learns the translations of words and small groups of words by training with source texts and their human translations [1]. This method isn't able to learn the structure of the text, causing translations between two languages with different structures, like Japanese and English, results in sentences that don't look or sound right. However, it was very popular before Neural Machine Translation.

2.1.2 Neural Machine Translation

In 2016, Google Translate decided to start the change to Neural Machine Translation. Instead of piecing a sentence together from known words, the Neural Machine Translation System is able to understand, rearrange and adjust the sentence until it matches the target language structure and grammar [2]. This produces a more readable sentence.

2.2 Codes of Ethics

A Code of Ethics describes the ethics that a worker should follow in order to appear professional. But what is Professionalism? Defined in the dictionary as the conduct and aims that characterize a profession or professional person [3], it is the behaviour the public expects of the workers. For example, it is expected that information disclosed during a doctor appointment will not be spread publicly. This would be laid out in the Code of Ethics for people working in the medical profession. Placing this information into a code helps workers understand the standards they need to uphold, and supports them when they need to act ethically.

2.2.1 Engineering NZ Code of Ethical Conduct

The Engineering NZ Code of Ethical Conduct [4] is a list of standards that all New Zealand engineers should be adhering to. The Engineering profession covers a wide range, from structural engineers who work in architecture, to software engineers who develop software applications and systems. Due to this, it contains broad ethical principles which are able to encompass the high standard that engineers are expected to uphold. The main idea behind all the principles is to act thoughtfully and appropriately in both the workplace and in their code.

The standard *take reasonable steps to safeguard health and safety* could be talking about the effects that the program could have on people. For example, a software program which signals a rocket to arms it's ejection charge could fail to send the signal. This would prevent a controlled descent of the rocket, possibly causing serious damage to an individual or property. However, the same standard could also be referring to practising health and safety in the office. For example, not coming into work while sick or tidying the electrical wires protects the health and safety of the other workers and clients.

The last principle of the ENZ Code of Ethical Conduct requests that any breaches of the Code be reported. If a worker is being threatened or pressured to ignore a breach of ethics, they can use their commitment to up keeping the Code to protect them.

2.2.2 NZ Medical Association's Code of Ethics

Working in the medical profession requires a lot of knowledge and a high level of professionalism. The NZ Medical Association's Code of Ethics [5] makes sure to include descriptive principles and recommendations in order to properly guide medical professionals. Similar to the Engineering NZ Code of Ethical Conduct, it requires workers to work thoughtfully and kindly. However, as medical professionals are faced with more ethical dilemmas, the Code of Ethics also includes recommendations which outline specific scenarios in order to "convey an overall pattern of professional behaviour".

The sixth principle states that *the professional should be improving their knowledge and skills, and attending to their health and well-being, in order to give the best advice and treatment to the patient.* After all, a sound judgement

can't be made if the professional has just worked a 10 hour shift and haven't had time to study the particular problem. However, it also requires employers to allow the medical practitioners time for their well-being and to learn.

A recommendation, *render medical service to a patient without discrimination*, is explicitly stated in the principles. However, it is an extension of the first principle which states *consider the health and well-being of the patient to be your first priority.* Thus, medical professionals should lay aside any beliefs, including racism or sexism, to treat the patient to the best of their abilities. While these standards might not prevent unconscious bias, it encourages the professionals to be consciously aware of how they are treating people.

3 Analysis

The ENZ Code of Ethical Conduct uses broad principles to cover a large area of ethics. Machine Translation is a sub-profession of Software Engineering, so should fall under the umbrella of this Code. Thus, similar to the NZMA Code of Ethics, the Machine Translation Code of Ethics will use recommendations to paint a picture of the expected behaviour of a machine translation professional.

Principles:

1. Respect client privacy.
2. Don't provide incorrect or misleading training data.
3. Avoid learning biases.
4. Learn from mistakes.

The designed set of recommendations, building on the principles from the ENZ Code of Ethical Conduct:

1. If the Machine Translation System uses online learning, the saved data should contain no identifiers.
2. Machine Translation Systems shouldn't store the whole phrases that they learn.
3. Translated phrases which use substitutions shouldn't be used to train Machine Translation Systems.
4. Engineers should ensure that the training set uses the best translation word choices available.
5. If the Machine Translation System uses online learning, the engineer should regularly test that it making correct decisions.
6. Clients should be able to quickly indicate that something is wrong with their translations.
7. Initial training sets should avoid presenting biases.

8. Biases should be considered when designing output.

9. If translation mistakes are known, the Machine Translation System should receive training to avoid continuing these mistakes.

3.1 Respect Client Privacy

Principle 9 of the NZ Privacy Act states that personal information cannot be kept for longer than is required [6]. For Machine Translation Systems, this should mean that the statements and translations should be deleted after the Machine Translation process has finished. However, free Machine Translation Services, like Google Translate use the provided statements as 'payment' to improve their translation services and other purposes, like advertising or marketing [7].

Recording the data for advertising or marketing uses would be against the 9^{th} Principle. However, improving the translation services with user input data is "further processing" of the data [8] which may be allowed, and allows the Machine Translation System to develop along side the human language (e.g. new slang). However, this data should be recorded in a manner to erase all personal connections to the data. Only the statement entered, and possibly the country of origin, should be kept, with statements kept separately. Thus, the first two recommendations, that saved data from online learning shouldn't contain identifiers, to remove it as much as possible from the client, and that saved phrases shouldn't be kept whole, in order to keep the original text as private as possible.

3.2 Don't Provide Incorrect or Misleading Training Data

Based off the following examples, problems do arise in Machine Translation System's translations. The fifth recommendations states that any online learning Machine Translation System should be regularly tested to check for translation mistakes to avoid real-world problems. However, if a client manages to find a translation mistake before the engineers, the sixth recommendation states that they should be able to quickly

3.2.1 Substitutions

When translating, humans are able to consider the audience. Thus, different words might suit one situation better than another, in order for the audience to understand the message in the text. However, while Machine Translation Systems don't have this ability, they can also be tripped up by substitutions being used.

In Franz Och's Google TechTalk [9], he mentioned a problem which occurred with Google Translate translating 'Heath Ledger' in to 'Tom Cruise'. The original text where this mistake was learnt was translated for Argentinians who might not have recognised the example famous male movie star 'Heath Ledger'.

The problem was, even outside of this scenario, Google Translate thought this was the correct translation.

Based off this example, imagine a translated news paper heading stating that 'Tom Cruise' would be visiting only for 'Heath Ledger' to show up. This could lead to potential uproar and backlash against the company and Heath Ledger. Thus the third recommendation, for training sets to not include substitutions.

3.2.2 Ambiguous Word Choice

Similar to substituting more appropriate words based on the audience, there are many translations where the word choice can greatly change the meaning. During the translation of the Treaty of Waitangi, specific words were chosen to misdirect the Māori into believing they would still reign over their lands.

In the First Article, the translation of the sentence *the Chiefs ... give absolutely to the Queen of England for ever the complete government over the land* [10] used the Māori word *kawanatanga* to mean government. Unfortunately, it was a missionary neologism based on the English word governer, and not widely known. Further more, the Māori had no concept of a governor, nor how much authority one would have. Thus, it wasn't clear to the Māori that this sentence was saying they were to hand over their sovereignty to the Queen.

In the Second Article, the chiefs were promised "chieftainship" or *tino rangatiratanga* over their lands, villages and treasures. This words was known to the Māori to represent the power, rights and authority of the chief, i.e. the sovereignty. However, the English version used the word possession [11] meaning that the chiefs would have the legal rights to the land, but still must follow the Queen's rule on that land.

This particular case was done out of malice towards the Māori chiefs. However, a Machine Translation System can only derive meaning from it's learning material. From this example, the fourth recommendation can be designed: engineers should ensure that the training set uses the best translation word choices available.

3.2.3 Solving Problems

Based off these examples, problems do arise in Machine Translation System's translations. The fifth recommendations states that any online learning Machine Translation System should be regularly tested to check for translation mistakes to avoid real-world problems. However, if a client manages to find a translation mistake before the engineers, the sixth recommendation states that they should be able to quickly report this problem to the service provider. Finally, any mistake should be trained out of the Machine Translation System, as the ninth recommendation states.

3.3 Avoid Learning Biases

Jobs are an "interesting window into the nature of gender bias" [12] due to the unequal proportion of male representation in the work place. This is also reflected in university where only 35% of STEM students are female [13].

In 2018, a study decided to explore Google Translate's biases towards genders related to certain jobs. This was achieved by translating sentences about wide varieties of jobs from gender-neutral languages to English and examining the resulting sentence.

Hungarian	**English**
ő egy ápolónő	she's a nurse
ő egy mérnök	he is an engineer
ő egy tudós	he is a scientist
ő egy pék	she's a baker

Most languages had the majority of translated sentences containing male pronouns. Female pronouns were less widely represented, with some languages translating only 1% of sentence with female pronouns. Finally, gender-neutral pronouns were occasionally used, with most languages having preference to either gender-neutral or female pronouns for it's non-male sentences. These results were compared to the percentage of females in the job sectors, expecting a similar trend. However, the trends didn't match. Since Google Translate uses online learning, it has obviously developed the same bias towards male pronouns that English speakers tend to have.

Thankfully, Google Translate decided to change its policy to offer both male and female translations of sentences in an attempt to debias their machine. This was the inspiration for the eighth recommendation, which asks for biases to be considered in the output of the Machine Translation System.

As an important note, biases could be considered important in some decisions. For example, it has been proven that males are more likely to cause or be in motor vehicle accidents, with the total number of male deaths caused by motor vehicle accidents in 2019 equalling to almost 2.5 times higher than the number of female deaths [14]. Thus, car insurance tends to cost more for male drivers than females.

3.4 Learn from Mistakes

Facebook uses it's own Machine Translation System to allow people's posts to reach a wider audience. In 2017, a Palestinian man posted يسبِيهُم, or *yusbihuhum* which should have translated to "good morning" alongside a photo of him leaning against a bulldozer. However, Facebook translated it to "hurt them" in English or "attack them" in Hebrew. As bulldozers had been recently

used in multiple terrorist attacks, along with the mistranslated phrase, the police were alerted and the man arrested.

While the police hadn't checked the untranslated version of the man's post before arresting, the truth came to light before worse could happen. Facebook later released a statement admitting their fault and stating that other mistakes have happened before. They also admitted that Arabic is hard for Machine Translation System to learn.

In 2020 Facebook again had to apologise when Burmese posts translated into English had the Chinese leader, Xi Jinping, translated as *Mr. Shithole* [15]. After running tests, it was discovered that Facebook's system, when faced with an unknown word, attempts to replace it with another word. Unfortunately, multiple words starting with "xi" and "shi" were translated to the same vulgarity in English. The mistranslation was discovered in a post about the Chinese leader visiting Myanmar, a south-east Asian country, and could have damaged the new relationship between the two countries.

Both incidents could have lead to devastating effects for the posters. However, Facebook clarified, after the second event, that they would strive to ensure that it wouldn't happen again. This is in line with the ninth recommendation and last principle, to learn from mistakes. A trend of mistakes, as mentioned in the first incident, should be noticed quickly and the system should be trained more before the public can translate those phrases again to avoid real-world problems arising, like these two.

4 Conclusion

A Code of Ethics for Machine Translation needs to consider the problems that arise from using machines to translate text from one language to another. As explored already, this could rise to losing the meaning of the original text which could lead to jail or worse for the original writer. To avoid this, the principles offer guidance for the engineer and Machine Translation System: respect client privacy, don't provide incorrect or misleading training data, avoiding learning biases, and learn from mistakes.

In conclusion, the Machine Translation Code of Ethics expects systems and engineers to make decisions in order to provide the best possible experience for the client.

References

[1] D. Kenny, "The ethics of machine translation", *Proceedings of the XI NZSTI National Conference*, 2011. [Online]. Available: `https://core.ac.uk/download/pdf/11311284.pdf` (Accessed: Apr. 7, 2021).

[2] B. Turovsky, "Found in translation: More accurate, fluent sentences in google translate", *The Keyword, Google*, Nov. 2016. [Online]. Available: https://blog.google/products/translate/found-translation-more-accurate-fluent-sentences-google-translate/ (Accessed: Apr. 24, 2021).

[3] *Professionalism*. [Online]. Available: https://www.merriam-webster.com/dictionary/professionalism (Accessed: Apr. 12, 2021).

[4] *Code of ethical conduct*, 2021. [Online]. Available: https://www.engineeringnz.org/engineer-tools/ethics-rules-standards/code-ethical-conduct/ (Accessed: Apr. 7, 2021).

[5] *Code of ethics for the new zealand medical profession*. Wellington, New Zealand: New Zealand Medical Association, Jun. 2020, ISBN: 9780473288877. [Online]. Available: https://assets-global.website-files.com/5e332a62c703f6340a2faf44/5fbd645fe15640fa981fa469_Code%20of%20Ethics%20Redesign%202020%20version%204.pdf (Accessed: Apr. 7, 2021).

[6] *Implementing the privacy principles*, Digital.Govt.NZ, Aug. 2020. [Online]. Available: https://www.digital.govt.nz/standards-and-guidance/governance/managing-online-channels/security-and-privacy-for-websites/designing-for-security-and-privacy/implementing-the-privacy-principles/ (Accessed: Apr. 24, 2021).

[7] M. Linder, *The data security issues around public machine translation - a translator perspective*, Excel Translations. [Online]. Available: https://exceltranslations.com/data-security-public-machine-translation/ (Accessed: Apr. 24, 2021).

[8] P. Kamocki and J. O'Regan, "Privacy issues in online machine translation services - european perspective", in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, May 2016, pp. 4458–4462. [Online]. Available: https://www.aclweb.org/anthology/L16-1706.pdf (Accessed: Apr. 7, 2021).

[9] F. Och, "Statistical machine translation", *Google Tech Talk*, Jun. 2009. [Online]. Available: http://www.youtube.com/watch?v=y_PzPDRPwlA (Accessed: Apr. 7, 2021).

[10] P. Moon and S. Fenton, "Bound into a fateful union: Henry williams' translation of the treaty of waitangi into maori in february 1840", *The Journal of the Polynesian Society*, vol. 111, no. 1, pp. 51–63, Mar. 2002. [Online]. Available: http://www.jstor.org/stable/20707042 (Accessed: Apr. 7, 2021).

[11] "Read the treaty", *Ministry for Culture and Heritage*, Jun. 2020. [Online]. Available: https://nzhistory.govt.nz/politics/treaty/read-the-treaty/english-text (Accessed: Apr. 7, 2021).

[12] M. O. R. Prates, P. H. Avelar, and L. C. Lamb, "Assessing gender bias in machine translation: A case study with google translate", *Neural Computing and Applications*, vol. 32, no. 10, pp. 6363–6381, Mar. 2019. DOI: 10.1007/s00521-019-04144-6. [Online]. Available: https://link.springer.com/article/10.1007/s00521-019-04144-6 (Accessed: Apr. 7, 2021).

[13] E. López-Iñesta, C. Botella, S. Rueda, A. Forte, and P. Marzal, "Towards breaking the gender gap in science, technology, engineering and mathematics", *IEEE Revista Iberoamericana de Tecnologias del Aprendizaje*, vol. 15, no. 3, pp. 233–241, 2020. DOI: 10.1109/RITA.2020.3008114. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9137264 (Accessed: Apr. 7, 2021).

[14] *Fatality facts 2019: Males and females*, Mar. 2021. [Online]. Available: https://iihs-prod.iihs.org/topics/fatality-statistics/detail/males-and-females (Accessed: Apr. 14, 2021).

[15] M. Padilla, "Facebook apologizes for vulgar translation of chinese leader's name", *The New York Times*, Jan. 2020. [Online]. Available: https://www.nytimes.com/2020/01/18/world/asia/facebook-xi-jinping.html.