Machine translation for te reo Māori

Te Taka Keegan

Introduction

What is Machine Translation? Machine Translation (MT) is a process where computer software is used to translate texts from one natural (or spoken) language to another. Early research centred on two distinct approaches: Rules Based Machine Translation (RBMT) and Statistical Machine Translation (SMT). In simple terms RBMT makes use of large sets of linguistic rules that define languages whereas SMT uses statistical techniques to build language models from large language corpora. Increases in computing power and the amount of language corpora available has meant that SMT had become the preferred option with recent advances in neural networks also being applied to improve the accuracy of SMT. For commercial reasons, this is an area of research that has generated a lot of interest and funding support from some major international computer companies, including Google[1], Microsoft[2] and Facebook[3].

Why is Machine Translation important for te reo Māori? A te reo Māori purist may argue that it is not important to focus activities on having a machine undertake translations for te reo Māori; if people want to understand te reo Māori then they should put in the effort to learn the language. This line of reasoning is difficult to argue with. But, from a Māori language activist perspective the value of MT is not so much in the translation of te reo Māori to (say) English, but rather the translation of English to te reo Māori. If this translation can be done efficiently, with low costs, it will assist in the proliferation of te reo Māori into new contexts, new environments and will assist its normalisation in New Zealand's society.

At this time, two of the major international companies, Google and Microsoft, have invested significantly in MT for te reo Māori. This paper summarises their endeavours and reports on the quality of translations they have been able to generate.

Google's support for te reo Māori

Google's support for te reo Māori began in 2001 when it enabled the Māori language as one of the languages in its *Google in Your Language* programme. This programme was an avenue for local language communities to translate interface strings so that the Google Web Search home page could appear in a local language. As translations were undertaken part time, it took 7 years before the Google home page appeared in te reo Māori.

A year later, in 2009, the Māori language was enabled on the Google Translator Toolkit (GTT). The GTT is a Translation Memory tool to assist translators undertake translation work by placing at their fingertips dictionaries, word lists and previous translations. Testing was undertaken on this tool that suggested it could increase the productivity, quality and consistency of Māori language translations. If 70% of Māori translators used this tool ,a further 20 million words could be translated per annum

(Keegan & Evas, 2011). However, significant uptake of this tool by reo Māori translators is yet to be realised.

One of the issues with typing te reo Māori is the ability to type the macron character over a vowel to indicate that the lengthening of that vowel, which is important to indicate word meaning and pronunciation. In 2010 Google released a keyboard to allow the typing of the macron on its web search home page.

In 2017, Google worked with Vodafone to create a feature that allows for the entry of correct pronunciation of Māori place names[4].

While the above tools and features are important for the use of te reo Māori with Google tools, perhaps the most significant facility released by Google for te reo Māori was Google Translate. Google Translate became available in te reo Māori in December 2013. However, it is important to be aware that the significance and usefulness of Google Translate for te reo Māori clearly lies with the accuracy of translations that it can produce. A brief investigation on the usefulness of Google Translate in te reo Māori was undertaken and is described next.

Google Translate in te reo Māori investigation

In the New Zealand summer of 2013-2014, a Master's student at Waikato University, Whitiaua Ropitini, undertook a brief investigation into the support for the Māori language inclusion in Google Translate (Ropitini, 2014). Six Māori language participants were identified with mixed backgrounds: four had a high knowledge of te reo Māori, four had experience with Māori language translation, and four had experience using computers. The participants were introduced to Google Translate (see *Figure 1*) and asked for qualitative, somewhat anecdotal, feedback.

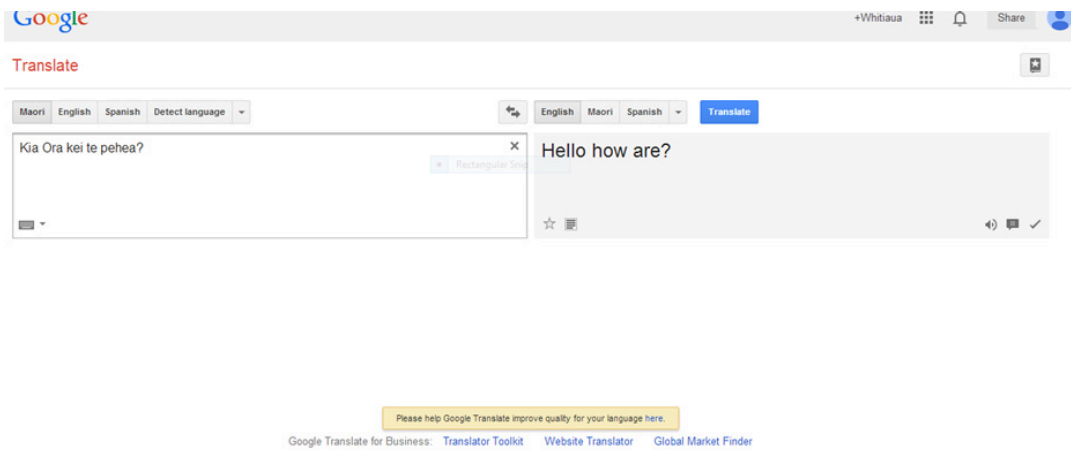

Figure 1: *Google Translate*

When asked how they felt about the inclusion of te reo Māori in Google Translate five of six were supportive, suggesting the mana o te reo Māori was uplifted by being available through this tool. When asked about the quality of translations all of the participants commented

that while the tool was able to translate (to/from Māori) there were errors. The participants questioned how these errors could be corrected. When asked if Google Translate could assist in the long term survival of te reo Māori, all six participants were positive in the benefits that this tool could provide with most participants suggesting improvements and corrections would be necessary, and suggesting these corrections should be driven by Māori language supporters.

Microsoft's support for te reo Māori

Microsoft's support for te reo Māori began in 1998 when it acknowledged an operating system keyboard was needed to produce the macron character. This was subsequently created and made available in 2003. It has been available in all subsequent versions of the Microsoft Windows operating systems.

Microsoft's next major milestone was the translation of Windows XP and Office 2003 into te reo Māori. This was, perhaps, the most ambitious Māori language IT project ever completed, involving the translation of over 900,000 words in 180,000 separate strings. Subsequent version of Windows and Office have also been translated into te reo Māori. The translation was significant because it now allows a computer (running Microsoft Windows) to be used totally in te reo Māori. One would expect that a key user of this software would be Māori medium education, however studies have shown that, for a number of reasons, the uptake in this sector has not been as high as expected (see Keegan & Mato, 2011; and Mato, Keegan, Cunliffe, & Dalley, 2012).

Microsoft has also supported the translation of Minecraft coding sessions into te reo Māori for the Hour of Code as part of 2016 Computer Science Education Week.

Pokapū Whitireo Charitable Trust

In 2013, Microsoft began discussions with some Māori language supporters to set up a trust whose functions include the gathering of a body of translated works that can be used for language research and revitalisation. This led to the registration of the Pokapū Whitireo Charitable Trust in 2015 that has three trustees and lists Professor Pare Keiha as the contact. While Microsoft do not have a member on the Trust, they provided legal support to set up the trust. It is envisaged that all resources collected by the Trust are available to all who are undertaking research on te reo Māori, including Microsoft and the Microsoft Translator Hub.

Microsoft Translator Hub

The Microsoft Translator Hub (MTH) is an extension of the Microsoft Translator service. It is an online service that provides the facility for language communities to create their own machine translation tool. This facility was made available by Microsoft to the Māori language community in 2013. Perhaps due to a limited capacity in the Māori language community the service was never activated until 2017, when a Waikato University undergraduate computer science student, Jasmin Cairns, created a translation system and

compared it with the Google Translation System.

Microsoft Translator Hub comparison study

The study undertaken by Jasmin Cairns (2017) had two components. First she built a translation system using MTH, then she compared the quality of Māori language translations produced by the MTH to those produced by Google Translate.

To build the MTH system, Cairns sourced data in the form of translated sentence pairs: English to Māori or Māori to English. She sourced this data from the Pokapū Whitireo Charitable Trust, from corpora that were gifted by the Faculty of Māori and Indigenous Development of Auckland University of Technology, and from corpora that were gifted by the Computer Science Department of Waikato University. In all she collected approximately 134,000 translated pairs. The data was uploaded to the MTH, and after a training process, the online system was created. It was made available for testing (see *Figure 2*), but was never deployed publically.

Maori	English
Mauria he hāmarara, kai te tīmata te kōua.	**Ref:** Take an umbrella, it's starting to spit. **MT:** Take an umbrella and start the kōua.
Ki te iwi o Ngāti Porou, he maunga rangatira a Hikurangi.	**Ref:** To the people of Ngāti Porou, Hikurangi is a grand mountain. **MT:** To the people of Ngāti Porou, a captain of the Sky.
Na nga pakeke i whakarato nga paraihe o te kura.	**Ref:** The elders gave out the school prizes. **MT:** The age of the prize of the school.
I tonoa mai māku e torō te nama wīni.	**Ref:** I was asked to draw the winning number. **MT:** Sent me draw a win.
E ngākau kore ana ahau ki te whai atu i tēnā ara.	**Ref:** I am disinclined to follow that course of action. **MT:** I have no heart to follow that path.

Figure 2: *Sample output from Microsoft Translator Hub (Māori to English)*

Cairns used two methods to undertake comparison translation quality. She used a machine translation evaluation software called Asiya which compared the translations over 9 metrics including BLEU, GTM, IO, WER and PER. She also had the 1,000 outputs manually checked by two human verifiers proficient in te reo Māori. The combined results showed that there were no significant differences in the output of the two systems, Microsoft Translation Hub and Google Translate.

The usefulness of machine translation for te reo Māori

For modern translations systems to translate with accuracy they require large amounts of language corpora, and in particular digitised translated pairs of data. The MT systems built by Google and Microsoft produce accurate translations for the largest languages of the world because there are large datasets available that these systems utilise to refine translation accuracy.

The amount of Māori language digitised data available for MT is not sufficient enough at this time for the MT systems to produce accurate translations. Consequently the publically available Google Translate in Māori MT system and the, yet to be deployed

publically, Microsoft Translator Hub system for te reo Māori should be used with that in mind. These Māori MT systems can perform translations in te reo Māori, but these translations need to be verified before being used with authority. This is something that Hamilton mayoral candidate James Cameron discovered, when in 2016 he used Google Translate to translate his personal profile into Māori and the resulting output was posted to every home in Hamilton (Smallman, 2016).

But the outlook is bright. More digitised Māori language data is being made available all the time. A recent web crawl for Māori language data found over 2,200 web pages with Māori language texts on them[5], an amount that would be inconceivable just a few years ago. This, coupled with the knowledge that neural networks in Machine Translation is constantly improving the quality of translations, and suggests there is a future for MT in te reo Māori. The question is not 'Will this happen?', but rather 'When will this happen and how can Māori language activists be aware of this so that it can have a positive impact on the normalisation of te reo Māori?'

Endnotes

1. https://research.googleblog.com/2016/09/a-neural-network-for-machine.html
2. https://blogs.msdn.microsoft.com/translation/2016/11/15/microsoft-translator-launching-neural-network-based-translations-for-all-its-speech-languages/
3. https://code.facebook.com/posts/289921871474277/transitioning-entirely-to-neural-machine-translation/
4. See: https://www.sayittika.co.nz/
5. Personal correspondence from Kevin Scannell, August 2017.

References

Cairns, J. (2017), *Comparison of Google Translate with Microsoft Translator Hub in Māori* (Unpublished report for University of Waikato, Computer Science COMP477 research paper). Hamilton, New Zealand: Computer Science Department, University of Waikato.

Keegan, T., & Evas, J. (2011). Can computer assisted translation tools improve the productivity and quality of minority languages translations? *He Pukenga Kōrero: A Journal of Māori Studies, 10*(2), 34-41.

Keegan, T., & Mato, P. (2011, February). *Are computer applications being used in Te Re Māori? Some initial findings.* Paper presented at the 2nd International Conference on Language Conservation and Documentation: Strategies for Moving Forward, University of Hawaiʻi at Mānoa, Hawaiʻi. Retrieved from http://hdl.handle.net/10125/5184

Mato, P., Keegan, T. T., Cunliffe, D., & Dalley, T. P. M. (2012) Perception and use of software with a Māori-language interface in Māori-medium schools of New Zealand. In T. Kaʻai, M. Ó Laoire, N. Ostler, R. Kaʻai- Mahuta, D. Mahuta, & T. Smith (Eds.), *Proceedings of the Conference FEL XVI, Language Endangerment in the 21st Century: Globalisation, Technology and New Media*

(pp. 87–92). Auckland, New Zealand: Te Ipukarea and Printsprint, AUT University.

Ropitini, W. (2014). *He kohinga kōrero mō Google Whakamāori* (Unpublished report for Summer Scholarship). Hamilton, New Zealand: Computer Science Department, University of Waikato.

Smallman, E. R. (2016, September 20). *Hamilton candidate's Google translation misguided gibberish, Maori likely to switch off.* Retrieved from Stuff: http://www.stuff.co.nz/national/politics/84398001/hamilton-candidates-google-translation-misguidedgibberish-maori-likely-to-switch-off