Institution: University of Edinburgh

Unit of Assessment: 11

Title of case study: Neural machine translation improves translation quality

Period when the underpinning research was undertaken: 2015 – 2018

Details of staff conducting the underpinning research from the submitting unit:

Name(s):	Role(s) (e.g. job title):	Period(s) employed by submitting HEI:
Alexandra Birch-Mayne	Senior Research Fellow	2010 – present
Ulrich Germann	Senior Researcher	2012 – present
Barry Haddow	Senior Research Fellow	2005 – present
Kenneth Heafield	Reader	2015 – present
Rico Sennrich	Lecturer	2015 – present

Period when the claimed impact occurred: 2016 – 2020

Is this case study continued from a case study submitted in 2014? No

1. Summary of the impact

Research at the University of Edinburgh (UoE) into neural machine translation (NMT) has produced a pioneering translation toolkit, which has been widely adopted for commercial and public service use. The software, named Marian, uses the UoE's novel *byte pair encoding* (BPE) and back translation techniques to increase existing machine translation capabilities to include niche vocabulary and unusual words. It was made open source in 2016, and has since been adopted by Microsoft to underpin the Microsoft Translator embedded translation services, and by specialist translation companies such as Lingo 24. It also formed the foundation for the software WIPO Translate, which is used by WIPO, a United Nations (UN) agency that verifies patents for the UN's 192 member states.

2. Underpinning research

Researchers at the University of Edinburgh (UoE) pioneered neural machine translation (NMT): the use of end-to-end neural network models for large-scale machine translation. During 2015 – 2016 the UoE team developed a number of technical innovations which resulted in NMT becoming the preferred approach to machine translation both in research and in practical applications.

The first innovation addressed the issue of ensuring that NMT systems could translate languages with very large vocabularies, including rare words. The UoE team introduced the approach of *byte pair encoding* (BPE) which enables effective language modelling at the sub-word level [3.1]. BPE enables infinite vocabulary sizes to be modelled using a finite set of sub-words, and results in significant improvements in translation quality. This unsupervised method is language independent, which makes it easy and quick to apply to any language. This extremely highly cited work is now the standard approach for language representation in machine translation and has been used widely in other areas, for example speech recognition.

A second innovation addressed the problem of training NMT systems with data from only a single language. Training machine translation systems requires parallel source and target

language texts. For many language pairs there is a relatively small amount of such parallel data, although there may be extensive data from either or both of the individual languages. The UoE team introduced a monolingual technique, *back-translation*, which enables target language data to be used when training NMT systems, without the need for parallel source language data [3.2]. Back-translation is of particular use for less well-resourced language pairs, and again has become a standard approach in NMT.

A third innovation addressed the need for NMT systems to handle correctly issues such as capitalisation, numbers, and dates. The UoE team introduced a *factored representation* for NMT [3.3], based on linguistic features, which enables the system to learn such information in a compact and efficient way. Again, this has become a very widely used technique.

These three innovations were integrated in number of systems developed by the UoE team, and released as open source software. In 2016, the UoE team used their NMT systems for the Workshop on Statistical Machine Translation (WMT) international evaluation of MT systems [3.4], which includes the major research teams in the area. The UoE systems were the most accurate in 7 out of 8 language pairs, often by a significant margin. This was the first time that neural models had conclusively beaten previous statistical machine translation models across a variety of language pairs.

The open source Nematus [3.5] and subsequent Marian [3.6] toolkits, developed by the UoE team, provide efficient implementations of NMT models and algorithms, including BPE, back-translation, and factorised representations.

3. References to the research

3.1. Sennrich, R., Haddow, B., & Birch, A. (2016). Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (pp. 1715-1725). Berlin, Germany: Association for Computational Linguistics (ACL). https://doi.org/10.18653/v1/P16-1162 (**3187 citations**)

3.2. Sennrich, R., Haddow, B., & Birch, A. (2016). Improving Neural Machine Translation Models with Monolingual Data. *In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (pp. 86-96). Berlin, Germany: Association for Computational Linguistics (ACL). https://doi.org/10.18653/v1/P16-1009 (**1119 citations**)

3.3. Sennrich, R., & Haddow, B. (2016). Linguistic Input Features Improve Neural Machine Translation. In *Proceedings of the First Conference on Machine Translation, Volume 1: Research Papers* (pp. 83-91). Berlin, Germany: Association for Computational Linguistics (ACL). https://doi.org/10.18653/v1/W16-2209 (**301 citations**)

3.4. Sennrich, R., Haddow, B., & Birch, A. (2016). Edinburgh Neural Machine Translation Systems for WMT 16. In *Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers* (pp. 371-376). Berlin, Germany: Association for Computational Linguistics. https://doi.org/10.18653/v1/W16-2323 (**387 citations**)

3.5. Sennrich, R., Firat, O., Cho, K., Birch-Mayne, A., Haddow, B., Hitschler, J., Junczys-Dowmunt, M, ... Nadejde, M. (2017). Nematus: a Toolkit for Neural Machine Translation. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 65-68). Valencia, Spain : Association for Computational Linguistics (ACL). https://doi.org/10.18653/v1/E17-3017 (**330 citations**)

3.6. Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Neckermann, T., ... Birch-Mayne, A. (2018). Marian: Fast Neural Machine Translation in C++. In *Proceedings of ACL 2018, System Demonstrations* (pp. 116–121). Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-4020 (**236 citations**)

Citation counts obtained from Google Scholar 2020-12-14.

Key research grants
European Commission: QT21 (645452, GBP275,551); HimL (644402, GBP600,545); TraMOOC (644333, GBP356,391); SUMMA (688139, GBP1,332,472); MMT (645487, GBP469,472)

4. Details of the impact

The research resulted in the software toolkit Marian, which was made open source on Github in 2016, with an updated version released in November 2017. Marian has proved extremely adaptable in practice, and has demonstrated positive effects on a wide range of organisations, meeting very different needs.

Its most high-profile commercial adopter is Microsoft. In 2019, Microsoft blogged about adopting Marian to underpin the full range of the company's embedded translation tools [5.1], which include those in Outlook, MS Word and Office 365. That year alone, the applications employing Marian reached a combined user base of 380,000,000 [5.2, para. 15].

Marian's novel features – specifically the BPE and back-translation features – have enabled far more language pairs available for instant translation on the Microsoft platforms than previously possible, currently more than 70 [5.3]. Microsoft has publicly called Marian "one of the most efficient NMT toolkits available" [5.1, para. 21], directly citing four of the UoE papers on its blog, and elaborating on the specifics of how Marian has enhanced its applications. According to Microsoft, Marian is extremely efficient, not requiring GPUs during translation time, and is "very efficient at training time. Due to its self-contained nature, it is quite easy to optimize Marian for NMT specific tasks" [5.1, para. 20]. As evidence of the importance to Microsoft of its translation tools, for its Christmas 2019 advert, the company chose to focus on Microsoft Translator as a flagship product [5.4].

The uptake by Microsoft demonstrates the broad commercial reach of Marian, as it has been tailored for speedy use by general users. Significantly, Marian has also improved the translation capabilities for specialist translation companies. This is evidenced through its adoption by Lingo 24, and its targeted use at the UN's WIPO.

Lingo 24 is the UK's third-largest translation company, with an annual turnover of approximately [text removed for publication] [5.5, para. 1]. Since 2018, it has used Marian in bespoke translation services for a variety of clients, and confirms that "the switch from our previous systems to Neural MT using Marian has generated a marked increase in MT quality" [5.5, para. 4]. This improvement in the quality of MT has enabled the company to increase the percentage of revenue generated by MT-based work, from [text removed for publication] in 2018 to [text removed for publication] in 2019. Marian's high performance on CPUs – negating the need for more expensive GPU hardware – additionally saves the company money [5.5, para. 6].

Lingo 24 has been able to pass these benefits onto their various clients. In some cases clients whose services have been provided through Marian have "gained a market advantage…being able to release products more quickly" [5.5]. In the case of one distributor of home electronics, faster translation enabled the company "to reduce their time to market

… allowing them to grow international revenues from [text removed for publication] to [text removed for publication]" [5.5, para. 7]. Lingo 24 is also explicit in crediting to Marian increased orders from customers due to better performance and faster turnaround [5.5, para. 8].

In public service, the UN and the US Airforce are among the agencies to have adopted Marian. The US Airforce has credited "Marian's ease of use and high performance" in processing non-English video, transcriptions of speech, and translation of documents [5.6]. The UN uses Marian through its WIPO agency, an organisation tasked with protecting intellectual property across the UN's 192 member states. WIPO deals in complex patents, written in legalese, and uses technical vocabulary. It runs a database named PATENTSCOPE, which contains 91,000,000 patent records [5.7].

Marian enabled WIPO to build its own in-house translation tool, WIPO Translate, which has led to increased accuracy of patent translation, to a level that beats its commercial competition [5.8, paras. 2-4]. The Marian-based WIPO Translate is also used as the instant translator for PATENTSCOPE, allowing inventors and patent offices to translate previously filed patents from and to 14 language pairs, at rapid speed [5.9, p. 9].

Marian generates a more natural word order when translating between so-called 'distant' language pairs, such as English and Chinese. Currently, more than 50% of worldwide filings for patents are in Chinese, Japanese or Korean [5.8, para. 5].

In a 2016 article, WIPO's former director-general confirmed that WIPO Translate "outperforms any other technology for translating the complex language used in patents" [5.8, para. 1]. These claims are supported by the BLEU scores of WIPO Translate, which have topped Google Translate in at least 18 language pairs [5.10, p. 3]. This means that WIPO is not reliant on an external provider for quality translation services, but can instead rely on their in-house tool which was enabled by UoE research.

WIPO Translate has been so well-received at the UN that other agencies have adopted it too, including the World Trade Organisation. WTO's Director of Languages states:

> *WIPO Translate plays an important role supporting multilingualism at WTO, allowing Secretariat staff to quickly comprehend texts which cannot be translated by humans due to resource constraints. Thanks to the great progress made in the quality of machine output, WIPO Translate is also increasingly called on by translators as an important tool in their toolbox.* [5.11]

The benefits Marian has brought to machine translation are evidenced by the swiftness of the software's uptake by those in the translation field, and by the variety of translation uses to which it has been put. It has been adapted to fill a vast range of modern translation needs, from global public service organisations, to mainstream quick-translation apps, and specialist translation agencies. Its impact as a tool has been both significant and far-reaching.

5. Sources to corroborate the impact

5.1. Microsoft. (2019, June 17). Neural Machine Translation Enabling Human Parity Innovations In the Cloud. Retrieved March 17, 2020, from

https://www.microsoft.com/en-us/translator/blog/2019/06/17/neural-machine-translation-enabling-human-parity-innovations-in-the-cloud/
5.2. Nadella, S. (2019, October 16). Microsoft Annual Report 2019. Retrieved December 9, 2019, from https://www.microsoft.com/investor/reports/ar19/index.html
5.3. Microsoft. (2020). Microsoft Translator Languages - Microsoft Translator. Retrieved November 4, 2020, from https://www.microsoft.com/en-us/translator/languages/
5.4. Microsoft. (2019). Microsoft Holiday Ad 2019 – "Holiday Magic: Lucy & the Reindeer". Retrieved December 9, 2019, from https://www.microsoft.com/en-us/empowering
5.5. Letter of corroboration from Lingo24
5.6. Statement from United States Air Force Research Laboratory
5.7. WIPO. (2020). Search International and National Patent Collections. Retrieved November 5, 2020, from https://patentscope.wipo.int/search/en/search.jsf.
5.8. Cookson, C. (2016, October 31). Patent translator flies artificial intelligence flag for public sector. Retrieved September 23, 2019, from https://www.ft.com/content/d6be65f0-9e8d-11e6-891e-abe238dee8e2
5.9. WIPO. (2017). PATENTSCOPE. Retrieved March 18, 2020, from https://www.wipo.int/publications/en/details.jsp?id=4203
5.10. Pouliquen, B (2017, October 25). WIPO Translate Updates, Confederacy of European Patent Information User Groups. Retrieved 3 November, 2020 from http://cepiug.org/public/index.php?page=WIPO_round_Table_2017
5.11. WIPO. (2020). WIPO Translate. Retrieved November 5, 2020, from https://www.wipo.int/wipo-translate/en/