View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by DCU Online Research Access Service

Chapter 3.3: Machine Translation: where are we at today?

Andy Way

I. Introduction

Machine translation (MT) usage today is staggering. Consider Google Translate,[1] which as of May 2016 was translating an average of 143 billion words a day – 20 words/day for every person on the planet, just for a *single* (albeit the largest) MT service provider – across 100 language combinations, a doubling in translation volume in just 4 years. This number alone already means that MT quality is 'good enough' for a range of use-cases, so continuing to question the utility of MT is moot.

The aim of this chapter is to explain to translation/interpreting students and academics, professional translators and other industry stakeholders how MT works today, and how the field has altered in the last 30 years. I describe the underlying reasons why MT engine-building changed from being underpinned by grammatical rules to the situation today where it is almost entirely data-driven; while for some time most of the research in academia was corpus-based, the leading MT engines in industry remained almost wholly rule-based, but this dichotomy has now largely disappeared, principally due to the introduction of the Moses Statistical MT (SMT) toolkit (Koehn et al., 2007), and the subsequent rise of Neural MT (NMT).

While it was already the case that the dominant paradigm was SMT, a performance ceiling was reached relatively quickly, such that for the past ten years or so, MT system developers have been 'smuggling in' linguistic information in order to improve performance as demonstrated by both automatic and human evaluation. Until just three or four years ago, SMT was undoubtedly state-of-the-art, but NMT has recently emerged, and in academic circles at least, appears to be so promising that many protagonists are already claiming it to have surpassed the performance of SMT. In this chapter, I will consider the extent to which it is appropriate at this juncture to make this call; SMT remains dominant in the translation industry among many translation providers, but the big players like Google Translate and Bing Translator[2] have already launched NMT systems for many of their language pairs.

When SMT was launched, many practitioners advocated a 'pure' approach, where the strategy taken was "let the data decide"; no data cleaning or annotation was countenanced, at least initially, so that whatever quality was obtained was due entirely to the intrinsic characteristics of the approach rather than any pre-processing techniques. Nonetheless, as mentioned above, SMT system developers observed improvements in performance as measured by automatic evaluation metrics when introducing linguistic information into the engine-building process. With the advent of NMT, similarly 'pure' approaches to NMT are in vogue, but I question whether here too quality will improve only if syntactic, semantic and discourse features are integrated.

It is clear that human translators have for some time now been using Translation Memory (TM) systems (Heyn, 1998) to good effect. Many researchers have demonstrated that SMT and TM can be integrated to improve translator productivity (e.g. Ma et al., 2011; Bulté et al., 2018), and these benefits now appear fairly regularly in today's industry-leading CAT tools. TM integration has yet to be done for NMT, and given that I expect TM technology to remain as an essential tool in the translator's armoury for some time to come, I will consider how such integration might be brought about.

---

[1] https://translate.google.com
[2] https://www.bing.com/translator

I will also discuss how MT quality is measured, the extent to which 'traditional' MT evaluation is equipped to demonstrate improvements delivered by NMT, what human evaluations are currently adding to the mix, and how emerging use-cases where there is no place for human translators cause us to fundamentally question the notion of quality.

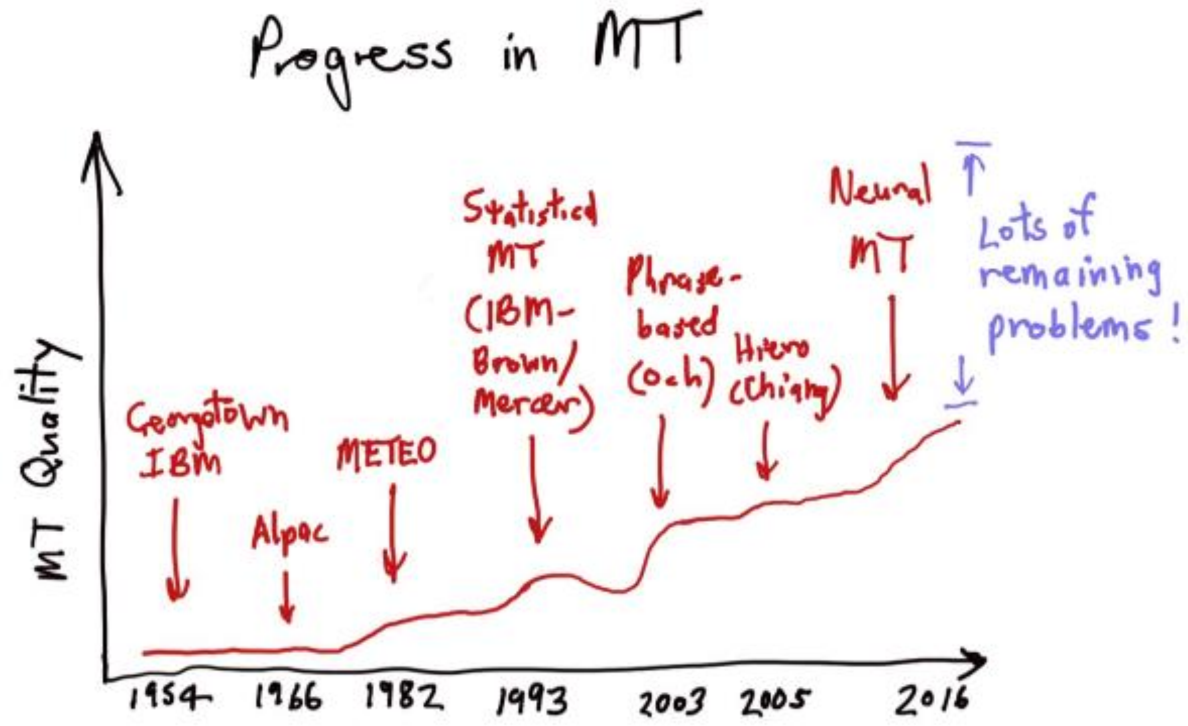


Figure 1: Progress in Machine Translation over the years (Luong et al., 2016)

II. The Rise and Fall of Different MT Paradigms

In this section, I provide a brief history of the major paradigms that have been put forward in MT. As Figure 1 illustrates schematically, different MT models have been in vogue at different times. Early MT systems were entirely rule-based, but in the 1980s corpus-based models came along and became the state-of-the-art by the mid-1990s. In just the last few years, the advent of NMT has really shaken up both academia and the wider MT and translation industry, and now appears to have taken over the mantle from SMT as the dominant paradigm today. In the next sections, I describe briefly the mechanics of each of these system types, as well as how the field reacted when they were suggested as competing paradigms to the dominant approaches of the day.

II.1 From Rule-based MT to SMT

As I describe in Way (2009), when Peter Brown of IBM (at the time) stood up at TMI in Pittsburgh and again at COLING in Budapest in 1988 and presented SMT as an alternative to rule-based translation, significant players in traditional approaches to MT were astonished. Pierre Isabelle's reaction was "*We were all flabbergasted. All throughout Peter's* *presentation, people were shaking their heads and spurting grunts of disbelief or even of hostility*.” Harold Somers noted “*The audience reaction was either incredulous, dismissive or hostile*”, while Walter Daelemans observed “*the Leuven Eurotra people weren’t very impressed by the talk and laughed it away as a rehash of ‘direct’ (word-by-word) translation*”.

Prior to Brown et al. (1988a/b), rule-based MT (RBMT) was divided into two camps: transfer-based MT and interlingual MT. The Vauquois Pyramid (see Figure 2) visualises schematically what was involved in building such systems quite succinctly, with the length of each arrow corresponding to roughly the amount of work needed by each component.

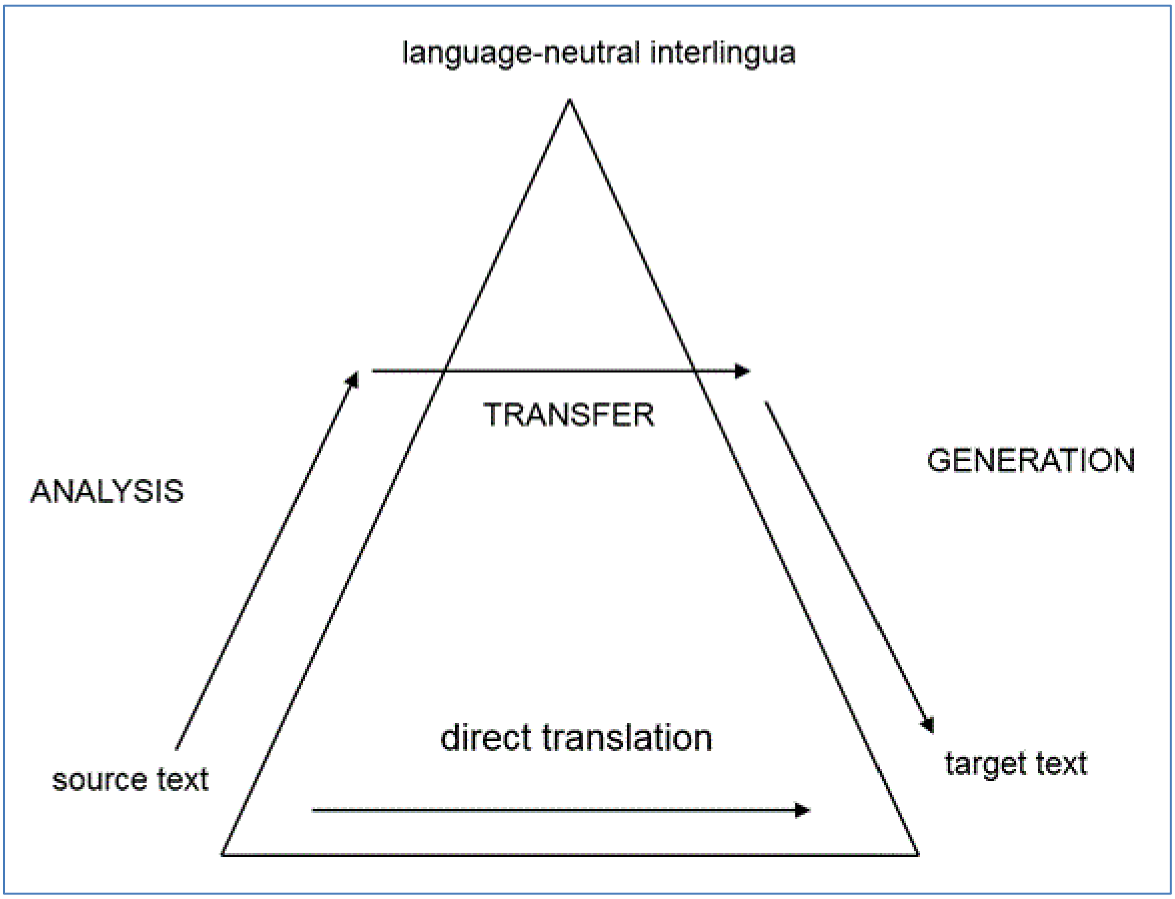


Figure 2: the Vauquois Pyramid depicting the three main approaches to RBMT: direct translation, transfer-based and interlingual MT (Vauquois, 1968)

Transfer and interlingual systems were both known as indirect, second-generation approaches to MT, and were compared to direct, first-generation MT systems. As can be seen in Figure 1, these latter did very little analysis (‘parsing’) of the source language or structural generation of the target language; that they worked at all was down to their very large bilingual dictionaries. Somewhat surprisingly, these systems enjoyed a relatively long shelf life, partly due to the fact that they were very robust, and compared to indirect systems, always produced *some* output, which between ‘similar’ languages (e.g. Portuguese and Spanish) could often be very reasonable indeed. In contrast, given that indirect systems depended on parsing (to different depths) the source-language input, they were explicitly designed to rule out ill-formed input; when I worked on Eurotra (King & Perschke, 1984) between 1988-1991, we wrote explicit test suites (Arnold et al., 1993) containing well-formed sentences that the analysis component ought to parse correctly and pass on an appropriate representation to the transfer stage, as well as ill-formed strings that the analysis stage should decree ungrammatical and cause further processing to cease. There were two main problems with such an approach: (i) there was a general assumption that people would always try to input well-formed sentences into an MT system, and (ii) given that the parser was based on a set of handcrafted rules by an expert linguist which was necessarily incomplete, the system could not tell the difference between a truly ill-formed string and well-formed input that simply was not covered by the set of linguistic rules in the grammar.

Note that in transfer-based systems, the three processes do ‘about the same’ amount of work: the source string is parsed into a syntactic (constituency or dependency) tree indicating the main actors in the sentence as well as any modifiers; this source-language representation is then passed to the transfer component *per se*, where appropriate lexical, syntactic and semantic rules generated a ‘meaning-equivalent’ target-language structure; this target dependency tree is then input into the generation (or ‘synthesis’) phase, where a set of target-language rules try to produce an appropriate translation.

In contrast, in interlingual systems, there was *no* explicit transfer phase, so that the output from the deep analysis phase was *exactly* the same as the input to the deep synthesis phase. While this was very attractive in theory, it proved impossible to bring about from a practical perspective. Languages simply don’t act the same way, with different languages having different ways of representing similar concepts. For example, in English, the periphrastic expression “to bake with cheese on top” has to be used, while French has a single lexical item – *gratiner* – to represent the same concept. Assuming an interlingual system involving French, English and Japanese, the amount of work that would need to be done in (say) a French-to-English engine just because Japanese has different words for “my mother”, “your mother” and “mothers in general” would be wasteful, given that neither French nor English has different lexical entries for these concepts (Hutchins & Somers, 1992).

At the onset of SMT in the late-80s, it was clear what camp you were in: either the transfer camp or the interlingual camp. However, in Way (2009), I noted that pretty quickly, these two camps merged to form a *de facto* alliance against this arrogant statistical newcomer which was set to undo all that they stood for. Despite this resistance, the language used early on by the new statistical practitioners was conciliatory, indicating a hope that the two communities would work together for the betterment of the discipline. I lamented that this did not happen, and that this impaired the creation and adoption of the syntax-based systems that came onstream in the late 2000s.

Nonetheless, certainly by the mid-1990s, SMT had come to be dominant, largely due to the very influential IBM models laid out in Brown et al. (1993), one of the seminal papers in the field. At this time, however, most SMT was word-based, which was odd when one considers that example-based MT (Nagao, 1984; Carl & Way, 2003) had from its very inception considered the phrase – not the word – as the primary linguistic construct to be used as the unit of translation. Koehn et al. (2003) demonstrated how SMT might work in a phrase-based manner, and with the advent of tools like Giza++ (Och & Ney, 2003, for word alignment) and the Moses toolkit (including phrase alignment) in 2007, phrase-based SMT (PBMT) became the dominant paradigm for the next ten years.

Larger and larger amounts of SMT training data came onstream (e.g. Europarl: Koehn, 2005), which allowed better and better PBMT models to be built, but only for those languages and genres where sufficiently large sets of aligned source—target sentences existed. The licence issued with Moses allowed it to be used commercially, so SMT systems were quickly deployed to good effect in real industrial scenarios.[3] SMT was robust, and capable of very good translation output, but suffered from problems such as the omission of target-language equivalents to parts of the source sentence (including on occasion really important words like *not*), and wrong target-language word order. As with RBMT, PBMT worked especially well between closely-related languages, and much less well when translating into morphologically complex languages (like German).

It was certainly the case when SMT first came along that most system developers relied solely on larger and larger amounts of training data to deliver improvements in translation quality, as measured by automatic metrics like BLEU (Papineni et al., 2002), METEOR (Banerjee & Lavie, 2005) and TER (Snover et al., 2006). However, little by little, SMT engine-builders began to realise that the only way to break through the performance ceiling –often a pretty good level of quality, mind – was to integrate additional syntactic and semantic information (e.g. Chiang, 2005). By 2015 or so, such linguistically informed PBMT systems were acknowledged to be the state-of-the-art in the field.

II.2 From SMT to NMT

However, around this time, researchers (including many newcomers to the field of MT) started to demonstrate that NMT systems could be built with good performance. While the preferred system set-ups were not so different from those that had been conceived some time before (e.g. Forcada & Ñeco, 1997), the hardware that facilitated the huge explosion in computation required was now sufficiently powerful to allow these systems to be built in practice.

The first NMT systems started off using convolutional neural nets (Kalchbrenner & Blunsom, 2013), but could not beat a PBMT baseline (cdec: Dyer et al, 2010). Improvements were seen quite quickly with the first encoder-decoder frameworks (Sutskever et al., 2014), which were subsequently extended with a source-language attention model (Bahdanau et al., 2014). While further improvements have been seen in the interim, this set-up – an encoder-decoder model with attention – remains pretty much the state-of-the-art today.

What really disrupted the field were the results achieved by NMT systems at the International Workshop on Spoken Language Translation in 2015.[4] Luong and Manning (2015) demonstrated clear wins over a range of different SMT systems for English-to-German, a significantly difficult language pair, in terms of automatic evaluation scores. Bentivogli et al. (2016) performed an in-depth human evaluation of exactly how the NMT model of Luong and Manning (2015) improved in terms of quality, noting that significantly fewer morphological, lexical and word order errors were made compared to SMT. They also demonstrated that NMT lowered overall post-editing effort by about 25%.

One of the main reasons why NMT improves compared to SMT on a range of use-cases is that once the source sentence has been processed by the encoder, the *full* context of the sentence is available to the decoder for consideration as to what target-language words and phrases should be suggested as part of the translation. That is, all source words and their context – what are known as "word embeddings" (cf. Mikolov et al., 2013), i.e. how each word relates to each other in the particular sentence at hand – are encoded in a *single* numerical representation (a vector of numbers indicating the final state of the encoder) which is sent to the decoder to generate a target-language string. In SMT, a source sentence is only translated using lexical and phrasal chunks; unless it is very short, it is never translated *en bloc*. Clearly having a window on the full source sentence is advantageous compared to a restriction of just a few words at a time, but managing all that information is non-trivial.

---

[3] The first commercial system, LanguageWeaver, was based not on Moses but rather on the SMT models of Kevin Knight and Dan Marcu at ISI (cf. Benjamin et al., 2003).

[4] http://workshop2015.iwslt.org/

The encoder-decoder architecture works well, but significant improvement came about when the source-language attention model was added. Rather than accepting that all source words are equally important in suggesting all target-language words, the attention model (similar to word and phrase alignments in SMT) demonstrates which source words are most relevant when it comes to hypothesising target-language equivalents. In practice, this means that each translation is generated from specific encoder states, with information which is much less relevant from other words – perhaps some distance away from the current word of focus and of little or no relevance to its translation – being ignored.

III. Is NMT the new state-of-the-art?

While the study by Bentivogli et al. (2016) was significant and far-reaching, it has to be noted that it only examined one language pair (English-to-German) and one use-case (TED talks). Further studies (e.g. Castilho et al. (2017)) have shown that there are situations where PBMT can still beat NMT in terms of both human and automatic evaluation. It is widely recognized that much larger amounts of training data are needed for good NMT performance compared to SMT (cf. Koehn & Knowles, 2017), and training and translation times remain much slower than for SMT.

Nonetheless, many MT practitioners believe that NMT is – or at least will be very soon – the new state-of-the-art, to the extent that all MT papers in the very top academic conferences feature only NMT models, and Moses scores are only given as comparative baseline levels of quality.

III.1 How is MT Quality Measured?

In Way (2018), I note that there are three ways in which MT quality is typically measured: via human evaluation, automatic evaluation, and task-based evaluation. In the former, human raters are asked to select from a (more or less) fine-grained numerical scale for “fidelity” (or “accuracy” or “adequacy”), the extent to which a translated text contains the same information as the source text; and (ii) intelligibility (or “fluency”), the extent to which the output sentence is a well-formed example of the target language.[5] While such evaluations are (usually) very informative, they are subjective, often inconsistent and take a long time to carry out. Accordingly, as is often the case in MT, insights from speech recognition were brought to bear in this field too, in particular Word Error Rate (WER: Levenshtein, 1966), and Position-Independent Word Error Rate (PER: Tillmann et al., 1997). However, it wasn’t until the BLEU metric came in that MT evaluation *per se* took off. BLEU (and NIST (Doddington, 2002), which came along around the same time) used different (but related) ways to compute the similarity between one or more human supplied ‘gold standard’ references and the MT output string based (largely) on *n*-gram co-occurrence.

---

[5] Here I give primacy to the terms originally used in the ALPAC report (Pierce et al., 1966), with more usual terms given in parentheses.

In Way (2018), I provide a number of problems with such metrics, as well as others arising from their (mis)use in the field. I will not rehash those here, but ultimately an MT system needs to be used for a particular use-case, which is where task-based evaluation comes in: who is the translation actually *for*? As I point out in that paper:

> "*WMT evaluations regularly include specific tasks nowadays, including medical translation (e.g. Zhang et al., 2014), automatic post-editing (e.g. Chatterjee et al., 2015) and MT for the IT domain (e.g. Cuong et al., 2016). We take this as evidence that the community as a whole is well aware of the fact that when evaluating MT quality, the actual use-case and utility of the translations therein need to be borne in mind*."

III.2 Does MT Evaluation need to change with NMT coming onstream?

It has to be acknowledged that the problem of MT quality assessment is an unsolved one, and research efforts are ongoing to improve on the metrics that are commonly used today. One question that is worth asking is the extent to which such metrics are sufficiently discriminative to accurately demonstrate the real improvement that NMT offers over SMT.

The translational improvements discovered by Bentivogli et al. (2016) are astonishing, especially bearing in mind that PBMT had been the dominant paradigm for 25 years or so, and that NMT has only come in as a realistic alternative in the past four years. In my opinion, *n*-gram-based metrics such as BLEU significantly underplay the real benefit to be seen when NMT output is evaluated. As I note in Way (2018), it simply *cannot* be the case that a 2-point improvement in BLEU score – almost an irrelevance on a real industrial translation use-case – which was typically seen in WMT-2016 where NMT systems swept the board on all tasks and language-pairs (Sennrich et al., 2016), can be reflective of the improvements in word order and lexical selection noted by Bentivogli et al. (2016). Note that Shterionov et al. (2018) actually computed the degree of underestimation in quality of three popular automatic evaluation metrics – BLEU, METEOR and TER – showing that for NMT, this may be up to 50%.

Metrics such as ChrF (Popović, 2015) which operate at the character level – or combinations of word- and character-based models (e.g. Chung et al. 2016; Luong & Manning, 2016)[6] – may be a move in the right direction, but the field will doubtless see new metrics tuned particularly to NMT in the very near future.

III.3 Is the Translation Industry ready to provide NMT?

Let us suppose that NMT either already is the new state-of-the-art in terms of MT quality, or very soon will be. The big free online players – Google Translate and Bing Translator – have both switched over at least some of their language-pairs to NMT models. Note that Amazon AWS only very recently announced their own NMT service,[7] so there is no doubt that where the largest multinational companies are concerned, the decision has been made to throw in their lot with NMT.

---

[6] In order to mitigate the problem of unknown words, character-based NMT models were proposed; if a word is unknown at the level of the lemma, some translation knowledge may be available at the subword level. Passban et al. (2018) demonstrate that splitting lemmas into roots and morphemes in a principled linguistic manner outperforms such arbitrary subword models.

[7] https://aws.amazon.com/translate/

Accordingly, those language service providers (LSPs) who rely on online MT provided by third-parties such as these will already have benefited from the improvements in quality afforded by NMT. But what about those MT providers who have developed services in-house around the Moses platform? I have already noted that neural MT engine training times are much slower – typically of the order of several weeks – than their SMT counterparts, so much so that people are claiming PBMT training times to be fast nowadays, although of course nothing has changed in that regard; it is simply the case that in comparison, NMT model training is incredibly slow, with billions of mathematical optimisations needed until the neural net converges to its optimal set-up. I have also noted that typically an order of magnitude larger data is needed to train a good NMT model compared to PBMT, and it is a fact that these datasets do not exist for almost all industry clients. In addition, the hardware needed to train an NMT system is expensive; GPUs contain thousands of CPUs, each of which can carry out its own calculation in parallel. Assuming most suppliers of customised MT engines do not have such hardware in-house, but rather rely on cloud-based services, the cost of additional MT engine training will have to be passed on to clients, although the latter should see most if not all of this returned by the huge improvements in MT quality and resultant decrease in post-editing effort required.

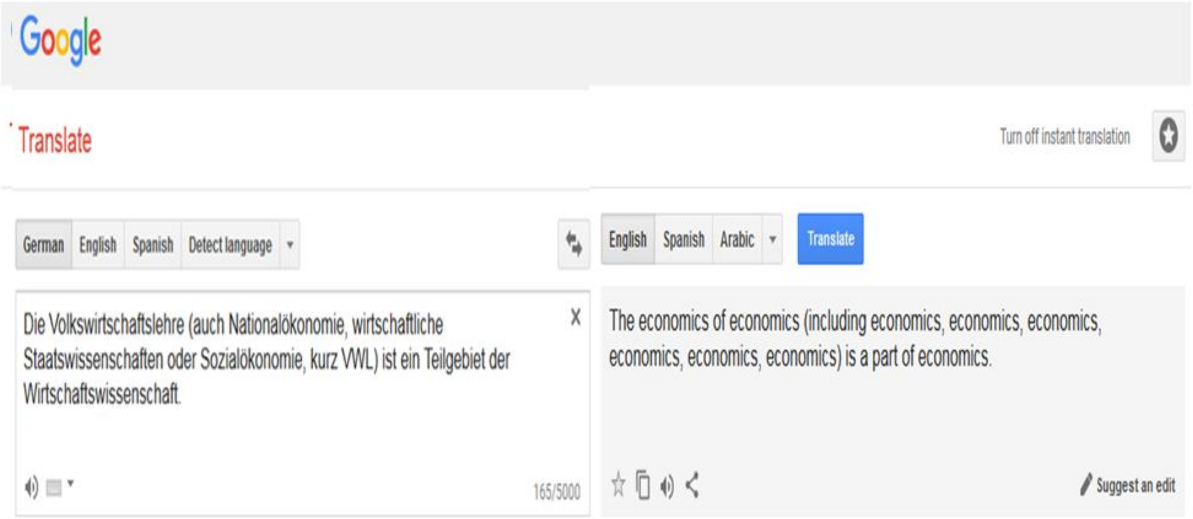
Those forward-thinking translators who have already integrated MT into their pipeline should benefit immediately from the improvements in MT quality to be seen. As I noted in the previous section, current MT quality assessment metrics are insufficiently discriminative to provide a realistic representation of the absolute improvement in quality seen with NMT, so it is open to doubt as to whether LSPs will be able to reflect this better quality in terms of higher levels of TM fuzzy matches (Sikes, 2007), with the concomitant reductions in pay to translators who are post-editing MT output. It seems to me that this is a good time for translators who have yet to use MT in their translation workflow to consider doing so without delay, as their productivity should rise pretty quickly, while LSPs are still tied in to post-editing rates of pay related to SMT.

MT has been integrated very well now with existing TM tools, with TM matches above well-defined thresholds being suggested to translators for post-editing, and MT used for all segments below such thresholds.[8] NMT shouldn't make too much of a difference, here, except that even more so than SMT, NMT output can be deceptively fluent; sometimes perfect target-language sentences are output, and less thorough translators and proofreaders may be seduced into accepting such translations, despite the fact that they may not be related to the source sentence at hand at all! In contrast, when the attention model provides too much focus on particular source-language words, errors such as that in Figure 3 can be seen; fortunately, these are easy to spot.

Finally, on the subject of quality, for a long time SMT models failed to deliver good enough quality for English-to-Japanese for Japanese translators to even consider post-editing MT output. Mike Dillinger (personal communication) now informs me that the quality seen from NMT is leading them to reconsider, and the sorts of questions being asked are *exactly* the same as those raised years ago in the scope of European languages.

---

[8] Moorkens and Way (2016) discuss the extent to which translation jobs should be carved up in this way, as well as how MT output is significantly preferred to TM matches when fuzzy match thresholding is removed.



Die Volkswirtschaftslehre (auch Nationalökonomie, wirtschaftliche Staatswissenschaften oder Sozialökonomie, kurz VWL) ist ein Teilgebiet der Wirtschaftswissenschaft.

The economics of economics (including economics, economics, economics, economics, economics, economics) is a part of economics.

Figure 3: Google NMT error due to an overly attentive Attention Model (18th July 2017)

IV. Informing research through the industry

As I stated at the outset, MT has never been as popular, and pretty much everyone in the industry knows that they have to embrace it as an enabling technology. Rather than outsource their MT requirements, companies like Google, Microsoft, Facebook, eBay and Amazon have (understandably) been recruiting leading academics to build their own internal MT offerings for some time now.

However, the result is that the relatively few MT centres of excellence that existed five years ago have become even rarer. I also took three years leave-of-absence to build industry-leading customised MT engines for two translation companies in the UK, but decided to return to my academic position to keep the MT team at my university together. One knock-on benefit to academic MT teams like my own is that excellent staff can be recruited from disbanded academic MT teams. However, at the same time, professionals with artificial intelligence (AI) and machine learning (ML) skills are highly prized, and the discrepancy between rates of pay which has always been there in academia and industry is widening at a rate of knots.[9]

Despite taking on leaders of academic teams into their companies, industry leaders are in the same breath bemoaning the fact that they are unable to recruit MT developers, as there are not enough trained experts coming from academic programmes to fill all the vacancies currently available. But they can't have it both ways: if they recruit the leaders of large, renowned academic groups, who used to train the MT developers of tomorrow, they shouldn't be surprised when the number of such potential recruits falls away. What MT academics want, therefore, is for industry to petition government to obtain more support for MT, AI, and ML in academia, so that the industrial community can be served to our mutual advantage.[10]

Note, however, that this cannot all be centrally funded by government. While there is no doubt that attracting research hubs of multinational companies pays off considerably – not just in terms of direct employment and return to the exchequer but also as it pertains to ancillary services – if those companies want a steady stream of suitably equipped new staff with up-to-date skill-sets trained by the best available lecturing staff, then they too will need to (at least part-)pay towards the tools and services required for their education.

---

[9] https://www.nytimes.com/2017/10/22/technology/artificial-intelligence-experts-salaries.html

[10] This is starting to happen in my own country, Ireland: https://irishtechnews.ie/irelands-first-industry-driven-masters-in-artificial-intelligence-is-launched/

As this problem resolves itself, further and deeper collaboration between industry and research is likely to be seen. More and more researchers are interested not just in an academic publication, but also in solving real problems of benefit not just to industry, but also relevant to their fellow academics. While many authors of papers at leading conferences in the field seldom consider potential end-users, it was recently announced that from 2018, the North American branch of the Association for Computational Linguistics (NAACL) will feature an industry track, focusing on disseminating results which apply cutting-edge research to real world problems. While plenty of such work exists already (e.g. Wang et al., 2016; Calixto et al., 2017), anything which explicitly gets more researchers to try to focus on truly impactful endeavours as opposed to being strictly of academic value is to be welcomed.

V. Informing the industry through research

There are differing views on whether users of a technology need to know the principles on which it is founded in order to (i) understand how the outputs are formed, and (ii) try to improve the underlying technology. Assuming that knowing how an MT system is built is useful, there is no doubt that non-experts found the principles of SMT hard to understand. In two companion papers (Hearne & Way, 2011; Way & Hearne, 2011), we provided an explanation of SMT for linguists and translators which attracted positive feedback.

Accordingly, translators that have already embraced MT have just about gotten their heads around SMT and how it works, but now NMT looks like eclipsing that framework. While PBMT quickly consolidated around the Moses toolkit, in contrast there is a proliferation of deep neural net tools in existence which NMT developers can use, including Tensorflow,[11] OpenNMT,[12] PyTorch,[13] and Nematus.[14]

Again, unsurprisingly, many non-experts – even those who have been around the language industry for some time – find recent research papers on NMT unintelligible. I have already provided a high-level explanation of how an NMT system works in Section II.2, and I hope that my description of the encoder-decoder system with attention is understandable to a broad audience; interested readers should consult Forcada (2017) for another explanation of NMT for non-experts.

---

[11] https://www.tensorflow.org/

[12] http://opennmt.net

[13] https://pytorch.org/

[14] https://github.com/rsennrich/nematus

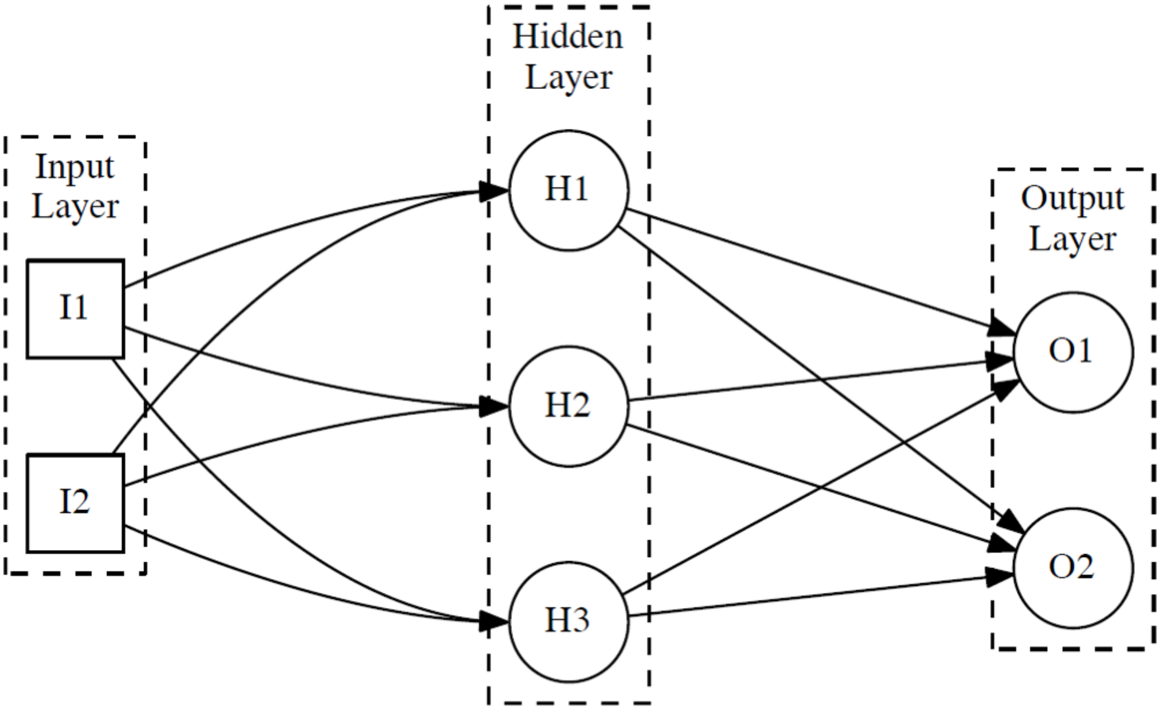


Figure 4: a schematic depiction of a feed-forward neural network

NMT is just one example of sequence-to-sequence learning using a neural network; others include text summarization and speech recognition. Essentially the neural network comprises (i) a set of input nodes, (ii) a set of hidden nodes (in one or more hidden layers; if there is more than one, the network is said to be ‘deep’, hence “deep learning”), and (iii) a set of output nodes, as in Figure 4. Each input node is connected to each hidden unit, and each hidden unit is connected to each output node; if there is more than one hidden layer, then each node in one hidden layer is connected to all the nodes in the next hidden layer, and so on. The mathematical complexity of deep learning comes about as the weight (or importance) of every connection between every node needs to be optimised. One way of doing this in the “training” phase is to (i) initially assign random weights to each connection,[15] (ii) supply a set of inputs where the outputs are known (e.g. if a vectoral representation of an image of a bird is input, the neural net can be expected to predict the label “bird”, not “dog”)), and (iii) percolate (via “back-propagation”: Rumelhart et al. (1986)) any errors back through the network until the weights are optimally adjusted and no further improvement in accuracy can be seen. The neural net is then frozen, and new inputs are provided to the system (in the “testing” phase) and the accuracy is evaluated by observing how often the neural net accurately predicts the correct label.

Most people acknowledge that MT is one of the hardest problems we are trying to address in computer science, as so many inputs (each word in each sentence) are required, many hidden units in many hidden layers are needed, and many outputs (i.e. possible translations) may be observed. Note too that unlike the feed-forward neural network in Figure 4, most state-of-the-art networks are “recurrent”, meaning that some units are linked to themselves; this permits inputs and outputs of any size, whereas feed-forward networks allow only fixedlength inputs and outputs. Accordingly, it can be seen quite quickly that the number of calculations is astronomical. Billions of tweaks of the weights are needed before the optimal configuration of the neural net is arrived at, and no further improvement can be seen. This is why GPUs – excellent at performing calculations on matrices, which are standard structures used in neural processing – are needed for network training, although decoding can be run fairly smoothly on CPUs.

---

[15] Including to “bias” nodes, which are connected to each hidden unit to prevent that hidden unit from being ‘switched off’ in case of a zero-sum input.

The mathematical underpinning of neural networks is fairly hair-raising, but I anticipate that the description of neural nets provided together with my description of the state-of-the-art NMT model today and the explanation provided in Forcada (2017) will suffice for most translators to know how the whole set-up works, and how they may help deep-learning engineers improve their systems. At the same time, I trust that this will prove useful to other industry players not *au fait* with the current technology, in order for them to consider using NMT to underpin their translation services.

VI. Concluding Remarks

This paper has undertaken several challenges: (i) to discuss different MT paradigms, (ii) to argue which of those paradigms might be considered state-of-the-art today, (iii) to explain to non-experts how neural MT works, (iv) to discuss whether today's automatic evaluation metrics are sufficiently fine-grained to accurately reflect the dramatic improvements we have seen recently in terms of MT quality, and (v) to reflect on the relationship between academia and industry in the field of MT today.

We concluded that if NMT is not already the state-of-the-art in the field, it certainly has the potential to become so, and very soon. Accordingly, I trust that the description of the underlying deep-learning technology and the state-of-the-art NMT configuration may benefit a wide range of non-experts, who might be struggling to understand how this new paradigm actually works, why it outperforms PBMT, but also what problems remain to be solved. These can be examined both from a research perspective – including providing fine-grained MT evaluation metrics to accurately reflect the considerable improvement in MT quality that has recently been seen – and an optimisation point of view, especially in terms of improving NMT engine training times.

By providing an insight into how academia and industry need to help each other in these turbulent times, I anticipate that this will contribute to the building of stronger bridges between academic research and the language industry. As we concluded in Way & Hearne (2011), this collaboration is sorely needed if the field as a whole is to benefit. While SMT appears to have only a limited future, with NMT having emerged as the dominant force in MT, such collaboration is needed more than ever. It is encouraging to see that the lesson learned by SMT practitioners regarding the improvements to be seen by incorporating linguistic information seem to be being taken onboard, by some NMT practitioners at least (cf. García-Martínez et al., 2016; Sennrich & Haddow, 2016).

With the considerable improvements in MT quality that have been seen in recent times has come an increase in hype, most notably from journalists, most of whom don't understand how the technology works, but also from MT developers such as the Google and Microsoft NMT teams; the claim by Wu et al. (2016) that Google NMT was "bridging the gap between human and machine translation [quality]" led to considerable hyperbole and hysteria from different camps, which was amplified recently by the claim by Hassan et al. (2018) that

Microsoft had “achieved human parity” in terms of translation quality.[16] Those of us who have seen many paradigms come and go know that overgilding the lily does none of us any good, especially those of us who have been trying to build bridges between MT developers and the translation community for many years. The human-in-the-loop will *always* remain the most important link in the chain, at least where dissemination of translated material is required; all MT system developers are trying to do is improve the output from their systems to make technology-savvy translators more productive. MT systems are unlikely *ever* to “bridge the gap” or “achieve human parity” with human quality translation. Just because a new paradigm is in vogue does not mean that MT has become easy, or a solved problem, as some would like to make out (e.g. Goodfellow et al., 2016:473). Let’s see how many of the newcomers to MT are still here in a decade; my prediction is that a good percentage of them will indeed discover that MT is too difficult, and that certain problems remain hard to solve, just like they’ve always been … which is why translators are very much still needed, and always will be.

Acknowledgements

Many thanks indeed to the anonymous reviewers of this paper, as well as to Mikel Forcada, whose comments served to improve this paper considerably. This work has been supported by the ADAPT Centre for Digital Content Technology which is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

References

Arnold, D., D. Moffat, L. Sadler, and A. Way. (1993), Automatic Generation of Test Suites. *Machine Translation* **8**:29—38.

Bahdanau, D., K. Cho, and Y. Bengio. (2014), Neural Machine Translation by Jointly Learning to Align and Translate. eprint arXiv:1409.0473 (https://arXiv.org/abs/1409.0473).

Banerjee, S., and A. Lavie. (2005), METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *ACL 2005, Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43rd Annual Meeting of the Association for Computational Linguistics*, Ann Arbor, Michigan, pp.65—72.

Benjamin, B., L. Gerber, K. Knight, and D. Marcu. (2003), Language Weaver: the next generation of machine translation. *MT Summit IX*, New Orleans, USA, pp.229—231.

Bentivogli, L., A. Bisazza, M. Cettolo, and M. Federico. (2016), Neural versus Phrase-Based Machine Translation Quality: a Case Study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, Texas, pp.257—267.

---

[16] Even more recently, SDL announced on June 19th 2018 that they had “cracked Russian to English Neural Machine Translation”: https://www.sdl.com/about/news-media/press/2018/sdl-cracks-russian-to-english-neural-machine-translation.html. As with the other similar claims, this has met with some incredulity on social media platforms, including by this author (@tarfandy).

Brown, P., J. Cocke, S. Della Pietra, V. Della Pietra, F. Jelinek, R. Mercer, and P. Roossin. (1988a), A statistical approach to French/English translation. In *Second International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*, Pittsburgh, Pennsylvania, 16pp.

Brown, P., J. Cocke, S. Della Pietra, V. Della Pietra, F. Jelinek, R. Mercer, and P. Roossin. (1988b), A statistical approach to language translation. In *Coling Budapest: Proceedings of the 12th International Conference on Computational Linguistics* (vol.1), Budapest, Hungary, pp.71—76.

Brown, P., S. Della Pietra, V. Della Pietra, and R. Mercer. (1993), The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics* **19**(2): 263—311.

Bulté, B., T. Vanallemeersch, and V. Vindeghinste. (2018), M3TRA: integrating TM and MT for professional translators. In *EAMT2018, Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, Alacant, Spain, pp.69—78.

Calixto, I, Q. Liu, Qun, and N. Campbell. (2017), Doubly-Attentive Decoder for Multi-modal Neural Machine Translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vancouver, Canada, pp.1913—1924.

Carl, M., and A. Way (eds). (2003), *Recent Advances in Example-Based Machine Translation*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Castilho, S., J. Moorkens, F. Gaspari, I. Calixto, J. Tinsley, and A. Way. (2017), Is Neural Machine Translation the New State-of-the-Art? *Prague Bulletin of Mathematical Linguistics* **108**:109—120.

Chatterjee, R., M. Turchi, and M. Negri. (2015), The FBK Participation in the WMT15 Automatic Post-editing Shared Task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, Lisbon, Portugal, pp.210—215.

Chiang, D. (2005), A hierarchical phrase-based model for statistical machine translation. In *ACL-2005: 43rd Annual meeting of the Association for Computational Linguistics*, Ann Arbor, MI, pp.263—270.

Chung, J., K. Cho, and Y. Bengio. (2016), A character-level decoder without explicit segmentation for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany, pages 1693—1703.

Cuong, H., S. Frank, and K. Sima'an. (2016), ILLC-UvA Adaptation System (Scorpio) at WMT'16 IT-DOMAIN Task. In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany, pp.423—427.

Doddington, G. (2002), Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *HLT 2002: Human Language Technology Conference*: *proceedings of the second international conference on human language technology research*, San Diego, California, pp.138—145.

Dyer, C., A. Lopez, J. Ganitkevitch, J. Weese, F. Ture, P. Blunsom, H. Sewatian, V. Eidelman, and P. Resnik. (2010), cdec: a decoder, alignment, and learning framework for finitestate and context-free translation models. In *Proceedings of the ACL 2010 System Demonstrations*, Uppsala, Sweden, pp.7—12.

Forcada, M. (2017), Making sense of neural machine translation. *Translation Spaces* **6**(2):291—309.

Forcada, M., and R. Ñeco. (1997), Recursive hetero-associative memories for translation. In *Biological and Artificial Computation: From Neuroscience to Technology (International Work-Conference on Artificial and Natural Neural Networks, IWANN'97, Proceedings),* Lanzarote, Canary Islands, Spain, pp.453—462.

García-Martínez, M., L. Barrault, and F. Bougares. (2016), Factored Neural Machine Translation. eprint arXiv:1609.04621 (https://arxiv.org/abs/1609.04621)

Goodfellow, I., Y. Bengio, and A. Courville. (2016), *Deep Learning*. Cambridge, MA.: MIT Press.

Hassan, H., A. Aue, C. Chen, V. Chowdhary, J. Clark, C. Federmann, X. Huang, M. Junczys-Dowmunt, W. Lewis, M. Li, S. Liu, T-Y. Liu, R. Luo, A. Menezes, T. Qin, F. Seide, X. Tan, F. Tian, L. Wu, S. Wu, Y. Xia, D. Zhang, Z. Zhang, and M. Zhou. (2018), Achieving Human Parity on Automatic Chinese to English News Translation. eprint arXiv: 1803.05567 (https://arxiv.org/abs/1803.05567).

Hearne, M., and A. Way. (2011), Statistical Machine Translation: A Guide for Linguists and Translators. *Language and Linguistics Compass* **5**:205—226.

Heyn, M. (1998), Translation Memories – Insights & Prospects. In L. Bowker, M. Cronin, D. Kenny and J. Pearson (Eds.) *Unity in Diversity? Current Trends in Translation Studies*, Manchester: St. Jerome, pp.123—136.

Hutchins, W.J., and H.L. Somers. (1992), *An Introduction to Machine Translation*. London: Academic Press.

Kalchbrenner, N., and P. Blunsom. (2013), Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, Seattle, WA., USA, pp.1700—1709.

King, M., and S. Perschke. (1984), EUROTRA. In M. King (ed.) *Machine translation today: the state of the art. Proceedings of the Third Lugano Tutorial*, Lugano, Switzerland, Edinburgh: Edinburgh University Press, 1987, pp.373—391.

Koehn, P. (2005), Europarl: a parallel corpus for statistical machine translation. In *MT Summit X, Conference Proceedings: the tenth Machine Translation Summit*, Phuket, Thailand, pp.79—86.

Koehn, P., H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. (2007), Moses: open source toolkit for statistical machine translation. In *ACL 2007: proceedings of demo and poster sessions*, Prague, Czech Republic, pp.177—180.

Koehn, P., and R. Knowles. (2017), Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, Vancouver, Canada, pp.28—39.

Koehn, P., F. Och, and D. Marcu. (2003), Statistical phrase-based translation. In *HLT-NAACL 2003: conference combining Human Language Technology conference series and the North American Chapter of the Association for Computational Linguistics conference series*, Edmonton, Canada, pp.48—54.

Levenshtein, V.I. (1966), Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady* **10** (8): 707—710.

Luong, M-T., K. Cho, and C. Manning. (2016), Neural Machine Translation. Tutorial presented at *54th Annual Meeting of the Association for Computational Linguistics*, Berlin, Germany (https://nlp.stanford.edu/projects/nmt/Luong-Cho-Manning-NMT-ACL2016-v4.pdf).

Luong, M-T., and C. Manning. (2015), Stanford neural machine translation systems for spoken language domains. In *Proceedings of the 12th International Workshop on Spoken Language Translation*, Da Nang, Vietnam, pp.76—79.

Luong, M-T., and C. Manning. (2016), Achieving Open Vocabulary Neural Machine Translation with Hybrid Word-Character Models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Berlin, Germany, pp.1054—1063.

Ma, Y., Y. He, A. Way, and J. Van Genabith. (2011), Consistent Translation using Discriminative Learning: A Translation Memory-inspired Approach. In *ACL-HLT 2011: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, Portland, Oregon, USA, pp.1239—1248.

Mikolov, T., K. Chen, G. Corrado, and J. Dean. (2013), Efficient Estimation of Word Representations in Vector Space. In *Proceedings of a Workshop at the International Conference on Learning Representations (ICLR-13)*, Scottsdale, AZ, USA, 12pp.

Moorkens, J., and A. Way. (2016), Comparing Translator Acceptability of TM and SMT Outputs. *Baltic Journal of Modern Computing* **4**(2):141—151.

Nagao, M. (1984), A framework of a mechanical translation between Japanese and English by analogy principle, *Artificial and human intelligence: edited review papers presented at the international NATO Symposium*, October 1981, Lyons, France; ed. A. Elithorn and R. Banerji. Amsterdam: North Holland, pp.173—180.

Och, F., and H. Ney. (2003), A systematic comparison of various statistical alignment models. *Computational Linguistics* **29** (1):19—51.

Papineni, K., S. Roukos, T. Ward, and W-J. Zhu. (2002), BLEU: a method for automatic evaluation of machine translation. In *ACL-2002: 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, PA., pp.311—318.

Passban, P., Qun Liu, and A. Way. (2018), Improving Character-based Decoding Using Target-Side Morphological Information for Neural Machine Translation. In *Proceedings of HLT-NAACL 2018, the 16th Annual Conference of the North American Chapter of the*

Association for Computational Linguistics: Human Language Technologies, New Orleans, LA, USA (to appear).

Pierce, J., J. Carroll, E. Hamp, D. Hays, C. Hockett, A. Oettinger, and A. Perlis. (1966), Language and Machines — Computers in Translation and Linguistics. ALPAC report, National Academy of Sciences, National Research Council, Washington, DC.

Popović, M. (2015), ChrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, Lisbon, Portugal, pp.392—395.

Rumelhart, D.E., Hinton, G.E., and Williams, R.J. (1986), Learning representations by back-propagating errors. *Nature* **323**: 533—536.

Sennrich, R., and B. Haddow. (2016), Linguistic Input Features Improve Neural Machine Translation. In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany, pp.83—91.

Sennrich, R., B. Haddow, and A. Birch. (2016), Edinburgh Neural Machine Translation Systems for WMT 16. In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany, pp.371—376.

Shterionov, D., R. Superbo, P. Nagle, L. Casanellas, T. O'Dowd, and A. Way. (2018), Human vs. Automatic Quality Evaluation of NMT and PBSMT. *Machine Translation* **32**(3-4) (in press).

Sikes, R. (2007), Fuzzy matching in theory and practice. *Multilingual* **18**(6):39—43.

Snover, M., B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul. (2006), A study of translation edit rate with targeted human annotation. In *AMTA 2006, Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, Cambridge, MA., pp.223—231.

Sutskever, I., O. Vinyals, and Q.V. Le. (2014), Sequence to Sequence Learning with Neural Networks. In *NIPS 2014: Advances in Neural Information Processing Systems*, Montréal, Canada, pp.3104—3112.

Tillmann, C., S. Vogel, H. Ney, H. Sawaf, and A. Zubiaga. (1997), Accelerated DP-based search for statistical translation. In *Proceedings of the 5th European Conference on Speech Communication and Technology (EuroSpeech '97)*, Rhodes, Greece, pp.2667—2670.

Vauquois, B. (1968), A Survey of Formal Grammars and Algorithms for Recognition and Transformation in Machine Translation. In *Information Processing 68, Proceedings of IFIP Congress 1968*, Edinburgh, UK, pp.254—60.

Wang, L., Z. Tu, X. Zhang, H. Li, A. Way, and Q. Liu. (2016), A Novel Approach to Dropped Pronoun Translation. In *NAACL HLT 2016: Proceedings of The 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, San Diego, CA, pp.983---993.

Way, A. (2009), A Critique of Statistical Machine Translation. *Journal of translation and interpreting studies: Special Issue on Evaluation of Translation Technology* **8**:17—41.

Way, A. (2018), Quality expectations of machine translation. In S. Doherty, S. Castilho, J. Moorkens, and F. Gaspari (eds.) *Human and Machine Translation Quality and Evaluation - From Principles to Practice*, Springer: Berlin (in press).

Way, A., and M. Hearne. (2011), On the Role of Translations in State-of-the-Art Statistical Machine Translation*. Language and Linguistics Compass* **5**:227—248.

Wu, Y., M. Schuster, Z. Chen, Q, V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, L. Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean. (2016), Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. eprint arXiv:1609.08144 (https://arxiv.org/abs/1609.08144).

Zhang, J., X. Wu, I. Calixto, A. Hosseinzadeh Vahid, X. Zhang, A. Way, and Q. Liu. (2014), Experiments in Medical Translation Shared Task at WMT 2014. In *Proceedings of WMT 2014: the Ninth Workshop on Statistical Machine Translation*, Baltimore, MD., pp.260—265.