# Editors' foreword to the invited issue on SMT and NMT

Andy Way. ADAPT Centre, School of Computing, Dublin City University Ireland

Mikel L. Forcada. Departament de Llenguatges i Sistemes Informàtics Universitat d'Alacant, Spain

Until quite recently, phrase-based statistical machine translation (PB-SMT) (Koehn et al. 2003, 2007; Koehn 2010) was indisputably the dominant paradigm in the field of MT. Papers suggesting how neural networks could be used for MT had been published twenty years ago (Chalmers 1990; Chrisman 1991; Castano and Casacuberta 1997; Forcada and Ñeco 1997; Ñeco and Forcada 1997), but the hardware around at the time was insufficient to support the amount of computation required for realistic experimentation.

However, the advent of GPUs over the past few years has allowed practical large-scale neural MT (NMT) models to be built. The first NMT systems to show promise used convolutional neural nets (Kalchbrenner and Blunsom 2013), but the quality achieved, as measured with the BLEU score (Papineni et al. 2002), was not better than that of the PB-SMT baseline (cdec: Dyer et al. (2010)). The first encoder–decoder frameworks (Sutskever et al. 2014), which were not that different from some of those architectures from the nineties, improved significantly, especially when these models were extended with a source-language attention model (Bahdanau et al. 2015). While more and more improvements are coming out all the time, the encoder–decoder model with attention remains pretty much the state-of-the-art today.

At IWSLT 2015,[1] the NMT system of Luong and Manning (2015) demonstrated clear wins (more than 2 BLEU points) over a range of different SMT systems for English-to-German. Bentivogli et al. (2016) performed an in-depth human evaluation of the NMT model of Luong and Manning (2015), demonstrating significantly fewer morphological, lexical and word-order errors than the best SMT system. Bentivogli et al. (2016) also showed that the outputs generated by NMT required about 25% fewer edits compared to the best SMT system.

While NMT appeared to show a great deal of promise, many were already claiming it to be the new state-of-the-art in MT. However, while the findings of Bentivogli et al. (2016) were no doubt significant, the superiority of NMT had been demonstrated for just one language pair, and for a single vertical sector.

Unsurprisingly, therefore, a number of papers followed which set out to test which method was better—SMT or NMT—for a range of language pairs and use-cases.

We were both present at EAMT 2017[2] in Prague where many interesting papers addressing this topic were presented. We quickly decided that this research needed to be collected together in a single volume as soon as possible, as the field of MT has never moved as fast as it is at the moment; if these results were not quickly widely disseminated, their relevance as a set of informative case-studies would have been weakened.

Accordingly, we did not put out a call for inclusion in a special issue as would normally be the case. In this issue of the *Machine Translation* journal, we find revised, extended versions of those papers presented in Prague in May 2017.[3] Papers were reviewed in the normal way by at least two experts in the field. We provide here a short summary of each of the papers included:

- **[Klubička et al.]** extends the work by Toral and Sánchez-Cartagena (2017) by presenting a quantitative, fine-grained manual evaluation of three English-to-Croatian systems—PB-SMT, factored PB-SMT, and NMT—using the Multidimensional Quality Metrics (MQM) error taxonomy (Lommel et al. 2013). They find that NMT reduces the errors produced by the 'pure' PB-SMT system by more than half, and that NMT output is more fluent and more grammatical (e.g. it contains fewer agreement errors).
- **[Shterionov et al.]** set out to level the playing field by ensuring that all data used in their evaluation of KantanMT's PB-SMT and NMT engines is identical. They find that while the quality evaluation scores indicate that the PB-SMT engines perform better, human reviewers—all native speakers of the evaluated language pairs—show the opposite results, i.e. that NMT outperforms PB-SMT. This finding was corroborated by the fact that the majority of translators were most productive when post-editing NMT output. As a result, the authors hypothesize that traditional automatic MT evaluation metrics underestimate the adjudged quality of NMT output, and in a series of novel tests, determine that this underestimation can approach 50%, i.e. in half of those cases in which automatic metrics judged PB-SMT to be better, all three annotators judged NMT to be better.
- **[Popović]** conducts an extensive comparison on publicly available data between NMT and PB-SMT language-related issues for German-to-English, English-to-German and English-to-Serbian. Her findings corroborate those of Bentivogli et al. (2016), in that NMT is better at handling verb forms and avoiding verb omissions, as well as English noun collocations and German compound words. She also provides a comprehensive list of phenomena that NMT is still not good at dealing with, and suggests hybridization or combination of NMT and PB-SMT in view of the complementarity of their error profiles on the languages studied.
- **[Castilho et al.]** compare the English-to-German, -Greek, -Portuguese, and -Russian PB-SMT and NMT engines developed for translation of data from Massive Open Online Courses (MOOCs). Translation quality is evaluated using both automatic metrics and human evaluation, also using the MQM

error typology. While as expected NMT outperforms PB-SMT in terms of committing fewer inflectional morphology and word-order errors, the authors find that, contrary to previous results, NMT is better across all language pairs on both short and long sentences as regards fluency, while the results are less clear when it comes to adequacy.

In sum, the findings of the papers in this collection appear to demonstrate a definite edge of NMT over PB-SMT, so it can be argued with some conviction that NMT appears to have overtaken PB-SMT as the new state-of-the-art approach to MT. While other studies from the suppliers of freely available online MT systems have claimed that the perceived quality of NMT was approaching that of humans (Wu et al. 2016), and more recently that 'human parity' (defined as statistical indistinguishability in a specific set of subjective quality judgements) has been reached (Hassan et al. 2018), we hope that the comprehensive studies contained herein help readers to see precisely how and where NMT outperforms the previously dominant paradigm. At the same time, this research demonstrates quite clearly that no matter what technique is being applied, MT is not a solved problem, and the papers in this collection provide both reassurance to that effect to human translators/post-editors, as well as a list of remaining problems for system developers to tackle.

# Footnotes

1. http://workshop2015.iwslt.org/
2. https://ufal.mff.cuni.cz/eamt2017/
3. We did reach out to Bentivogli et al. to try to include their seminal study as well, but they had already submitted an extended version of their EMNLP 2016 paper elsewhere, appearing as Bentivogli et al. (2017).


### Acknowledgements

We are extremely grateful to Yvette Graham, Miquel Esplà-Gomis, Felipe Sánchez-Martínez, and Juan Antonio Pérez-Ortiz for helping us with the reviews of these papers. The ADAPT Centre for Digital Content Technology is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.


# References


1. Bahdanau D, Cho K, Bengio Y (2015) Neural machine translation by jointly learning to align and translate. In: Proceedings of the 6th international conference on learning representations (ICLR 2015), San Diego, CA, USAGoogle Scholar
2. Bentivogli L, Bisazza A, Cettolo M, Federico M (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 257–267Google Scholar
3. Bentivogli L, Bisazza A, Cettolo M, Federico M (2017) Neural versus phrase-based MT quality: an in-depth analysis on English-German and English-French. Comput Speech Lang 49:52–70CrossRefGoogle Scholar
4. Castano A, Casacuberta F (1997) A connectionist approach to machine translation. In: Proceedings of EUROSPEECH-1997. Fifth European conference on speech communication and technology, Rhodes, Greece, pp 91–94Google Scholar
5. Chalmers DJ (1990) Syntactic transformations on distributed representations. Connect Sci 2(1–2):53–62CrossRefGoogle Scholar
6. Chrisman L (1991) Learning recursive distributed representations for holistic computation. Connect Sci 3(4):345–366CrossRefGoogle Scholar
7. Dyer C, Lopez A, Ganitkevitch J, Weese J, Ture F, Blunsom P, Sewatian H, Eidelman P, Resnik P (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 7–12Google Scholar
8. Forcada M, Ñeco R (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 453–462CrossRefGoogle Scholar
9. Hassan H, Aue A, Chen C, Chowdhary V, Clark J, Federmann C, Huan X, Junczys-Dowmunt M, Lewis W, Li M, Liu S, Liu TY, Luo R, Menezes A, Qin T, Seide F, Tan X, Tian F, Wu L, Wu S, Xia Y, Zhang D, Zhang Z, Zhou M (2018) Achieving human parity on automatic Chinese to English news translation. arXiv:1803.05567v1
10. Kalchbrenner N, Blunsom P (2013) Recurrent continuous translation models. In: Proceedings of the 2013 conference on empirical methods in natural language processing, Seattle, Washington, USA, pp 1700–1709Google Scholar
11. Koehn P (2010) Statistical machine translation, 1st edn. Cambridge University Press, New York, NY, USAzbMATHGoogle Scholar
12. Koehn P, Och FJ, Marcu D (2003) Statistical phrase-based translation. In: Proceedings of the 2003 conference of the North American chapter of the association for computational linguistics on human language technology, vol 1, Edmonton, Canada, pp 48–54Google Scholar
13. Koehn P, Hoang H, Birch A, Callison-Burch C, Federico M, Bertoldi N, Cowan B, Shen W, Moran C, Zens R, Dyer C, Bojar O, Constantin A, Herbst E (2007) Moses: open source toolkit for statistical machine translation. In: Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions, Prague, Czech Republic, pp 177–180Google Scholar
14. Lommel AR, Burchardt A, Uszkoreit H (2013) Multidimensional quality metrics: a flexible system for assessing translation quality. In: ASLIB 2013: translating and the computer 35, Paddington, London, UKGoogle Scholar
15. Luong MT, Manning CD (2015) Stanford neural machine translation systems for spoken language domains. In: Proceedings of the 12th international workshop on spoken language translation (IWSLT), Da Nang, Vietnam, pp 76–79Google Scholar
16. Ñeco RP, Forcada ML (1997) Asynchronous translations with recurrent neural nets. In: International conference on neural networks, ICNN97, IEEE, Houston, TX, USA, vol 4, pp 2535–2540Google Scholar
17. Papineni K, Roukos S, Ward T, Zhu WJ (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 311–318Google Scholar
18. Sutskever I, Vinyals O, Le QV (2014) Sequence to sequence learning with neural networks. In: Proceedings of advances in neural information processing systems 27: annual conference on neural information processing systems, Montreal, Quebec, Canada, pp 3104–3112Google Scholar
19. Toral A, Sánchez-Cartagena VM (2017) A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions. In: Proceedings of the 15th conference of the European chapter of the association for computational linguistics, vol 1, Long Papers, Valencia, Spain, pp 1063–1073Google Scholar
20. Wu Y, Schuster M, Chen Z, Le QV, Norouzi M, Macherey W, Krikun M, Cao Y, Gao Q, Macherey K, Klingner J, Shah A, Johnson M, Liu X, Kaiser L, Gouws S, Kato Y, Kudo T, Kazawa H, Stevens K, Kurian G, Patil N, Wang W, Young C, Smith J, Riesa J, Rudnick A, Vinyals O, Corrado G, Hughes M, Dean J (2016) Google's neural machine translation system: bridging the gap between human and machine translation. CoRR arXiv:1609.08144