Hybrid MT in Academia & Industry

Andy Way

ADAPT Centre
School of Computing
Dublin City University
Dublin, Ireland

andy.way@adaptcentre.ie

-- *Abstract* --

These days pretty much every MT system developed is called a 'hybrid' system, especially in industry. In this talk, we firstly revisit the definition of Hybrid MT given in Way (2010) in order to set the boundaries for the presentation. We then recall Wu's (2005) “3-D Space of Hybrid Models of Translation”, demonstrating that this novel way of envisioning the space of possible MT systems can even today be enlightening. As one of the main developers of truly hybrid MT systems, we then revisit various novel models from my team at DCU which have permeated the field, especially in academia: EBMT & Word-based SMT, EBMT & Phrase-based SMT, Data-Oriented Translation, and Context-based models of PB-SMT.

We note that one of the reasons we got into Hybrid MT in the first place was due to receiving three rejections for a (good!) paper from three SMT-oriented reviewers. At the time, it was fashionable to 'let the data decide', while over the past 10 years we have been 'smuggling in' all sorts of linguistic information – syntax, semantic, and more recently discourse – in order to break through the quality ceiling. We note that exactly the same thing is happening now with the advent of deep learning approaches; we have already received a review of a (successful!) ACL 2016 paper where one reviewer bemoaned the fact that we had not compared our model with “state-of-the-art neural MT models”, and the new practitioners of this paradigm are repeating the mantra to 'let the data decide'.

We point out that it is not difficult to predict what will happen next. SMT practitioners *now* have a good idea of what SMT can and can't do, and there are many papers appearing that try to tackle SMT's problems in dealing with specific translational phenomena. We assert that it would be wise of the neural MT (NMT) protagonists to not repeat the mistakes of the past (cf. Way, 2009), and to try to find ways of solving the well-known errors that SMT makes with this new paradigm. Already some hybrid SMT-NMT papers are starting to appear; we suggest that progress could be made much more quickly by taking a more structured approach to such novel hybrid models: rather than throw the baby out with the bathwater, let us research how NMT can solve some of the problems of SMT, while at the same investigating whether SMT can fix some of the issues that NMT has.

Coming to MT in industry, as asserted above, pretty much all systems created by translation companies are dubbed 'hybrid', but this notion differs considerably from the use of the term in academic circles. For a start, all industrial MT systems are coupled together with Translation Memory (TM) systems, which is not typically the case in academia at all. We point out that such coupling of TM and MT systems is done in a rather arbitrary way, and many researchers – including ourselves – have demonstrated more effective ways of using the two systems in tandem.

As for the industrial MT systems themselves, what often proves effective is a separation of in-domain and out-of-domain data, where specific models are built using client-provided data and backoff models are used to fill in the gaps where coverage is incomplete from the in-domain data. Unlike in academia, where researchers try to squeeze every last BLEU point out of their systems, in

industry, there are bigger fish to fry, such as pre- and post-processing, glossary integration, and finding the best way of incorporating the human-in-the-loop, all with the ultimate aim of making money from MT.

Finally, we end the presentation with a question: what would you recommend to a company who wants to get into MT? Should they bypass SMT altogether? Use SMT first? Only use SMT? It is quite clear that while some companies are starting to dip their toe in the NMT water, contrary to what is espoused by practitioners of that paradigm, industry at least is not yet prepared to eschew their SMT models in favour of this new 'state-of-the-art'.

References

Andy Way. 2009. A Critique of Statistical Machine Translation. In W. Daelemans and V. Hoste (eds.) *Journal of translation and interpreting studies: Special Issue on Evaluation of Translation Technology*, Linguistica Antverpiensia (LANS 8/2009), Brussels: Academic and Scientific Publishers, Antwerp, Belgium, pp. 17—41.

Andy Way. 2010. Machine Translation. In A. Clark, C. Fox and S. Lappin (eds.) *The Handbook of Computational Linguistics and Natural Language Processing*, Wiley Blackwell, Chichester, UK, pp. 531—573.

Dekai Wu. 2005. MT model space: statistical versus compositional versus example-based machine translation. *Machine Translation* **19** (3): 213—227.