Chapter 1

Introduction

A major recent development in statistical machine translation is the adoption of neural networks. Neural network models promise better sharing of statistical evidence between similar words and inclusion of rich context. This chapter introduces several neural network modeling techniques and explains how they are applied to problems in machine translation

1.1 A Short History

Already during the last wave of neural network research in the 1980s and 1990s, machine translation was in the sight of researchers exploring these methods (Waibel et al., 1991). In fact, the models proposed by Forcada and Ñeco (1997) and Castaño et al. (1997) are striking similar to the current dominant neural machine translation approaches. However, none of these models were trained on data sizes large enough to produce reasonable results for anything but toy examples. The computational complexity involved by far exceeded the computational resources of that era, and hence the idea was abandoned for almost two decades.

During this hibernation period, data-driven approaches such as phrase-based statistical machine translation rose from obscurity to dominance and made machine translation a useful tool for many applications, from information gisting to increasing the productivity of professional translators.

The modern resurrection of neural methods in machine translation started with the integration of neural language models into traditional statistical machine translation systems. The pioneering work by Schwenk (2007) showed large improvements in public evaluation campaigns. However, these ideas were only slowly adopted, mainly due to computational concerns. The use of GPUs for training also posed a challenge for many research groups that simply lacked such hardware or the experience to exploit it.

Moving beyond the use in language models, neural network methods crept into other components of traditional statistical machine translation, such as providing additional scores or extending translation tables (Schwenk, 2012; Lu et al., 2014), reordering (Kanouchi et al., 2016; Li

et al., 2014) and pre-ordering models (de Gispert et al., 2015), and so on. For instance, the joint translation and language model by Devlin et al. (2014) was influential since it showed large quality improvements on top of a very competitive statistical machine translation system.

More ambitious efforts aimed at pure neural machine translation, abandoning existing statistical approaches completely. Early steps were the use of convolutional models (Kalchbrenner and Blunsom, 2013) and sequence-to-sequence models (Sutskever et al., 2014; Cho et al., 2014). These were able to produce reasonable translations for short sentences, but fell apart with increasing sentence length. The addition of the attention mechanism finally yielded competitive results (Bahdanau et al., 2015; Jean et al., 2015b). With a few more refinements, such as byte pair encoding and back-translation of target-side monolingual data, neural machine translation became the new state of the art.

Within a year or two, the entire research field of machine translation went neural. To give some indication of the speed of change: At the shared task for machine translation organized by the Conference on Machine Translation (WMT), only one pure neural machine translation system was submitted in 2015. It was competitive, but outperformed by traditional statistical systems. A year later, in 2016, a neural machine translation system won in almost all language pairs. In 2017, almost all submissions were neural machine translation systems.

At the time of writing, neural machine translation research is progressing at rapid pace. There are many directions that are and will be explored in the coming years, ranging from core machine learning improvements such as deeper models to more linguistically informed models. More insight into the strength and weaknesses of neural machine translation is being gathered and will inform future work.

1.2 Toolkits

There is an extensive proliferation of toolkits available for research, development, and deployment of neural machine translation systems. At the time of writing, the number of toolkits is multiplying, rather than consolidating. So, it is quite hard and premature to make specific recommendations. Nevertheless, some of the promising toolkits are:

- Nematus (based on Theano): `https://github.com/EdinburghNLP/nematus`
- Marian (a C++ re-implementation of Nematus): `https://marian-nmt.github.io/`
- OpenNMT (based on Torch/pyTorch): `http://opennmt.net/`
- xnmt (based on DyNet): `https://github.com/neulab/xnmt`
- Sockeye (based on MXNet): `https://github.com/awslabs/sockeye`
- T2T (based on Tensorflow): `https://github.com/tensorflow/tensor2tensor`