Practical Neural Machine Translation

Rico Sennrich and **Barry Haddow**
School of Informatics, University of Edinburgh
rico.sennrich@ed.ac.uk, bhaddow@inf.ed.ac.uk

Objectives

Neural Machine Translation (NMT) has achieved new breakthroughs in machine translation in recent years. It has dominated recent shared translation tasks in machine translation research, and is also being quickly adopted in industry. The technical differences between NMT and the previously dominant phrase-based statistical approach require that practictioners learn new best practices for building MT systems, ranging from different hardware requirements, new techniques for handling rare words and monolingual data, to new opportunities in continued learning and domain adaptation.

This tutorial is aimed at researchers and users of machine translation interested in working with NMT. The tutorial will cover a basic theoretical introduction to NMT, discuss the components of state-of-the-art systems, and provide practical advice for building NMT systems.

Outline

Part 1: An Introduction to Neural Machine Translation

In the first part, we will give a basic introduction to neural MT, discussing how the translation problem can be modelled via a neural network, introducing RNN language models, the attentional encoder-decoder architecture for NMT, and describe inference with greedy search and beam search.

Part 2: The State-of-the-Art in Neural Machine Translation

In the second part, we will describe the journey from a baseline attentional encoder-decoder with attention to our winning systems at WMT16, including techniques such as subword models, training with monolingual data, and bidirectional decoding. We will also provide an overview of recent analyses of NMT, highlighting its strengths and weaknesses.

Part 3: Practical Advice and Open Problems

The third part will cover practical advice for building NMT systems. We will discuss aspects such as training and decoding speed on different hardware and software, ensembling strategies, continued learning and domain adaptation, and advanced features such as Minimum Risk Training or factored models. We will also discuss recent research and open problems in NMT.

About the Speakers

Rico Sennrich (http://homepages.inf.ed.ac.uk/rsennric/) is a research associate at the Institute for Language, Cognition and Computation, University of Edinburgh. He received his PhD in Computational Linguistics from the University of Zurich in 2013, and has since worked at the University of Edinburgh. His recent research has focused on modelling linguistically challenging phenomena in machine translation, including grammaticality, productive morphology, domain effects, and pragmatic aspects. His work on syntax-based and neural machine translation has resulted in top-ranked submissions to the annual WMT shared translation task in three consecutive years.

Barry Haddow (http://homepages.inf.ed.ac.uk/bhaddow/) is a senior researcher, also at the Institute for Language, Cognition and Computation, University of Edinburgh, where he has worked since 2005. His research in machine translation has covered discriminative training, domain adaptation and evaluation, and most recently neural MT. He is the lead organiser of the Conference (formerly Workshop) in Machine Translation, and its associated shared tasks.