Neural Machine Translation: Basics, Practical Aspects and Recent Trends

Raj Dabre[1] **, Fabien Cromieres**[2]**,** and **Toshiaki Nakazawa**[2]
[1]Graduate School of Informatics, Kyoto University
[2]Japan Science and Technology Agency
raj@nlp.ist.i.kyoto-u.ac.jp, fabien@nlp.ist.i.kyoto-u.ac.jp,
nakazawa@nlp.ist.i.kyoto-u.ac.jp

Abstract

In just a few years, Neural Machine Translation (NMT)(Bahdanau et al., 2015; Cho et al., 2014) has become the main approach to Machine Translation as well as one of the most successful application of Deep Learning to NLP. It leverages powerful machine learning techniques to train complex translation models in an end-to-end manner. Although this area of research is pretty new, the many recent developments combined with the practical difficulties of deep learning can make it difficult for a researcher lacking the background and practical experience to develop state-of-the-art models. This tutorial is aimed at people who want to conduct NMT research but have little prior experience in this field. We hope that by the end of the tutorial the audience will have a working understanding of the basics, practical aspects and the recent advancements in NMT.

1 Tutorial Overview

This tutorial is primarily aimed at researchers who are fairly new to the world of NMT and want to obtain a deep understanding of NMT fundamentals. Because it will also cover the latest developments in NMT, it should also be useful to attendees with more experience in NMT. Roughly half of the tutorial will be spent on understanding the working of the sequence-to-sequence encoder-decoder with attention mechanism. This model introduced in (Bahdanau et al., 2015) has become the de facto baseline model in MT research.

The latter half of this tutorial will cover some practical aspects of applying NMT models such as preprocessing (especially in the case of Asian languages), model training and translation search (decoding). Some of these practical aspects are rarely explicitely described, but are important when one wants to obtain state-of-the-art results. This will be followed by a fairly comprehensive review of the recent advancements and trends in NMT that constitute the state of the art, in particular the recent trend trying to replace the recurrent components with more computation-efficient feed-forward components (as in (Vaswani et al., 2017)). This half of the tutorial will be useful to both NMT beginners as well as those with a fair amount of experience.

2 Structure

1. **Introduction: The appearance of NMT in the Machine Translation world** (15 min)
   - A quick review of the evolution of the approaches to Machine Translation
   - Applications of NMT beyond translation (POS Tagging, Parsing, etc) (Vinyals et al., 2015) and related topics (Image Captioning).
2. **The Encoder-Decoder Model** (45 min)

   The objective of this part of the tutorial is to give an indepth explanation of the recurrent NMT model that uses attention. We will give enough details to make sure that the audience has a good working idea of the NMT model and be in a position to try and implement it by themselves.
   - The general architecture of the recurrent sequence-to-sequence model.
   - Background information and notations (including linear algebra needed to understand).
   - Quick overview of recurrent neural network basics (LSTMs, GRUs, etc).
   - Generic sequence to sequence model.

Proceedings of the The 8th International Joint Conference on Natural Language Processing, pages 11–13,
Taipei, Taiwan, November 27 – December 1, 2017.

- Encoder-Decoder model without attention and the results.
- Attention mechanism and its results and implications.
- Variations of attention mechanisms like local attention and various attention strategies (dot product, linear combination, etc) (Luong et al., 2015).
- Visualizations of attention mechanisms.
- Model training in an end to end fashion.
- Limitations of the current model (Unknown words, gradient propagation issues in stacked RNNs etc).

(EXTRA) Overview of implementations of the NMT models and the frameworks: KNMT[1], Lamtram[2], Open NMT[3], Nematus[4], Tensor2Tensor[5].

Coffee Break (30 min)

3. **Practical NMT** (45 min)

   The objective of this part of the tutorial is to augment the audience's understanding of NMT with various practical ideas that can help improve the quality and speed of NMT as well as showcase the many black box applications of NMT.

   - Preprocessing and management of rare words.
   - Subword units to enable infinite vocabulary.
   - BPE (Byte Pair Encoding) and its impact on translation quality (Sennrich et al., 2016b).
   - Using monolingual corpora to improve NMT (Gülçehre et al., 2015; Sennrich et al., 2016a).
   - Training and Translation search.
   - Optimization algorithms (ADAM etc).
   - Residual connections.
   - Training schedules for optimal results (ADAM –> SGD –> annealing –> early stopping).
   - Regularization, dropout and hyperparameter tuning to improve results.
   - Beam search, model averaging and ensembling.

4. **Recent Developments** (45 min)

   The objective of this part of the tutorial is to bring the audience up to speed with the current SOTA (state-of-the-art) NMT models and advancements. We plan to enumerate the most important ones and thereby provide the audience members a roadmap to understanding the big picture.

   - Facebook's CNN (Convolutional Neural Network) based NMT model (Gehring et al., 2017).
   - Google's Transformer (Vaswani et al., 2017) that replies purely on attention and feedforward networks (SOTA for WMT tasks).
   - Results and speedup in training achieved by these architectures.
   - Multilingual Multiway NMT (MLNMT) (Firat et al., 2016) and Zero Shot NMT (Johnson et al., 2016).
   - Other advances (search-guided, latent graph, pointer networks)

5. **Summary and Conclusion**

3 About the Speakers

- **Raj Dabre**: Graduate School of Informatics, Kyoto University, Japan (raj@nlp.ist.i.kyoto-u.ac.jp)

  Raj Dabre is a 3rd year PhD student at Kyoto University. His research interests center on natural language processing, particularly neural machine translation for low resource languages and domain adaptation. He has MT-related publications in ACL, NAACL, COLING and WMT. He was a part of the organizing committee of COLING 2012 and has coordinated joint research between Kyoto University (Japan) and IIT Bombay (India).

- **Dr. Fabien Cromieres**: Japan Science and Technology Agency (JST), Japan (fabien@nlp.ist.i.kyoto-u.ac.jp)

  Fabien Cromieres is currently working with the Japan Science and Technology Agency in a project aiming at improve the translation of

[1] https://github.com/fabiencro/knmt
[2] https://github.com/neubig/lamtram
[3] http://opennmt.net
[4] https://github.com/EdinburghNLP/nematus
[5] https://github.com/tensorflow/tensor2tensor

technical documents between Japanese and Chinese. Initially focused on Example-Based Machine Translation, he has been working on Neural Machine Translation since the end of 2015. He is one of the authors of KyotoEBMT and KyotoNMT MT systems and has MT-related publications in EACL, NAACL, EMNLP and WAT.

- **Dr. Toshiaki Nakazawa**: Japan Science and Technology Agency (JST), Japan (nakazawa@nlp.ist.i.kyoto-u.ac.jp)

  Toshiaki Nakazawa is currently working for the Japan Science and Technology Agency (JST) as a researcher of Project on Practical Implementation of Japanese to Chinese and Chinese to Japanese Machine Translation. His research interests center on natural language processing, particularly language resource construction, linguistically motivated machine translation and NLP tools in human activities. He is one of the authors of KyotoEBMT and has MT-related publications in NAACL, EMNLP, COLING and WAT. He is also one of the organizers of the WAT workshops.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *In Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015)*, San Diego, USA. International Conference on Learning Representations.

Kyunghyun Cho, Bart van Merriënboer, ÇaÄ§lar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 866–875.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional Sequence to Sequence Learning. *ArXiv e-prints*.

Çaglar Gülçehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loïc Barrault, Huei-Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2015. On using monolingual corpora in neural machine translation. *CoRR*, abs/1503.03535.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda B. Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's multilingual neural machine translation system: Enabling zero-shot translation. *CoRR*, abs/1611.04558.

Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. *CoRR*, abs/1508.04025.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Oriol Vinyals, Ł ukasz Kaiser, Terry Koo, Slav Petrov, Ilya Sutskever, and Geoffrey Hinton. 2015. Grammar as a foreign language. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2773–2781. Curran Associates, Inc.