# The Roles of Language Models and Hierarchical Models in Neural Sequence-to-Sequence Prediction

**Felix Stahlberg**[1]
Department of Engineering
University of Cambridge
Trumpington St, Cambridge CB2 1PZ, UK
fs439@cantab.ac.uk

With the advent of deep learning, research in many areas of machine learning is converging towards the same set of methods and models. For example, long short-term memory networks (Hochreiter and Schmidhuber, 1997) are not only popular for various tasks in natural language processing (NLP) such as speech recognition, machine translation, handwriting recognition, syntactic parsing, etc., but they are also applicable to seemingly unrelated fields such as bioinformatics (Min et al., 2016). Recent advances in contextual word embeddings like BERT (Devlin et al., 2019) boast with achieving state-of-the-art results on 11 NLP tasks with the same model. Before deep learning, a speech recognizer and a syntactic parser used to have little in common as systems were much more tailored towards the task at hand.

At the core of this development is the tendency to view each task as yet another data mapping problem, neglecting the particular characteristics and (soft) requirements that tasks often have in practice. This often goes along with a sharp break of deep learning methods with previous research in the specific area. This thesis can be understood as an antithesis to the prevailing paradigm. We show how traditional symbolic statistical machine translation (Koehn, 2009) models can still improve neural machine translation (Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014; Bahdanau et al., 2015, NMT) while reducing the risk of common pathologies of NMT such as hallucinations and neologisms. Other external symbolic models such as spell checkers and morphology databases help neural models to correct grammatical errors in text.



[1]Now at Google Research.

We also focus on language models that often do not play a role in vanilla end-to-end approaches and apply them in different ways to word reordering, grammatical error correction, low-resource NMT, and document-level NMT. Finally, we demonstrate the benefit of hierarchical models in sequence-to-sequence prediction. Hand-engineered covering grammars are effective in preventing catastrophic errors in neural text normalization systems. Our operation sequence model for interpretable NMT represents translation as a series of actions that modify the translation state, and can also be seen as derivation in a formal grammar.

This thesis also focuses on the decoding aspect of neural sequence models. We argue that NMT decoding is very similar to navigating through a weighted graph structure or finite state machine, with the only difference that the state space may not be finite. This view enables us to use a wide range of search algorithms, and provides a strong formal framework for pairing NMT with other kinds of models. In particular, we apply exact shortest path search algorithms for graphs, such as depth-first search, to NMT, and show that beam decoding fails to find the global best model score in most cases. However, these search errors, paradoxically, often prevent the decoder from suffering from a frequent but very serious model error in NMT, namely that the empty hypothesis often has the global best model score.

The main contributions of this thesis are implemented in a novel open-source NMT decoding framework called SGNMT[2] which allows paring neural translation models with different kinds of constraints and symbolic models. SGNMT is compatible to a range of popular toolkits such as Ten-

[2]https://ucam-smt.github.io/sgnmt/html/

sor2Tensor (Vaswani et al., 2018) and fairseq (Ott et al., 2019) for neural models, KenLM (Heafield, 2011) for language modelling, and OpenFST (Allauzen et al., 2007) for finite state transducers. SGNMT has been used for: (1) teaching as SGNMT has been used for course work and student theses in the MPhil in Machine Learning and Machine Intelligence at the University of Cambridge, (2) research as most of the research work of the Cambridge MT group, including four successful WMT submissions, is based on SGNMT, and (3) technology transfer as SGNMT has helped to transfer research findings from the laboratory to the industry, eg. into a product of SDL plc.

The Apollo repository of the University of Cambridge provides open access to the full thesis (`https://doi.org/10.17863/CAM.49422`).

## Acknowledgements

The author would like to thank his Ph.D. supervisor, Bill Byrne, and his thesis examiners, Paula Buttery and Adam Lopez. The author was financially supported by the U.K. Engineering and Physical Sciences Research Council (EPSRC grant EP/L027623/1). Some of the work has been performed using resources provided by the Cambridge Tier-2 system operated by the University of Cambridge Research Computing Service[3] funded by EPSRC Tier-2 capital grant EP/P020259/1.

## References

Allauzen, Cyril, Michael Riley, Johan Schalkwyk, Wojciech Skut, and Mehryar Mohri. 2007. Openfst: A general and efficient weighted finite-state transducer library. In Holub, Jan and Jan Žďárek, editors, *Implementation and Application of Automata*, pages 11–23, Berlin, Heidelberg. Springer Berlin Heidelberg.

Bahdanau, Dzmitry, Kyung Hyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*, January.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Heafield, Kenneth. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland, July. Association for Computational Linguistics.

Hochreiter, Sepp and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Kalchbrenner, Nal and Phil Blunsom. 2013. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1700–1709, Seattle, Washington, USA, October. Association for Computational Linguistics.

Koehn, Philipp. 2009. *Statistical machine translation*. Cambridge University Press.

Min, Seonwoo, Byunghan Lee, and Sungroh Yoon. 2016. Deep learning in bioinformatics. *Briefings in Bioinformatics*, 18(5):851–869, 07.

Ott, Myle, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Sutskever, Ilya, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In Ghahramani, Z., M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc.

Vaswani, Ashish, Samy Bengio, Eugene Brevdo, Francois Chollet, Aidan Gomez, Stephan Gouws, Llion Jones, Łukasz Kaiser, Nal Kalchbrenner, Niki Parmar, Ryan Sepassi, Noam Shazeer, and Jakob Uszkoreit. 2018. Tensor2Tensor for neural machine translation. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers)*, pages 193–199, Boston, MA, March. Association for Machine Translation in the Americas.

[3] `http://www.hpc.cam.ac.uk`