Neural Machine Translation

Max Mustermann, and Hermann Ney
<surname>@i6.informatik.rwth-aachen.de

April 1st, 2017

Human Language Technology and Pattern Recognition
Computer Science Department, RWTH Aachen University

Outline

NMT History

Sequence to Sequence Translation

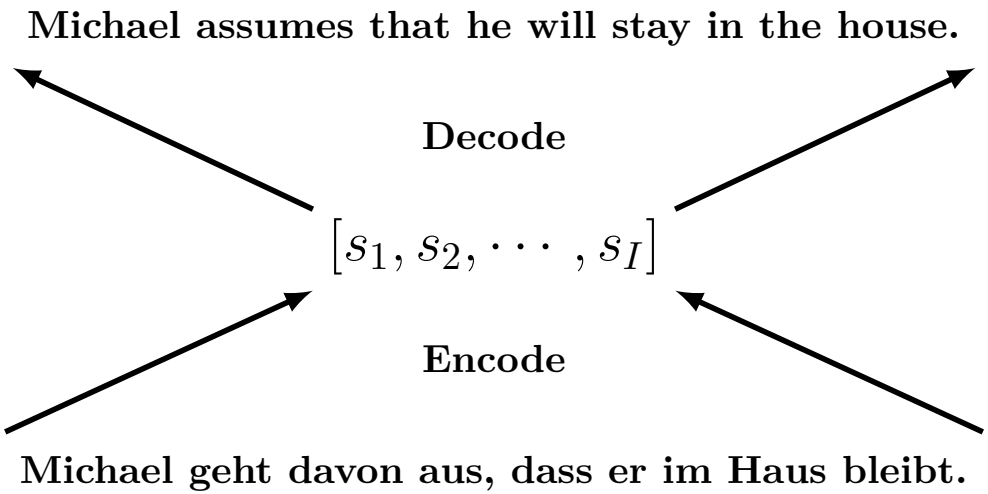
Comparison

NMT History

- **2014** Sequence to sequence translation [Sutskever & Vinyals+ 14, Cho & van Merrienboer+ 14b]
- **2015** Attention-based neural machine translation [Bahdanau & Cho+ 15, Luong & Pham+ 15a]
- **2016** Subword-NMT [Sennrich & Haddow+ 16b]
- Multi-lingual NMT [Firat & Cho+ 16]
- WMT 2016
- Google NMT [Wu & Schuster+ 16]
- Fully char-level NMT [Lee & Cho+ 16]
- **2017** ...

Sequence to Sequence Translation

- Basic principle behind neural machine translation
- Translation from one language to another can be done by a single neural network
- The input and output are both variable-length sequences
- An encoder reads the source sentence, encodes it into a vector
- Decoder uses this vector to predict the target words



Comparison to Best Systems

	WMT2015 De→En newstest2014		WMT2016 En→Ro newstest2016		BOLT Zh→En test1	
	BLEU %	TER %	BLEU %	TER %	BLEU %	TER %
Model A	28.1	53.2	24.5	59.3	17.9	67.7
Model B	24.0	57.2	26.7	56.6	17.9	67.2
Model C	28.2	53.0	25.3	58.4	17.6	68.9
Model D	29.2	52.9	27.1	55.4	21.7	63.7

- De→En and Zh→En:
  - Subwords and LSTM

Thank you for your attention

Max Mustermann, and Hermann Ney

<surname>@cs.rwth-aachen.de

Results: Cool Method with Overlay (IWSLT2013)

IWSLT De-En	dev		test		eval11		alignment-test	
Model	BLEU %	TER %	BLEU %	TER %	BLEU %	TER %	AER%	SAER%
Attention-Based	30.5	48.7	29.3	50.6	33.9	46.6	41.8	66.3
+ method	31.5	47.2	30.3	49.0	34.3	44.3	35.4	44.2

- Improves translation by an average of 0.8 BLEU on IWSLT2013
- Great improvement in AER and SAER

Results: Cool Method with Overlay (WMT2016)

WMT En-Ro	newsdev2016/1		newsdev2016/2		newstest2016	
Model	BLEU %	TER %	BLEU %	TER %	BLEU %	TER %
Attention-Based	19.8	62.0	21.3	58.1	20.3	60.4
+GA	21.0	61.1	23.6	56.4	21.8	59.4
+method + conv ($D = 10, M = 1$)	21.4	60.1	24.7	55.4	22.3	58.7

- Improves translation by an average of 0.8 BLEU on IWSLT2013
- Great improvement in AER and SAER
- Improves translation by an average of 1.7 BLEU on WMT2016
  - Adding convolutional feedback gives an additional 0.6 BLEU on average

Reference

- **D. Bahdanau, K. Cho, Y. Bengio.**
  Neural machine translation by jointly learning to align and translate.
  In *Proceedings of the International Conference on Learning Representations (ICLR)*, San Diego, CA, 2015.

- **Y. Bengio, R. Ducharme, P. Vincent, C. Jauvin.**
  A neural probabilistic language model.
  journal of machine learning research, Vol. 3, No. Feb, pp. 1137–1155, 2003.

- **M. Castana, E. Vidal, F. Casacuberta.**
  Inference of stochastic regular languages through simple recurrent networks.
  In *Grammatical Inference: Theory, Applications and Alternatives, IEE Colloquium on*, pp. 16–1. IET, 1993.

- **W. Chen, E. Matusov, S. Khadivi, J. Peter.**
  Guided alignment training for topic-aware neural machine translation.
  CoRR, Vol. abs/1607.01628, 2016.

- **W. Chen, E. Matusov, S. Khadivi, J.-T. Peter.**
  Guided alignment training for topic-aware neural machine translation.
  Austion, Texas, October 2016. Association for Machine Translation in the Americas.
- **K. Cho, B. van Merrienboer, D. Bahdanau, Y. Bengio.**
  On the properties of neural machine translation: Encoder-decoder approaches.
  In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pp. 103—-111, Doha, Qatar, October 2014.
- **K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio.**
  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)*, pp. 1724–1734, Doha, Qatar, October 2014. Association for Computational Linguistics.

- **K. Cho.**
  Natural language understanding with distributed representation.
  Technical report, New York University, 2015.
  Lecture Note for DS-GA 3001.
- **J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, Y. Bengio.**
  Attention-based models for speech recognition.
  In *Advances in Neural Information Processing Systems*, pp. 577–585, 2015.
- **J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, Y. Bengio.**
  Attention-based models for speech recognition.
  In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pp. 577–585. Curran Associates, Inc., 2015.

- **J. Chung, K. Cho, Y. Bengio.**
  A character-level decoder without explicit segmentation for neural machine translation.
  Proceedings of Association for Computational Linguistics (ACL), Vol., pp. 1693–1703, 2016.
- **J. Chung, K. Cho, Y. Bengio.**
  A character-level decoder without explicit segmentation for neural machine translation.
  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*, 2016.
- **T. Cohn, C. D. V. Hoang, E. Vymolova, K. Yao, C. Dyer, G. Haffari.**
  Incorporating structural alignment biases into an attentional neural translation model.
  CoRR, Vol. abs/1601.01085, 2016.

- **O. Firat, K. Cho, B. Sankaran, F. T. Y. Vural, Y. Bengio.**
  Multi-way, multilingual neural machine translation.
  Computer Speech & Language, Vol., 2016.

- **Ç. Gülçehre, O. Firat, K. Xu, K. Cho, L. Barrault, H. Lin, F. Bougares, H. Schwenk, Y. Bengio.**
  On using monolingual corpora in neural machine translation.
  CoRR, Vol. abs/1503.03535, 2015.

- **S. Jean, K. Cho, R. Memisevic, Y. Bengio.**
  On using very large target vocabulary for neural machine translation.
  In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1–10, Beijing, China, July 2015. Association for Computational Linguistics.

- **J. Lee, K. Cho, T. Hofmann.**
  Fully character-level neural machine translation without explicit segmentation.
  CoRR, Vol. abs/1610.03017, 2016.

- **M.-T. Luong, H. Pham, C. D. Manning.**
  Effective approaches to attention-based neural machine translation.
  Vol., 2015.
- **M. Luong, H. Pham, C. D. Manning.**
  Effective approaches to attention-based neural machine translation.
  In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1412—-1421, Lisbon, Portugal, 2015.
- **H. Mi, Z. Wang, A. Ittycheriah.**
  Supervised attentions for neural machine translation.
  arXiv preprint arXiv:1608.00112, Vol., 2016.
- **T. Mikolov, M. Karafiát, L. Burget, J. Cernockỳ, S. Khudanpur.**
  Recurrent neural network based language model.
  In *Interspeech*, Vol. 2, 3, 2010.

- **F. J. Och.**
  Minimum error rate training in statistical machine translation.
  In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pp. 160–167, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

- **H. Schwenk, M. R. Costa-Jussa et al.**
  Smooth bilingual n-gram translation.
  In *In EMNLP*. Citeseer, 2007.

- **H. Schwenk.**
  Continuous space language models.
  Computer Speech & Language, Vol. 21, No. 3, pp. 492–518, 2007.

- **R. Sennrich, B. Haddow, A. Birch.**
  Improving neural machine translation models with monolingual data.
  Association for Computational Linguistics (ACL), Vol., pp. 86–96, 2016.

- **R. Sennrich, B. Haddow, A. Birch.**
  Neural machine translation of rare words with subword units.
  Association for Computational Linguistics (ACL), Vol., pp. 1715–1725, 2016.
- **S. Shen, Y. Cheng, Z. He, W. He, H. Wu, M. Sun, Y. Liu.**
  Minimum risk training for neural machine translation.
  Association for Computational Linguistics (ACL), Vol., 2016.
- **S. Sukhbaatar, J. Weston, R. Fergus et al.**
  End-to-end memory networks.
  In *Advances in neural information processing systems*, pp. 2440–2448, 2015.
- **M. Sundermeyer, R. Schlüter, H. Ney.**
  Lstm neural networks for language modeling.
  In *Interspeech*, pp. 194–197, 2012.

- **I. Sutskever, O. Vinyals, Q. V. V. Le.**
  Sequence to sequence learning with neural networks.
  In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, K. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pp. 3104–3112. Curran Associates, Inc., Monteal, Canada, 2014.

- **Z. Tu, Z. Lu, Y. Liu, X. Liu, H. Li.**
  Coverage-based neural machine translation.
  CoRR, Vol. abs/1601.04811, 2016.

- **Z. Tu, Z. Lu, Y. Liu, X. Liu, H. Li.**
  Modeling coverage for neural machine translation.
  In *54th Annual Meeting of the Association for Computational Linguistics*, 2016.

- **A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, K. Lang.**
  Phoneme recognition: neural networks vs. hidden markov models vs. hidden markov models.
  In *Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on*, pp. 107–110. IEEE, 1988.

- **Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, Klaus et al.**
  Google's neural machine translation system: Bridging the gap between human and machine translation.
  CoRR, Vol. abs/1609.08144, 2016.

- **K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, Y. Bengio.**
  Show, attend and tell: Neural image caption generation with visual attention.
  In *International Conference on Machine Learning*, 2015.

- **L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, A. Courville.**
  Describing videos by exploiting temporal structure.
  In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4507–4515, 2015.