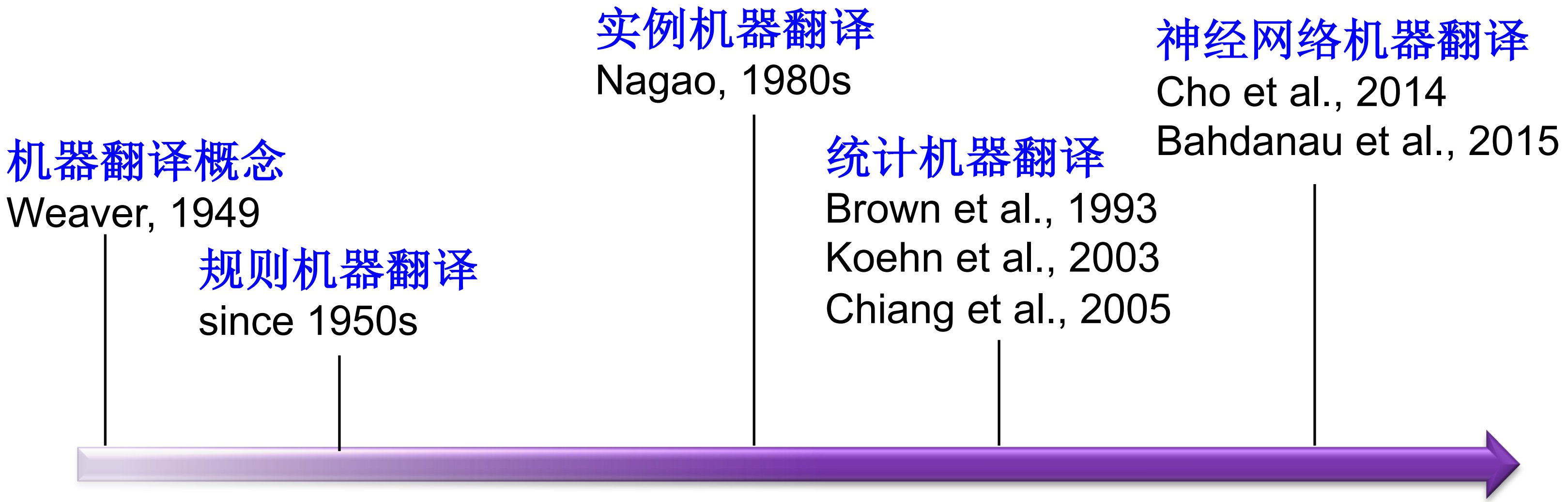
NLP 南京大学 自然语言处理研究组 Natural Language Processing Group

南京大學 NANJING UNIVERSITY

机器翻译的研究历程 --神经网络机器翻译

黄书剑

机器翻译的发展



机器翻译概念
Weaver, 1949

规则机器翻译
since 1950s

实例机器翻译
Nagao, 1980s

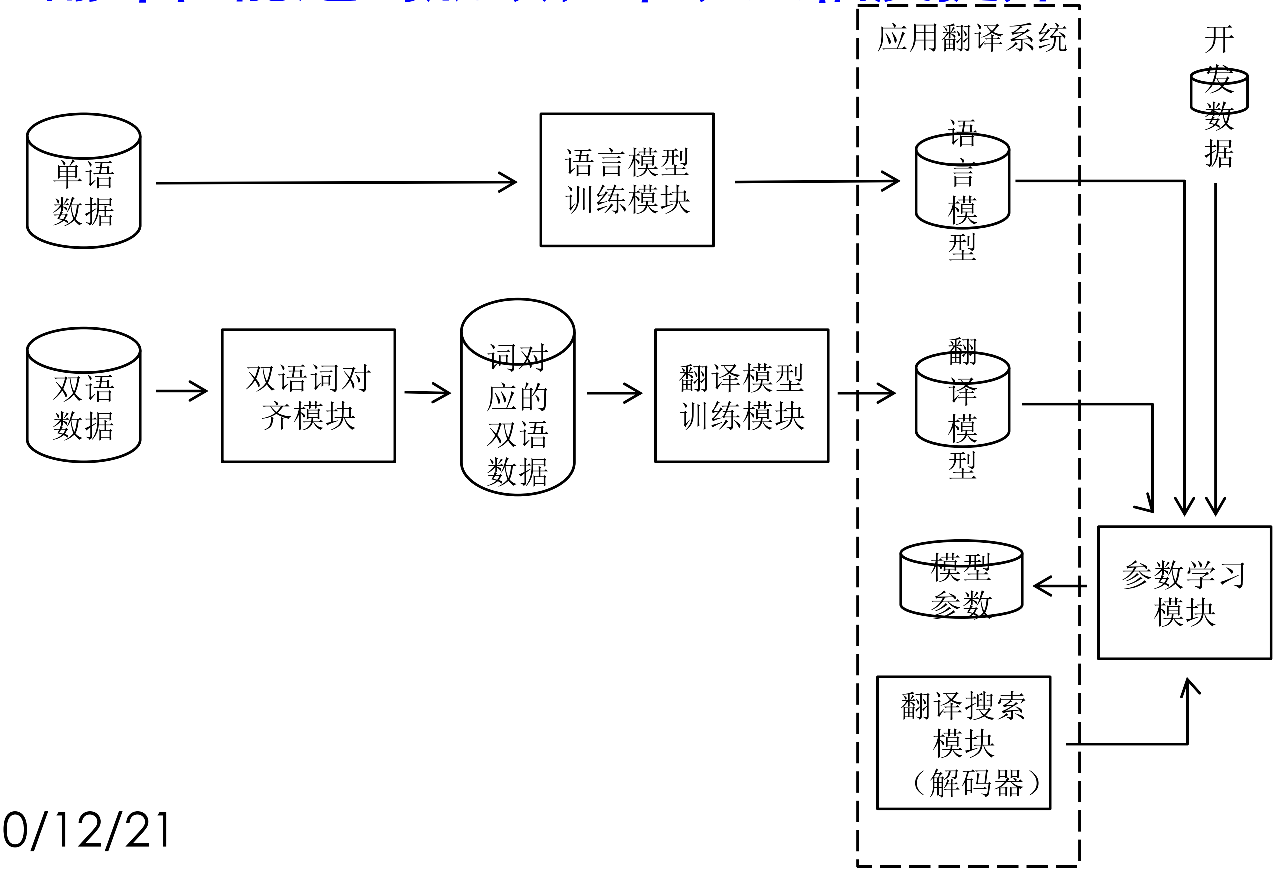
统计机器翻译
Brown et al., 1993
Koehn et al., 2003
Chiang et al., 2005

神经网络机器翻译
Cho et al., 2014
Bahdanau et al., 2015

数据资源越来越丰富
计算能力越来越强

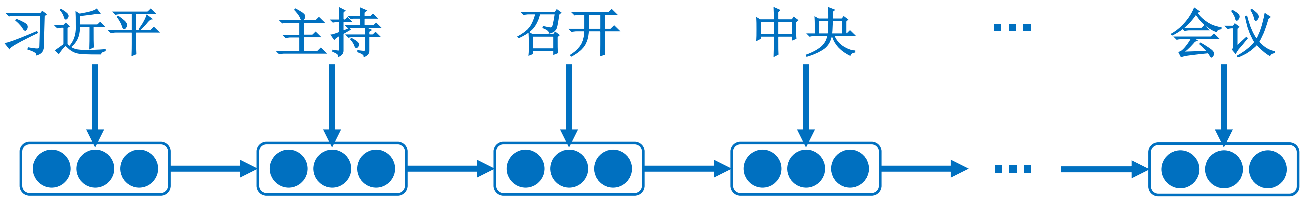
统计机器翻译回顾

- 可以一定程度上从数据中自动挖掘翻译知识
- 流程相对复杂，其中各个部分都不断被改进和优化
- 翻译性能遇到瓶颈，难以大幅度提升



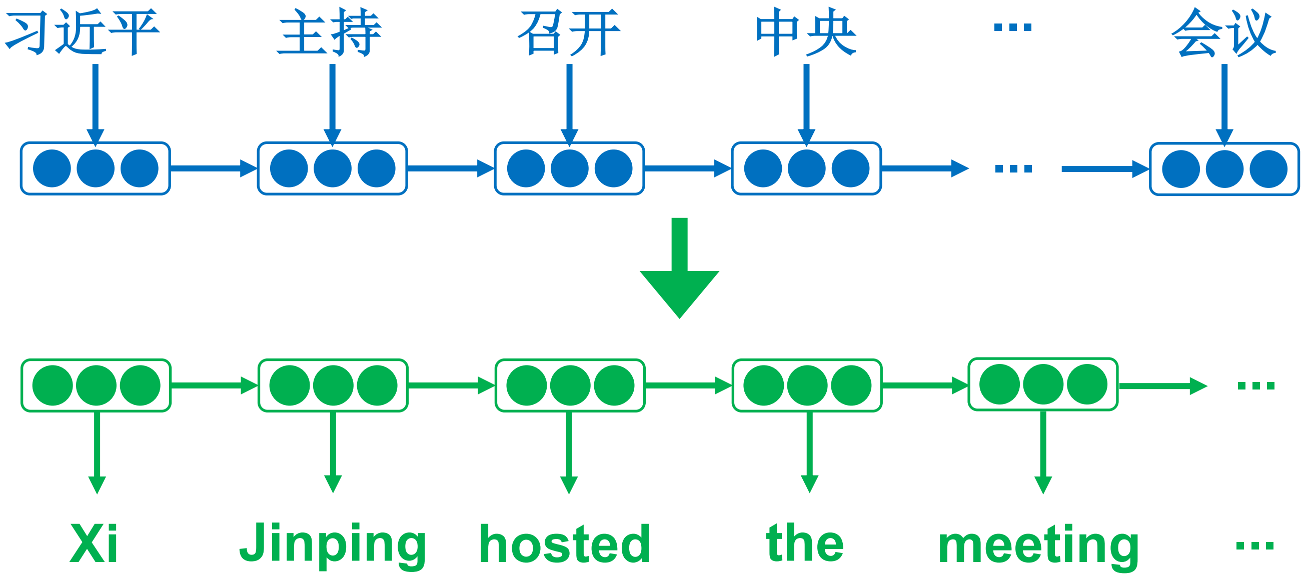
神经网络机器翻译

- 从单词序列到单词序列的翻译方式
  - 简单直接的把句子看做单词序列
  - 不再依赖大量从语料库中学习得到的有噪音规则
  - 例如：习近平主持召开中央全面深化改革领导小组会议



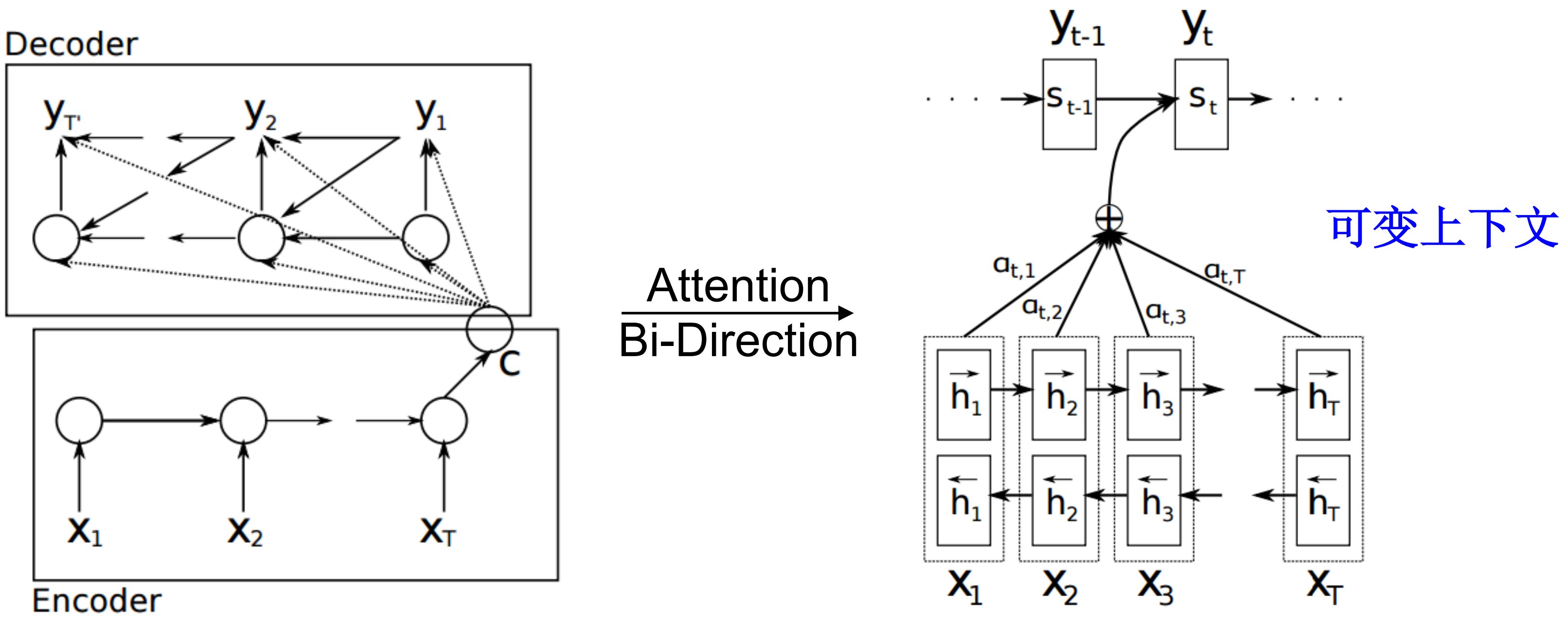
神经网络机器翻译

- 从单词序列到单词序列的翻译方式
  - 简单直接的把句子看做单词序列
  - 不需要建模规则的组合关系
  - 例如：习近平主持召开中央全面深化改革领导小组会议



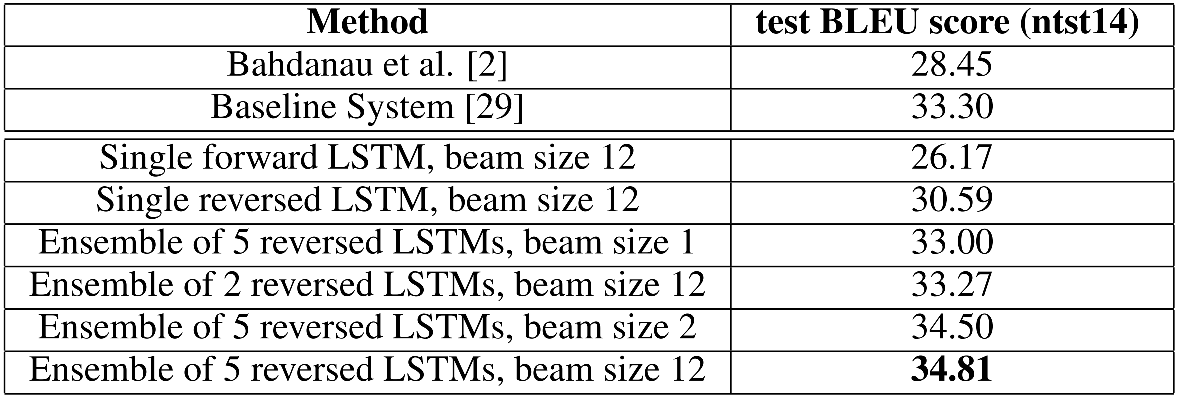
神经网络机器翻译

- **从单词序列到单词序列的翻译方式**
  - 简单直接的把句子看做单词序列
  - Bi-directional RNN + Attention



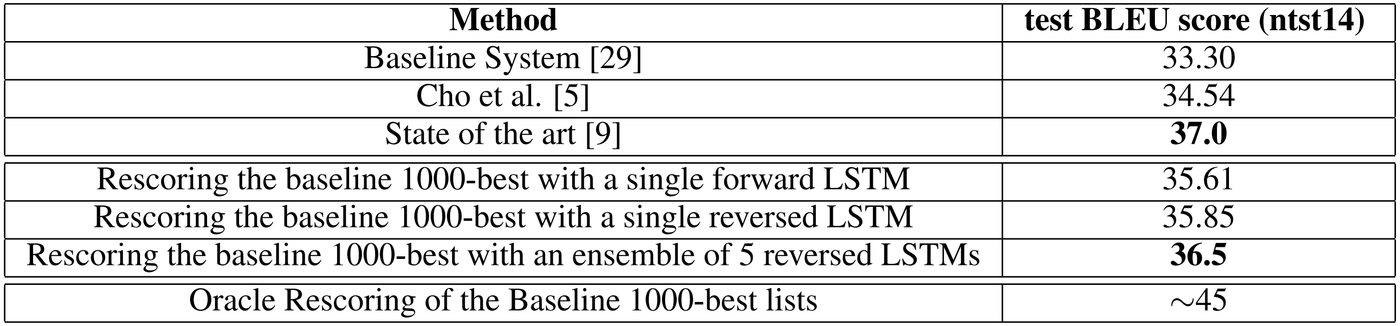
(Cho et al., 2014)

(Bahdanau et al., 2015)



Method	test BLEU score (ntst14)
Bahdanau et al. [2]	28.45
Baseline System [29]	33.30
Single forward LSTM, beam size 12	26.17
Single reversed LSTM, beam size 12	30.59
Ensemble of 5 reversed LSTMs, beam size 1	33.00
Ensemble of 2 reversed LSTMs, beam size 12	33.27
Ensemble of 5 reversed LSTMs, beam size 2	34.50
Ensemble of 5 reversed LSTMs, beam size 12	**34.81**

Table 1: The performance of the LSTM on WMT'14 English to French test set (ntst14). Note that an ensemble of 5 LSTMs with a beam of size 2 is cheaper than of a single LSTM with a beam of size 12.

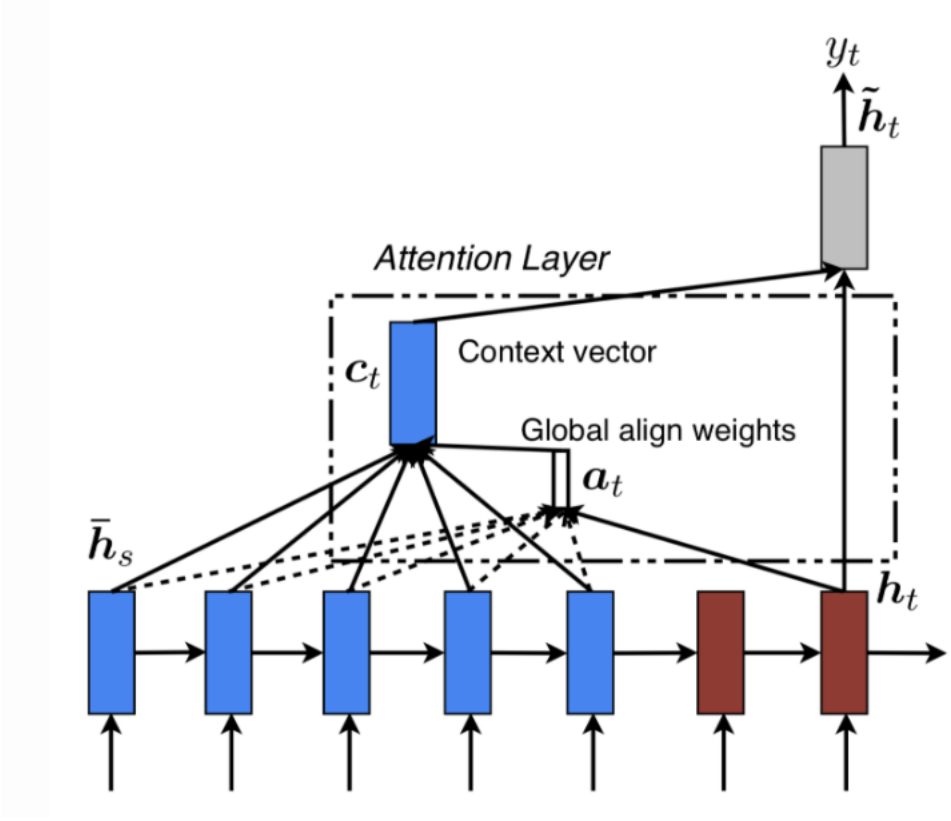


Method	test BLEU score (ntst14)
Baseline System [29]	33.30
Cho et al. [5]	34.54
State of the art [9]	**37.0**
Rescoring the baseline 1000-best with a single forward LSTM	35.61
Rescoring the baseline 1000-best with a single reversed LSTM	35.85
Rescoring the baseline 1000-best with an ensemble of 5 reversed LSTMs	**36.5**
Oracle Rescoring of the Baseline 1000-best lists	∼45

Table 2: Methods that use neural networks together with an SMT system on the WMT'14 English to French test set (ntst14).

[Sutskever et al. 2014]

Attention Mechanism



$$\begin{aligned}\boldsymbol{a}_t(s) &= \operatorname{align}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s)\\ &= \frac{\exp\left(\operatorname{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s)\right)}{\sum_{s'} \exp\left(\operatorname{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_{s'})\right)}\end{aligned}$$

$$\operatorname{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s) = \begin{cases} \boldsymbol{h}_t^\top \bar{\boldsymbol{h}}_s & \textit{dot}\\ \boldsymbol{h}_t^\top \boldsymbol{W_a} \bar{\boldsymbol{h}}_s & \textit{general}\\ \boldsymbol{W_a}[\boldsymbol{h}_t; \bar{\boldsymbol{h}}_s] & \textit{concat} \end{cases}$$

Thang Luong et al. 2015

神经网络机器翻译

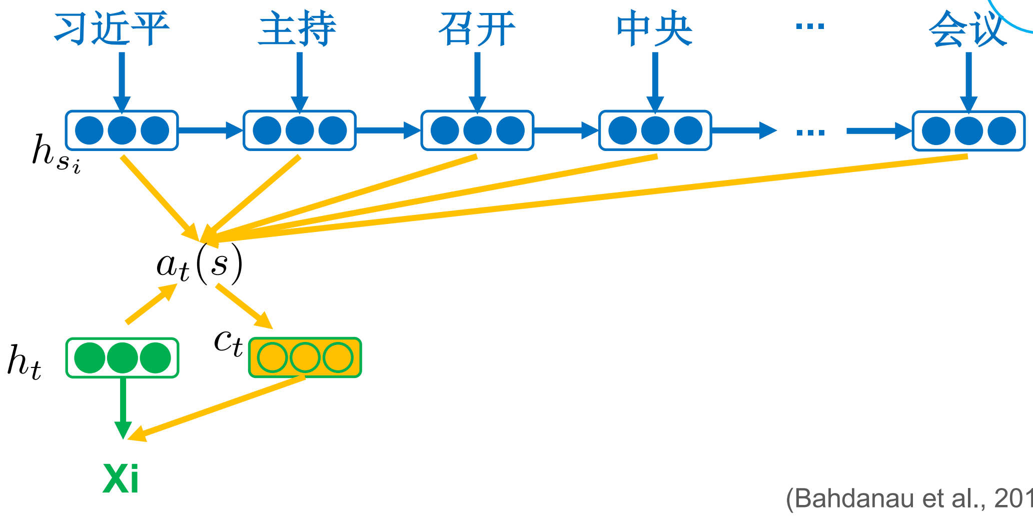
- 从单词序列到单词序列的翻译方式
  - 简单直接的把句子看做单词序列
  - 利用注意力机制动态获取信息
  - 例如：习近平主持召开中央全面深化改革领导小组会议

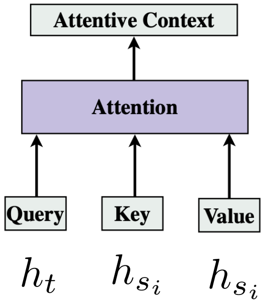
$$a_t(s_i) = \frac{\exp(score(h_t, h_{s_i}))}{\sum_j \exp(score(h_t, h_{s_j}))}$$

$$c_t(s) = \sum_j a_t(s_j) h_{s_j}$$

$$\mathrm{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s) = \begin{cases} \boldsymbol{h}_t^\top \bar{\boldsymbol{h}}_s & \textit{dot} \\ \boldsymbol{h}_t^\top \boldsymbol{W_a} \bar{\boldsymbol{h}}_s & \textit{general} \\ \boldsymbol{W_a}[\boldsymbol{h}_t; \bar{\boldsymbol{h}}_s] & \textit{concat} \end{cases}$$

(Thang Luong et al. 2015)

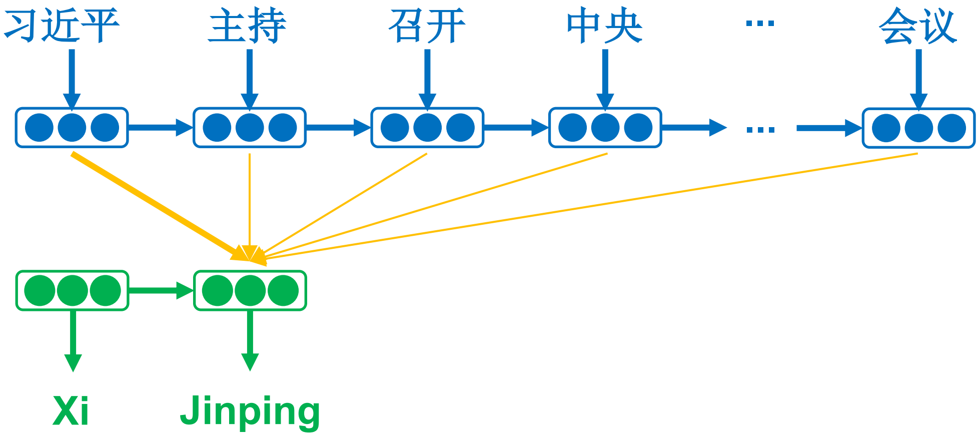




(Bahdanau et al., 2015)

神经网络机器翻译

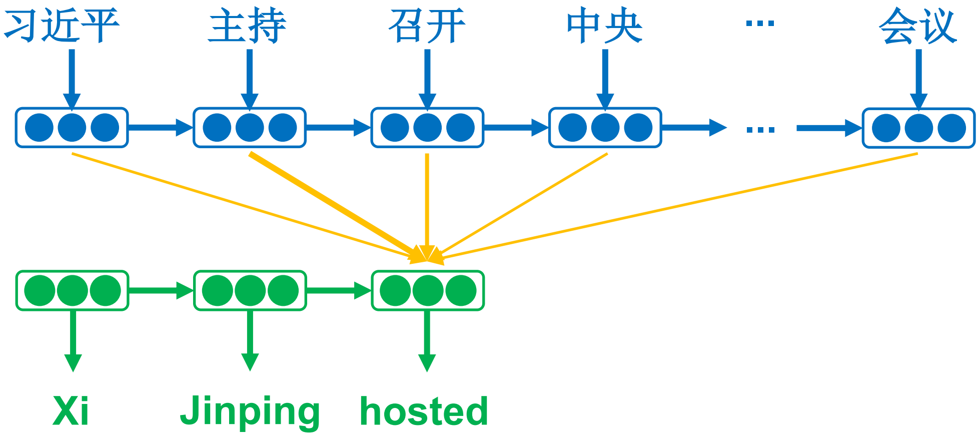
- 从单词序列到单词序列的翻译方式
  - 简单直接的把句子看做单词序列
  - 利用注意力机制动态获取信息
  - 例如：习近平主持召开中央全面深化改革领导小组会议



(Bahdanau et al., 2015)

神经网络机器翻译

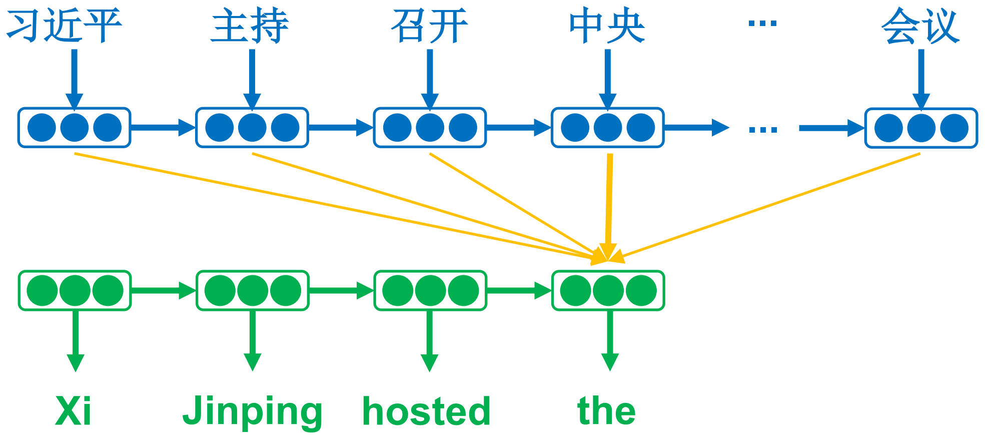
- 从单词序列到单词序列的翻译方式
  - 简单直接的把句子看做单词序列
  - 利用注意力机制动态获取信息
  - 例如：习近平主持召开中央全面深化改革领导小组会议



(Bahdanau et al., 2015)

神经网络机器翻译

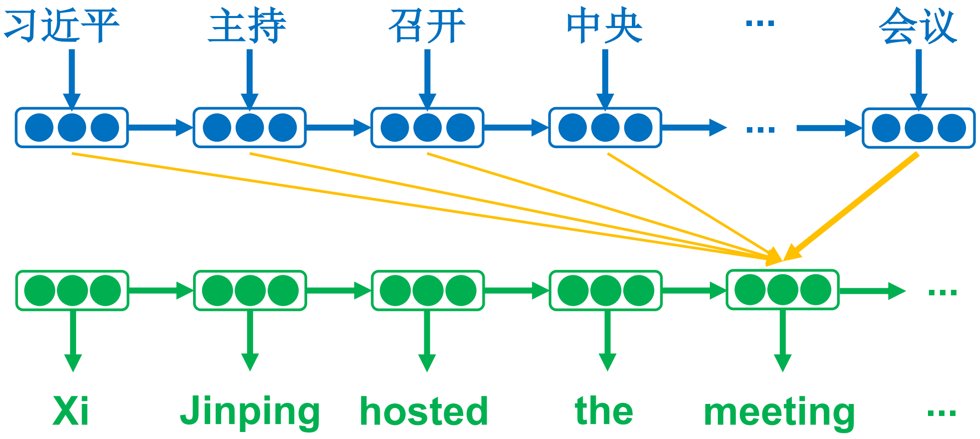
- 从单词序列到单词序列的翻译方式
  - 简单直接的把句子看做单词序列
  - 利用注意力机制动态获取信息
  - 例如：习近平主持召开中央全面深化改革领导小组会议



(Bahdanau et al., 2015)

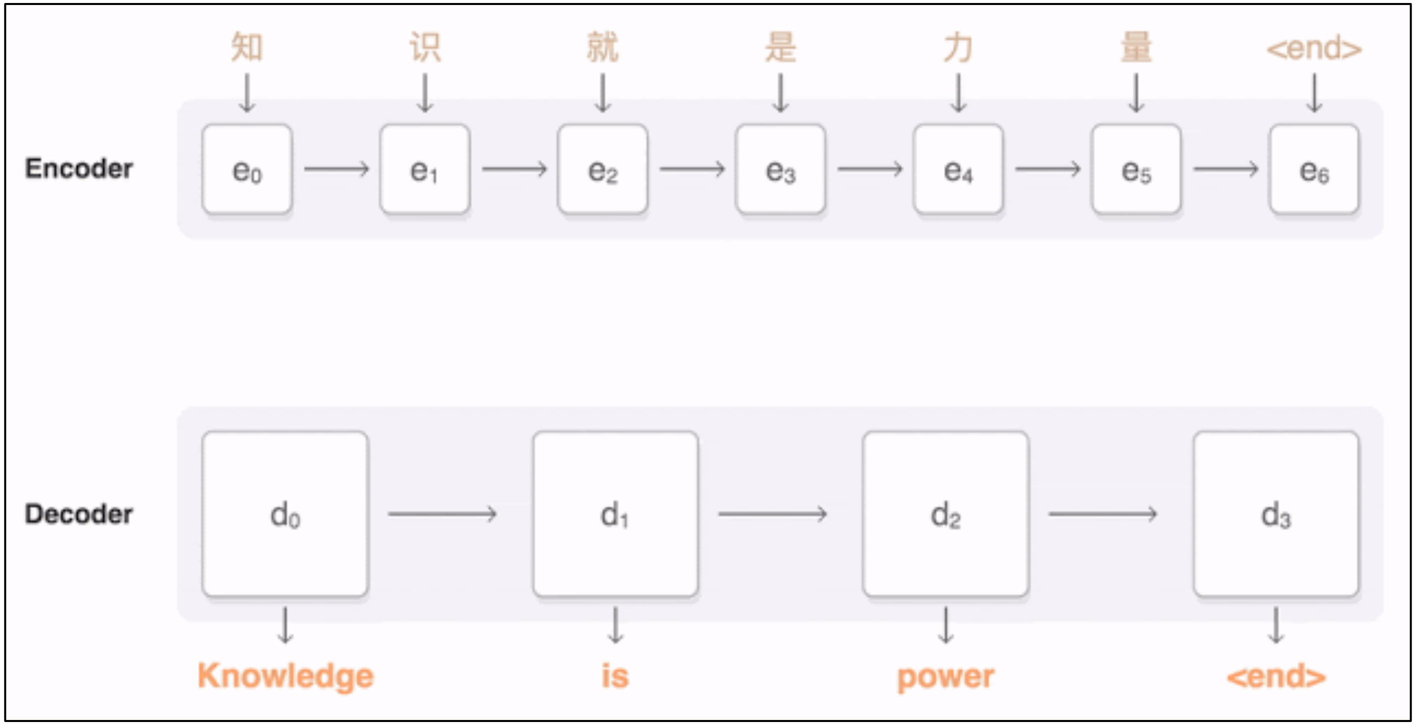
神经网络机器翻译

- 从单词序列到单词序列的翻译方式
  - 简单直接的把句子看做单词序列
  - 利用注意力机制动态获取信息
  - 例如：习近平主持召开中央全面深化改革领导小组会议

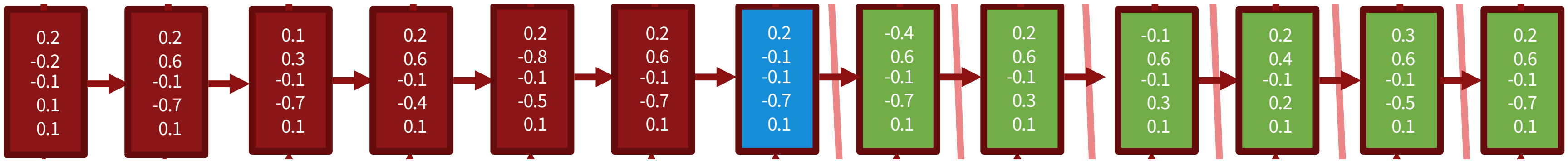


(Bahdanau et al., 2015)

Encoder-Decoder with Attention

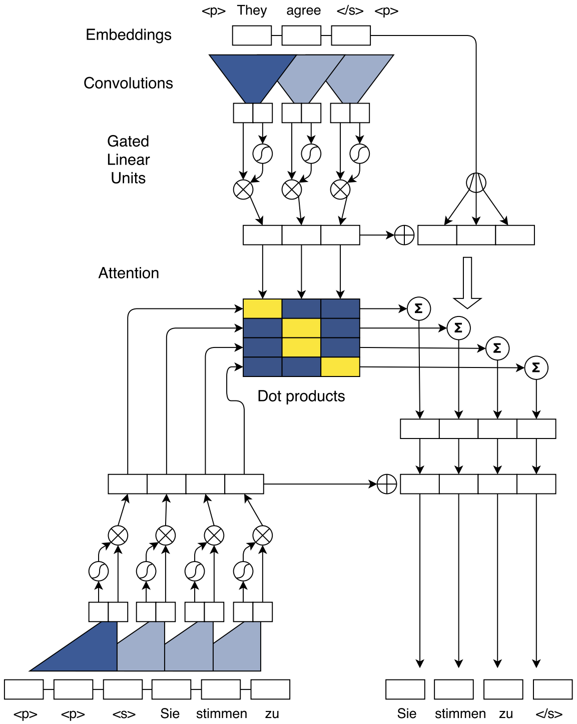


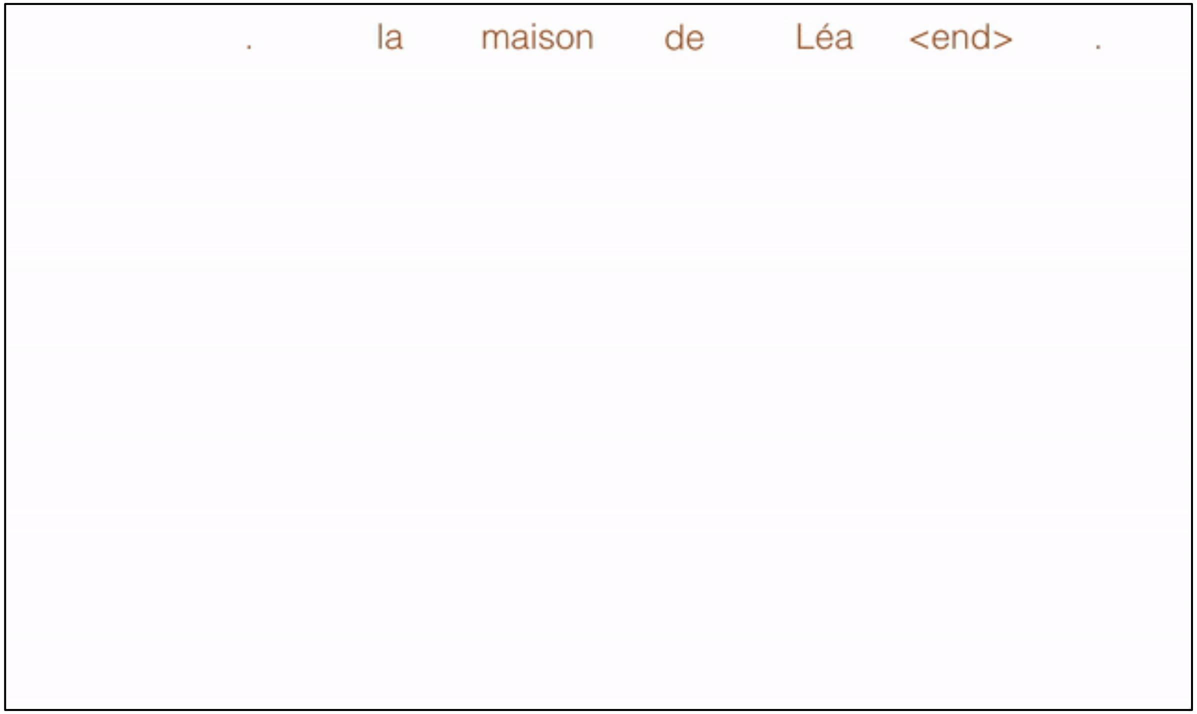
https://medium.com/analytics-vidhya/transformer-vs-rnn-and-cnn-18eeefa3602b



(Thang Luong et al., 2016)

CNN-based Encoder Decoder

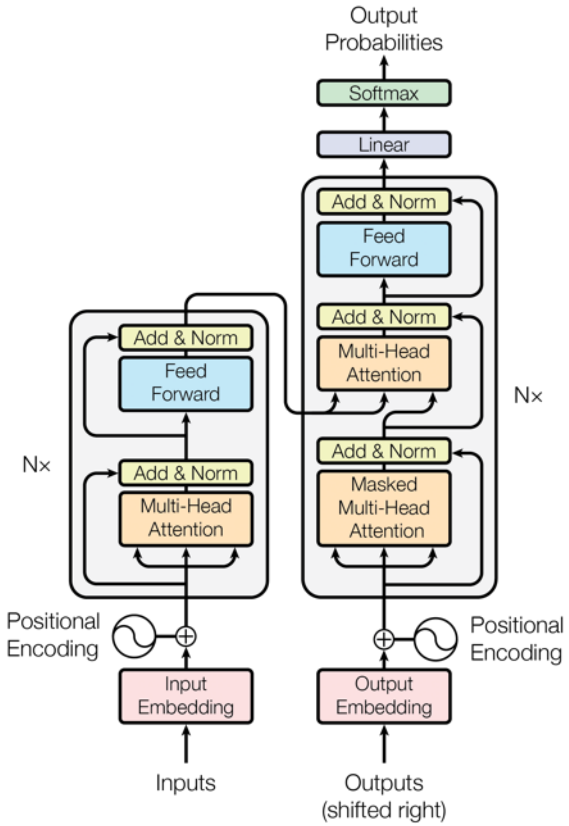




Gehring et al., (2017)

https://medium.com/analytics-vidhya/transformer-vs-rnn-and-cnn-18eeefa3602b

Self-Attention Networks (Transformer)



Vaswani et al., (2017)

https://medium.com/analytics-vidhya/transformer-vs-rnn-and-cnn-18eeefa3602b

Self-Attention v.s. RNN/CNN

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

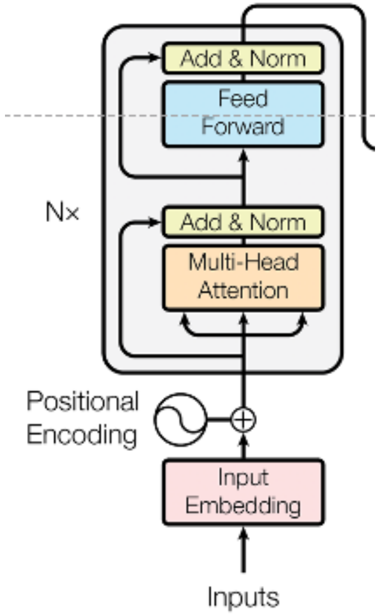
- Q=K=V for self-attention

- Total computational complexity
- The amount of computation that can be parallelized
- The path length between long-range dependencies in the network

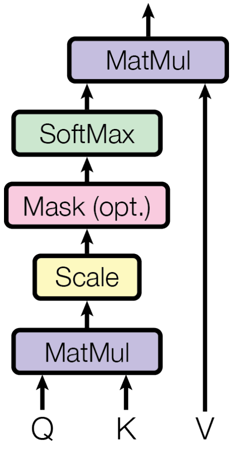
Vaswani et al., (2017)

其他实现技巧

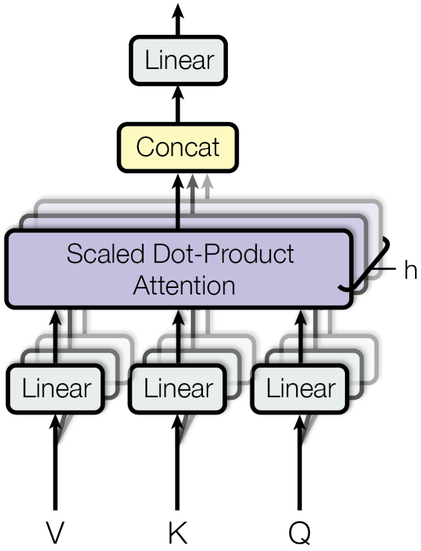
- Position, Layer Norm, Feed-Forward, Residual
- Scaled Dot-Product, Multi-Head
- Stacking



Scaled Dot-Product Attention



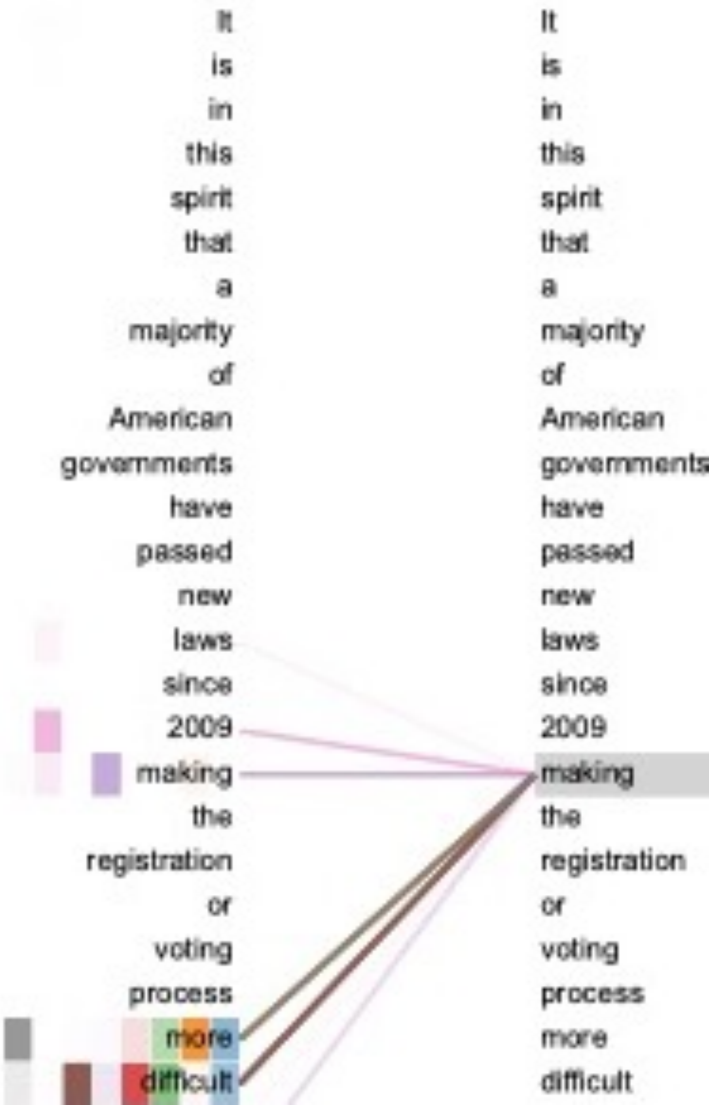
Multi-Head Attention



http://jalammar.github.io/illustrated-transformer/

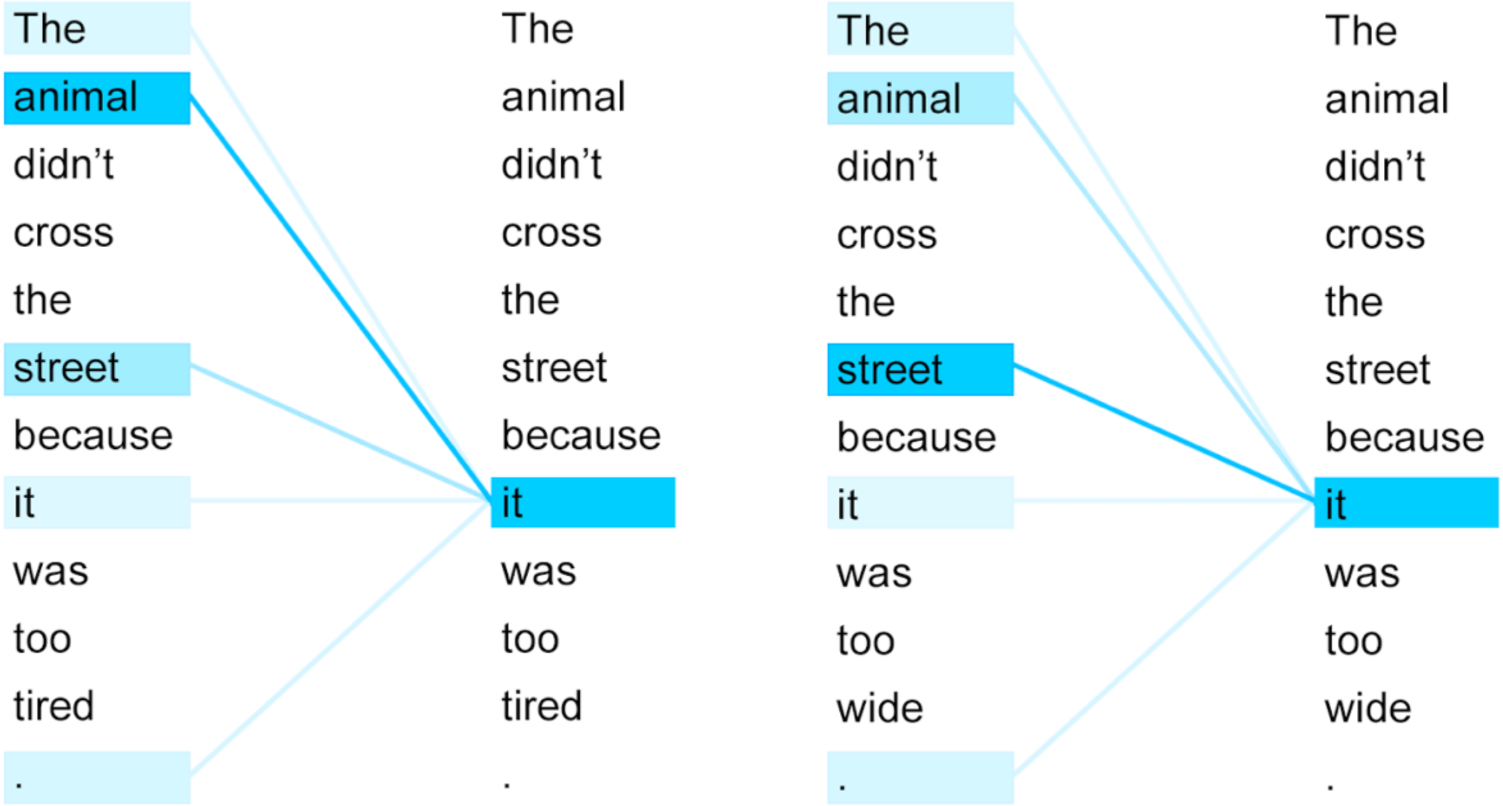
Self-Attention

- 捕捉相关的上下文



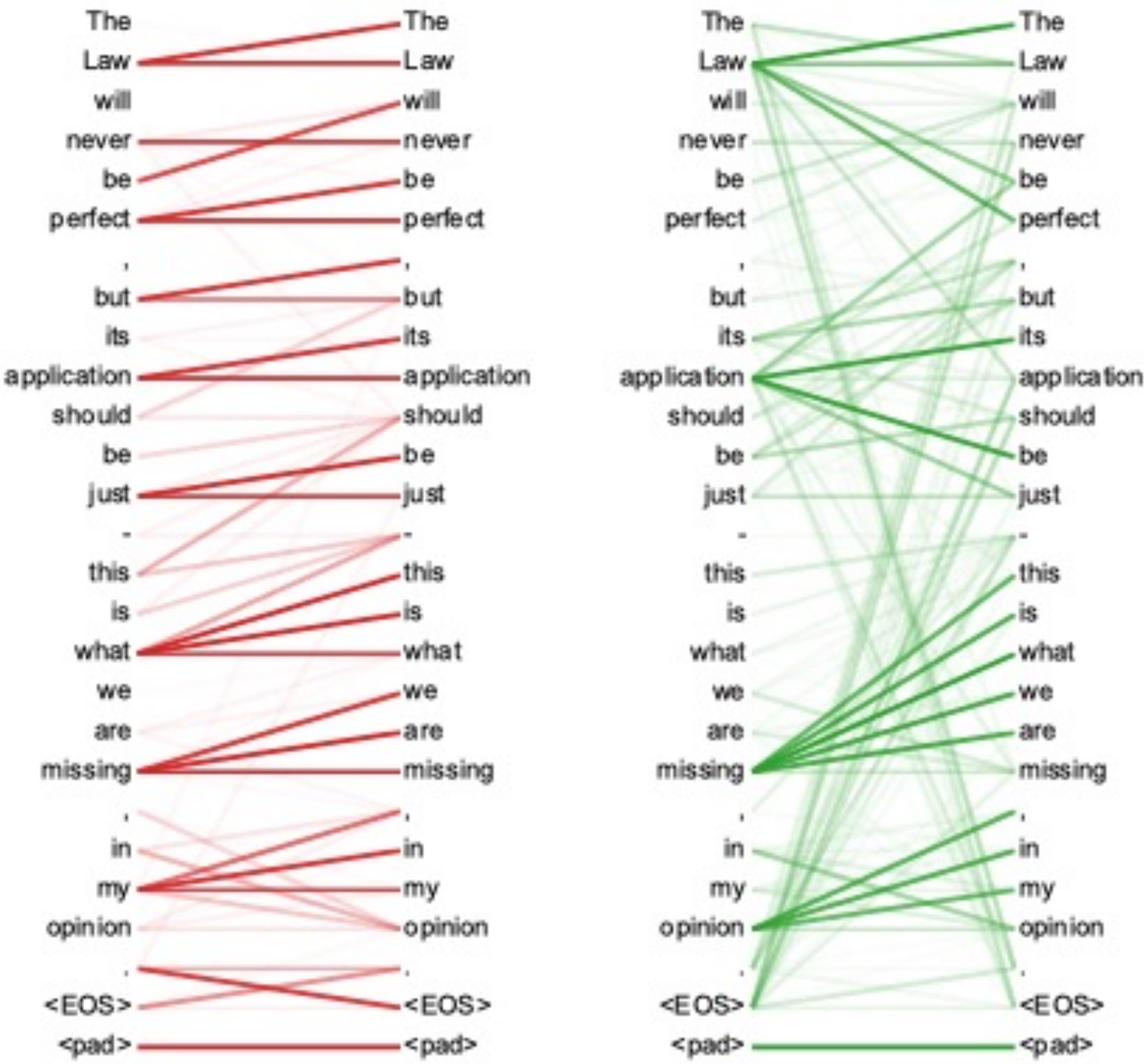
The animal didn't cross the street because it was too tired.
L'animal n'a pas traversé la rue parce qu'il était trop fatigué.

The animal didn't cross the street because it was too wide.
L'animal n'a pas traversé la rue parce qu'elle était trop large.



https://ai.googleblog.com/2017/08/transformer-novel-neural-network.html

Multi-Head Attention



小结

- 机器翻译能力随着机器计算能力的迅速发展而增长
  - 神经网络的引入从统计稀疏性和建模两个方面提升了机器翻译系统
  - 神经网络机器翻译是一种能够更加充分发挥机器长处的自动翻译方法
- 进一步的提升?

参考文献

- Wei Xu and Alex Rudnicky. Can artificial neural networks learn language models? In International Conference on Statistical Language Processing, 2000.
- Yoshua Bengio, Pascal Vincent, and Christian Janvin. A neural probabilistic language model. Journal of Machine Learning Research, 3(6):1137–1155, 2003.
- Ashish Vaswani, Yinggong Zhao, Victoria Fossum, and David Chiang. Decoding with large-scale neural language models improves translation. In Proceedings of the 2013 Con- ference on Empirical Methods in Natural Language Processing, pages 1387–1392, Seattle, Washington, USA, October 2013. Association for Computational Linguistics.

参考文献

- Michael Auli and Jianfeng Gao. Decoder integration and expected bleu training for re- current neural network language models. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 136–142, Baltimore, Maryland, June 2014. Association for Computational Linguistics.
- Yinggong Zhao, Shujian Huang, Huadong Chen, and Jiajun Chen, An Investigation on Statistical Machine Translation with Neural Language Models, CCL and NLP-NABD 2014, pp. 175–186, October 18-19， 2014, Wuhan, China
- Holger Schwenk. Continuous space translation models for phrase-based statistical machine translation. In Proceedings of COLING 2012: Posters, pages 1071–1080, Mumbai, India, December 2012. The COLING 2012 Organizing Committee.
- Jianfeng Gao, Xiaodong He, Wen-tau Yih, and Li Deng. Learning continuous phrase representations for translation modeling. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 699–709. Association for Computational Linguistics, 2014.

参考文献

- Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua 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), pages 1724–1734. 2014.
- Jiajun Zhang, Shujie Liu, Mu Li, Ming Zhou, and Chengqing Zong. Bilingually-constrained phrase embeddings for machine translation. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 111–121. 2014.
- Haiyang Wu, Daxiang Dong, Xiaoguang Hu, Dianhai Yu, Wei He, Hua Wu, Haifeng Wang, and Ting Liu. Improve statistical machine translation with context-sensitive bilingual semantic embedding model. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 142–146. 2014.
- Jinsong Su, Deyi Xiong, Biao Zhang, Yang Liu, Junfeng Yao, and Min Zhang. Bilingual correspondence recursive autoencoder for statistical machine translation. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1248– 1258. 2015.

参考文献

- Ke M. Tran, Arianna Bisazza, and Christof Monz. Word translation prediction for morphologically rich languages with bilingual neural networks. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1676– 1688. Association for Computational Linguistics, 2014.
- Peng Li, Yang Liu, Maosong Sun, Tatsuya Izuha, and Dakun Zhang. A neural reordering model for phrase-based translation. In Proceedings of COLING 2014, the 25th Inter- national Conference on Computational Linguistics: Technical Papers, pages 1897–1907, Dublin, Ireland, August 2014. Dublin City University and Association for Computational Linguistics.
- Antonio Valerio Miceli-Barone and Giuseppe Attardi. Non-projective dependency-based pre-reordering with recurrent neural network for machine translation. In The 53rd Annual Meeting of the Association for Computational Linguistics and The 7th International Joint Conference of the Asian Federation of Natural Language Processing, 2015.
- Yiming Cui, Shijin Wang, and Jianfeng Li. Lstm neural reordering feature for statistical machine translation. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 977–982, San Diego, California, June 2016. Association for Computational Linguistics.

参考文献

- Christian Hadiwinoto and Hwee Tou Ng. A dependency-based neural reordering model for statistical machine translation. In AAAI, pages 109–115, 2017.
- Jacob Devlin, Rabih Zbib, Zhongqiang Huang, Thomas Lamar, Richard Schwartz, John Makhoul， Fast and Robust Neural Network Joint Models for Statistical Machine Translation. ACL2014. Best paper award.
- Shujian Huang, Huadong Chen, Xin-Yu Dai, and Jiajun Chen. Non-linear learning for statistical machine translation. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on NaturalLanguage Processing (Volume 1: Long Papers), pages 825–835, Beijing, China, July 2015. Association for Computational Linguistics.

- Kalchbrenner and Blunsom. Recurrent Continuous Translation Models. EMNLP2013
- Sutskever Ilya, Vinyals Oriol, Le Quoc V.. Sequence to Sequence Learning with Neural Networks. NIPS 2014
- Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. ICLR 2015
- Minh-Thang Luong, Hieu Pham, Christopher D. Manning. Effective Approaches to Attention-based Neural Machine Translation. EMNLP2015.
- Thang Luong, Kyunghyun Cho, Christopher Manning. Neural Machine Translation. ACL tutorial. 2016