Sequence to Sequence Learning with Neural Networks

Ilya Sutskever, Oriol Vinyals, Quoc V. Le (2014)

NLP reading club 2021/06/10

Key Points

- Encoder - Decoder architecture based on LSTMs to transform input sequence to output sequence
- New training technique: reversed input sequence
- A deeper stack of LSTMs give better results
- Achieved BLEU score of 34.8 on English-French translation task

Dataset

- A clean subset of the WMT'14 English to French dataset. (348M French words and 304M English words)
- Fixed vocab: 160,000 of the most frequent words for the source language, 80,000 of the most frequent words for the target language. All other works replaced with 'UNK' token
- Start/end of sentence denoted by <START> and <EOS> tokens

Model structure

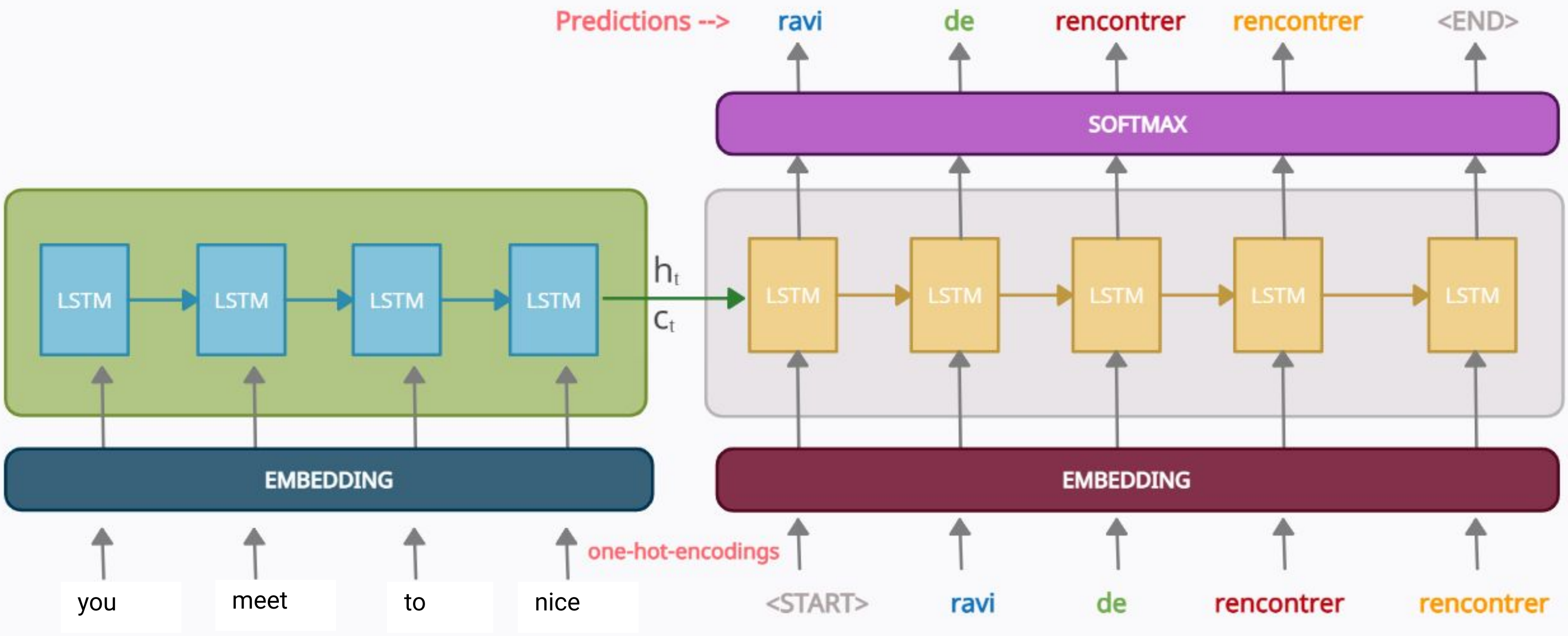
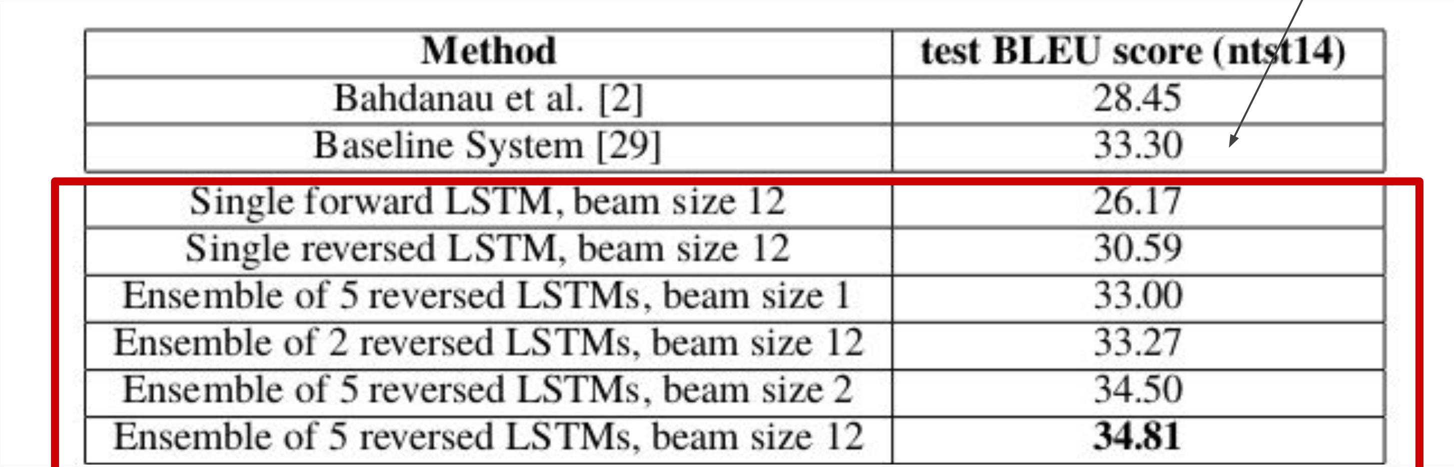


Image courtesy
https://medium.com/analytics-vidhya/encoder-decoder-seq2seq-models-clearly-explained-c34186fbf49b

Experiment results

Translation English to French:

SMT baseline



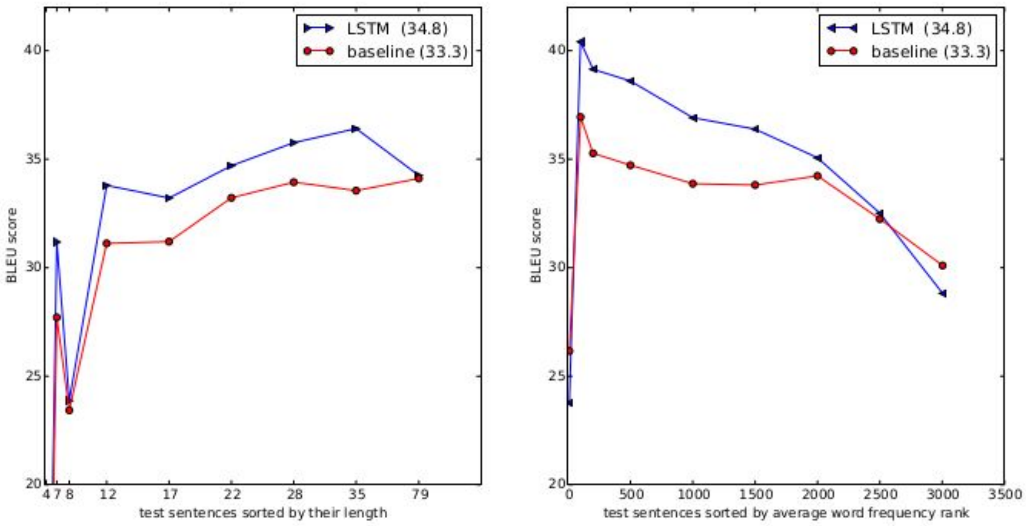
Method	test BLEU score (ntst14)
Bahdanau et al. [2]	28.45
Baseline System [29]	33.30
Single forward LSTM, beam size 12	26.17
Single reversed LSTM, beam size 12	30.59
Ensemble of 5 reversed LSTMs, beam size 1	33.00
Ensemble of 2 reversed LSTMs, beam size 12	33.27
Ensemble of 5 reversed LSTMs, beam size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

Results from paper

Experiment results

Sentence length:

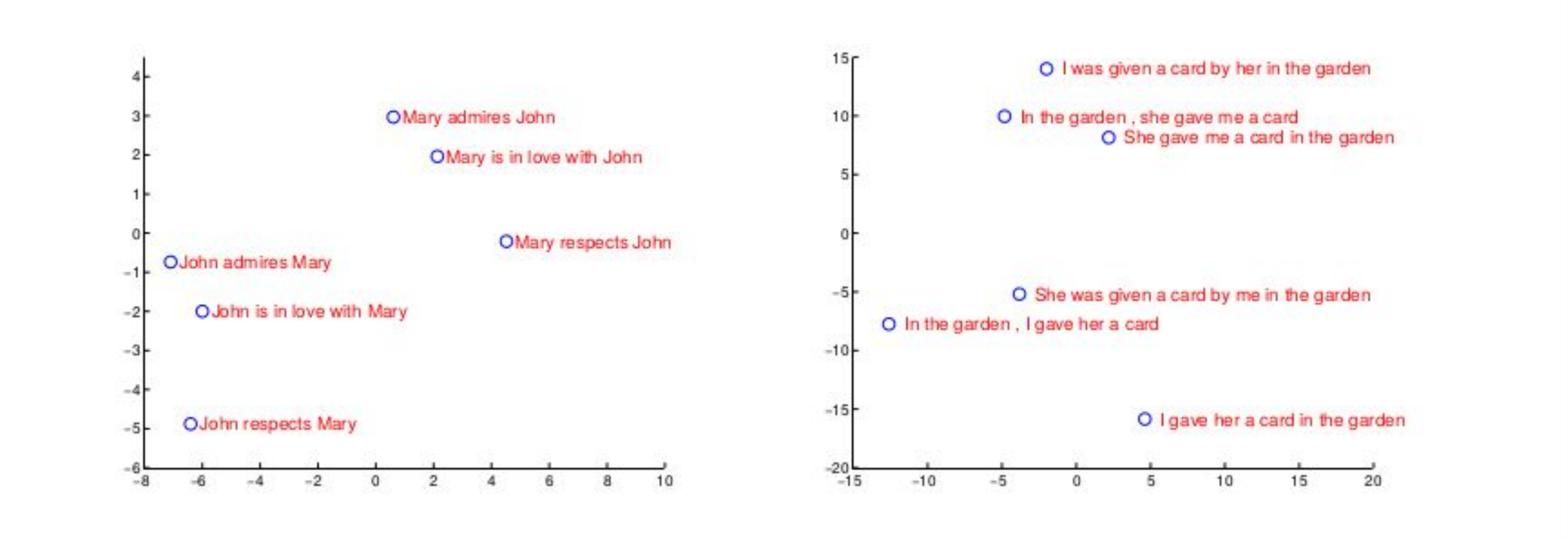
- Works well even for longer sentences
- Does less well for sentences with rare words, but still better than the SMT baseline



Experiment results

Word ordering

- Model captures meaning conveyed by word ordering
- Diagram shows PCA clustering of LSTM hidden states-- the sentences in which the subject/object are reversed are more separated than those in which they are not. This is not something e.g. bag of words can capture



Aside: Bleu score

- a metric for automatically evaluating machine-translated text
- Measures how similar the translated text output from the model is to the ‘ground truth translation’, runs from 0 (very poor translation) to 1 (very good translation)
- Similarity using the 1,2,3,4 -gram overlap + penalty for sentences that are too short

A good explanation can be found here

Further resources

- Link to original paper
- Jay alammar's blog post on Seq2Seq models
- Chris Olah's post on LSTMs
- Code walkthrough of LSTM
- Kritz Moses' blog post covering this paper specifically
- Oral presentation of the paper at Neurips 2014