# Sequence to Sequence Learning with Neural Networks

- Sutskever et al. 2014

Xinyu Zhou

March 15, 2018

# Encoder-Decoder

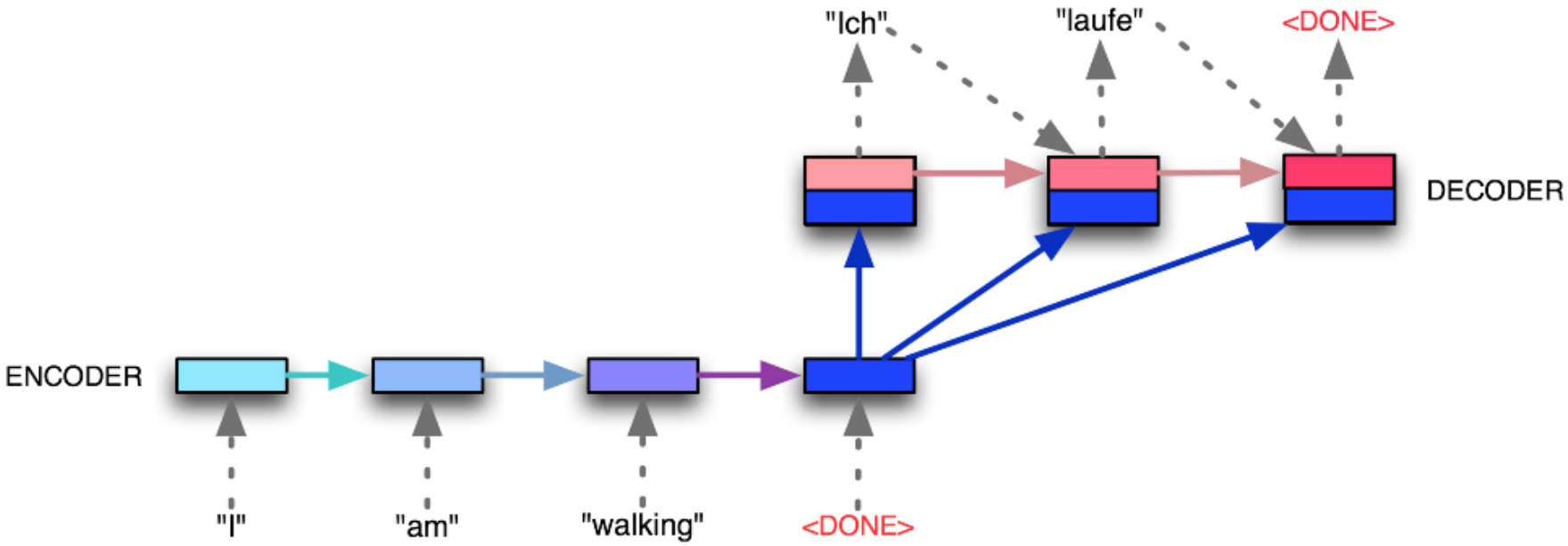



Figure: https://towardsdatascience.com/sequence-to-sequence-using-encoder-decoder-15e579c10a94

# 3-Line Summarize

- LSTM Is Better Than RNN
- Deep Is Better Than Shallow
- Reversed Is Better Than Original

# Notice

- This is NOT a paper for attention
- This is (one of) the FIRST paper for sequence-to-sequence in machine translation
- This paper claims neural MT can OUTPERFORM statistical MT

| 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** |

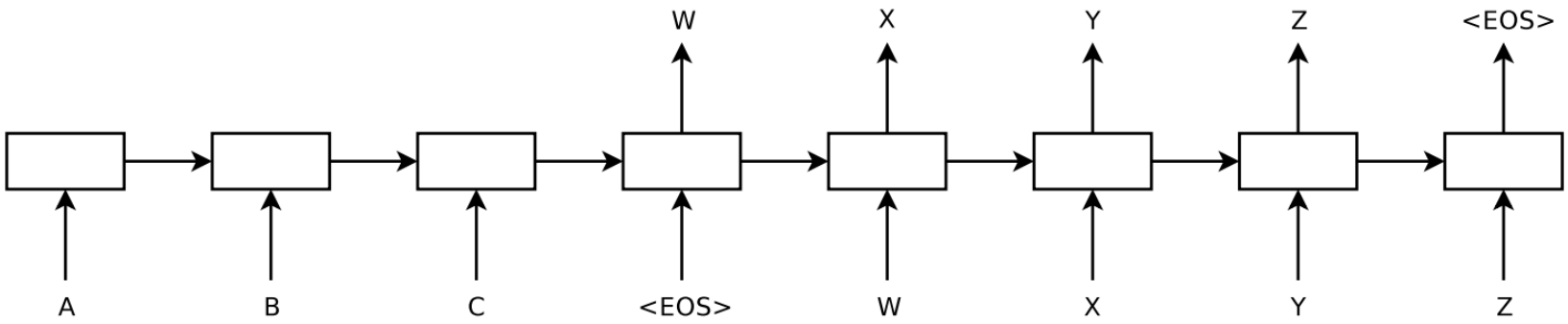
# Introduction

# Introduction

- Deep Neural Network (DNN) is powerful
- DNN can't solve sequential problems (input size is unknown)
- RNN can solve sequential problems
- RNN can also solve sequence to sequence problems
- How? Encoder-Decoder with LSTM

# Introduction

- Encoder: Map input sentence into fixed dimensional vector
- Decoder: Language model, conditioned with input sequence
- LSTM: Learn long range temporal dependencies
- Input $->$ Encoder $->$ Vector $->$ Decoder $->$ Sequence of Words



# Encoder-Decoder

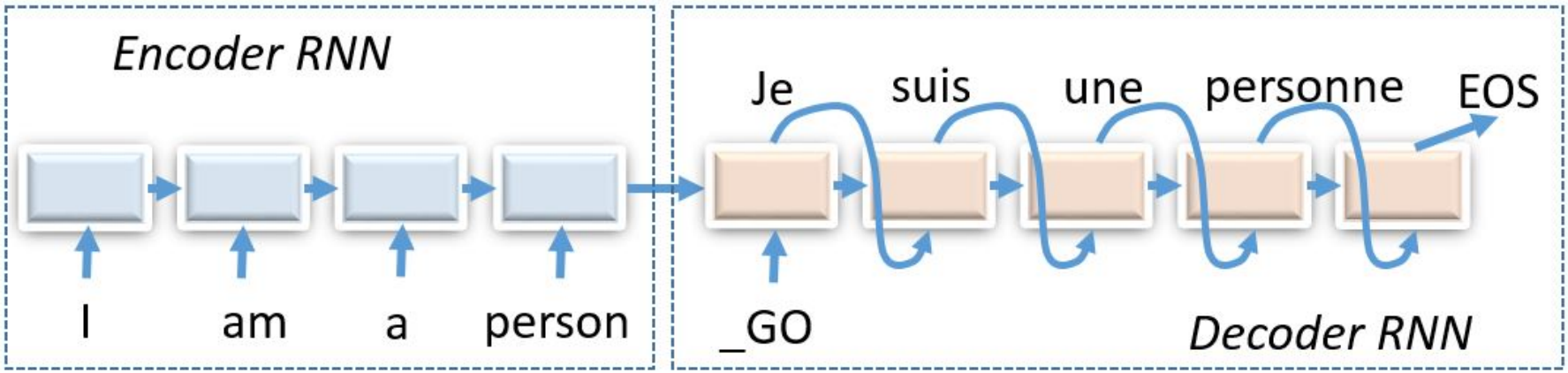



Figure: https://esciencegroup.com/2016/03/04/fun-with-recurrent-neural-nets-one-more-dive-into-cntk-and-tensorflow/

# Introduction

- BLEU on WMT' 14: 34.81 vs 33.30 (SMT baseline)
- BLEU for rescore 1000-best lists of the SMT baseline: 36.5, close to current best (37.0)

| Method | test BLEU score (ntst14) |
|---|---|
| Baseline System [29] | 33.30 |
| Cho et al. [5] | 34.54 |
| Best WMT'14 result [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 |

# Introduction

- LSTM did not suffer on very long sentences

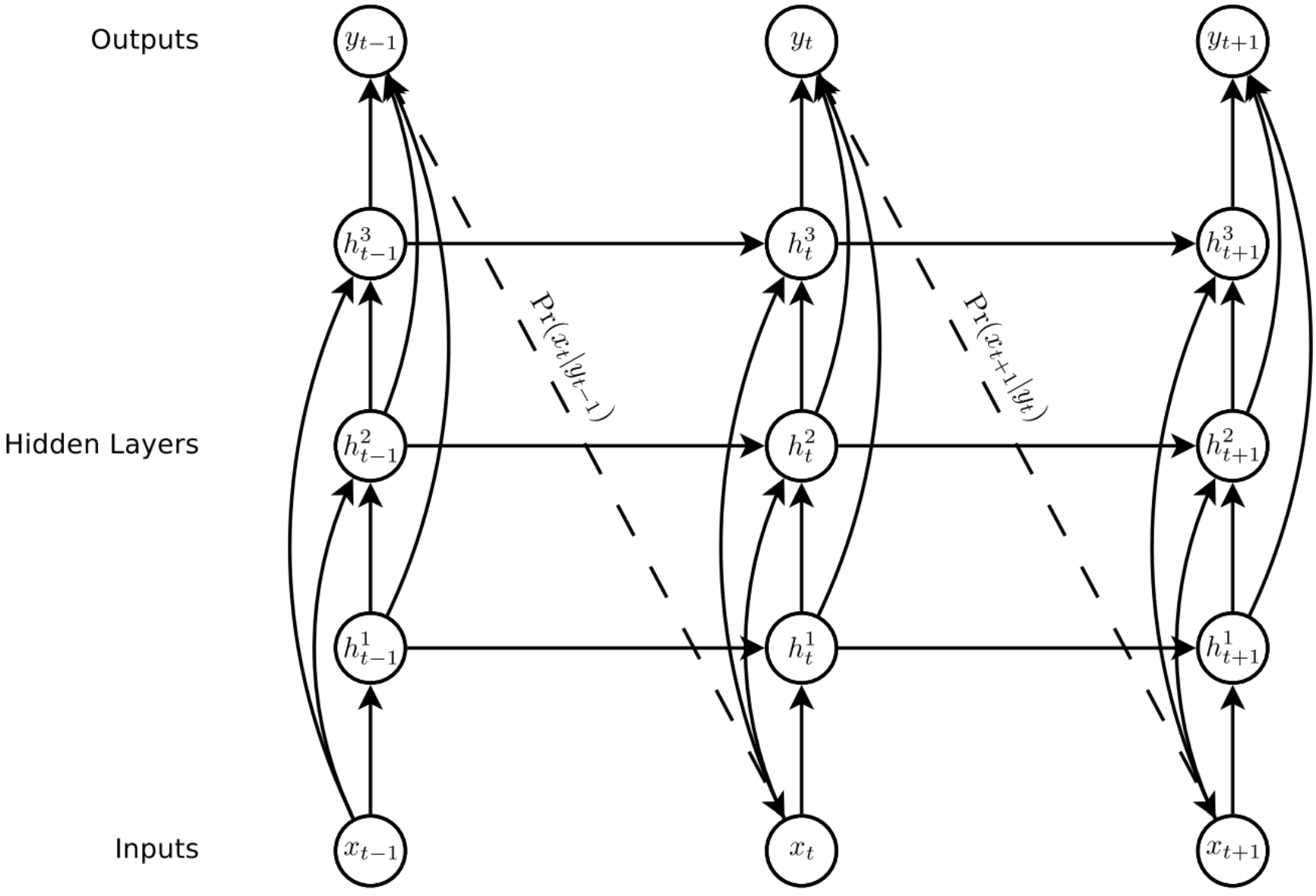
| Type | Sentence |
|---|---|
| **Our model** | Ulrich UNK , membre du conseil d' administration du constructeur automobile Audi , affirme qu' il s' agit d' une pratique courante depuis des années pour que les téléphones portables puissent être collectés avant les réunions du conseil d' administration afin qu' ils ne soient pas utilisés comme appareils d' écoute à distance . |
| **Truth** | Ulrich Hackenberg , membre du conseil d' administration du constructeur automobile Audi , déclare que la collecte des téléphones portables avant les réunions du conseil , afin qu' ils ne puissent pas être utilisés comme appareils d' écoute à distance , est une pratique courante depuis des années . |
| **Our model** | " Les téléphones cellulaires , qui sont vraiment une question , non seulement parce qu' ils pourraient potentiellement causer des interférences avec les appareils de navigation , mais nous savons , selon la FCC , qu' ils pourraient interférer avec les tours de téléphone cellulaire lorsqu' ils sont dans l' air " , dit UNK . |
| **Truth** | " Les téléphones portables sont véritablement un problème , non seulement parce qu' ils pourraient éventuellement créer des interférences avec les instruments de navigation , mais parce que nous savons , d' après la FCC , qu' ils pourraient perturber les antennes-relais de téléphonie mobile s' ils sont utilisés à bord " , a déclaré Rosenker . |
| **Our model** | Avec la crémation , il y a un " sentiment de violence contre le corps d' un être cher " , qui sera " réduit à une pile de cendres " en très peu de temps au lieu d' un processus de décomposition " qui accompagnera les étapes du deuil " . |
| **Truth** | Il y a , avec la crémation , " une violence faite au corps aimé " , qui va être " réduit à un tas de cendres " en très peu de temps , et non après un processus de décomposition , qui " accompagnerait les phases du deuil " . |

# Model

# Model

- RNN is proved to be powerless on long term dependencies
- LSTM is not
- LSTM formulation from *Generating Sequences With Recurrent Neural Networks* - Graves, 2013

# Stacked-LSTM



# Model

- Three differences:
- Two LSTMs, one for encoder, one for decoder
- Deep LSTMs is better than shallow LSTMs, so use LSTM with four layers
- Reverse the order of input words
  Better results: PPL from 5.8 to 4.7, BLEU score from 25.9 to 30.6.
  ”This way, *a* is in close proximity to $\alpha$, *b* is fairly close to $\beta$, and so on, a fact that makes it easy for SGD to establish communication between the input and the output.”

# Experiment

# Experiment

- Some parameters you may not care
- WMT14 EN-FR
- Train on 12M sentences, 348M French words, 304M English words.
- Vocabulary: 160,000 ENG, 80,000 FR, ”UNK”
- 4 LSTM layers, 1000 dimensional embedding, 1000 dimensional hidden
- Init with uniform distribution [-0.08, 0.08]
- 7.5 epochs in total
- Learning rate: 0.7 for first 5 epochs, then halve for every half epoch
- Batch size 128
- Hard constraint for gradient: if $\|g\|_2 > 5$, set $g = \frac{5g}{\|g\|_2}$
- Put sentence with similar length together to speed up (2x)
- 8 GPU parallelization, 6300 words per sec, 10 days to train

# Experiment

- BLEU score

| 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.

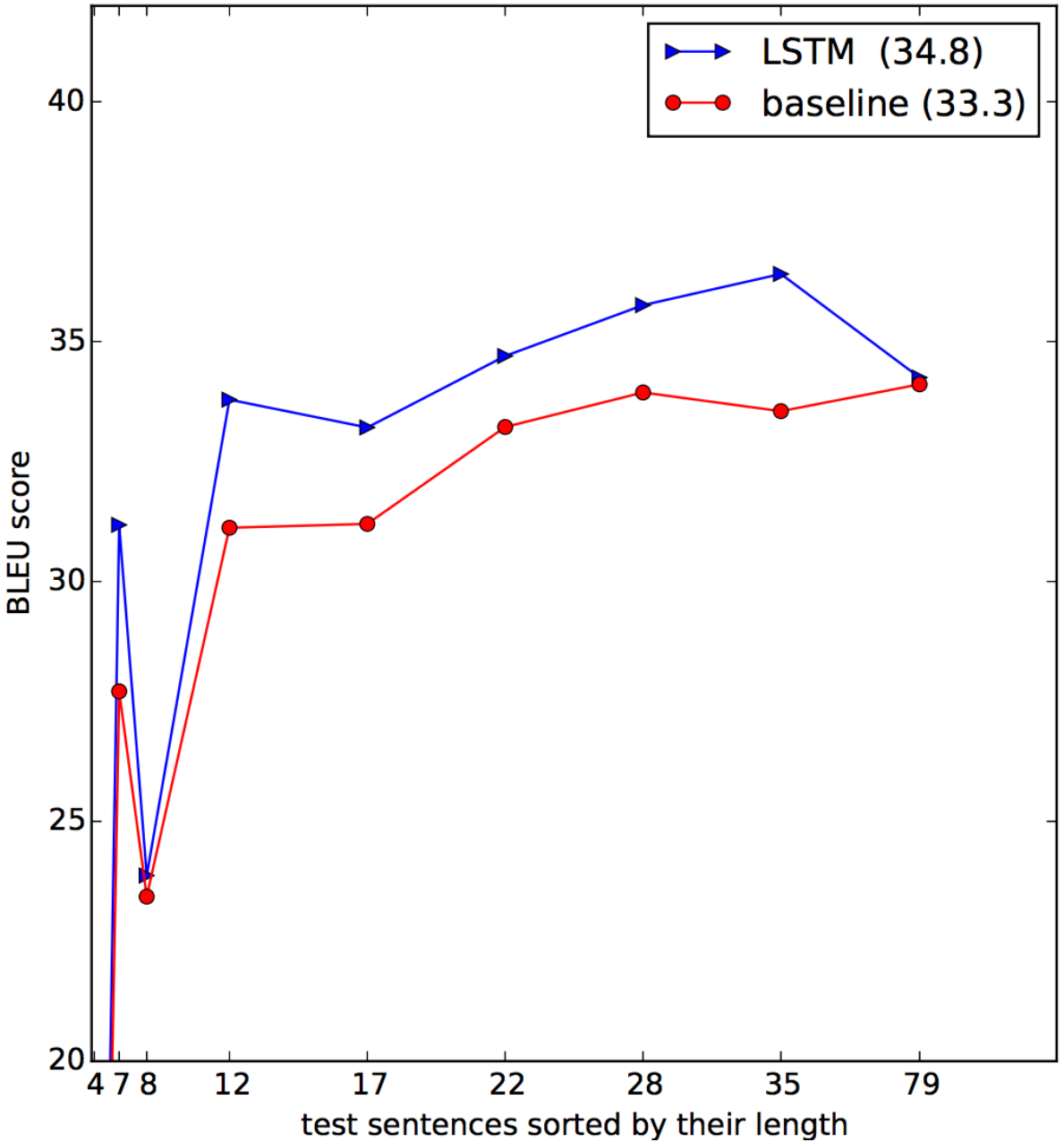
| Method | test BLEU score (ntst14) |
|---|---|
| Baseline System [29] | 33.30 |
| Cho et al. [5] | 34.54 |
| Best WMT'14 result [9] | **37.0** |
| Rescoring the baseline 1000-best with a single forward LSTM | 35.61 |
| Rescoring the baseline 1000-best with a single reversed LSTM | 35.85 |
| Rescoring 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).

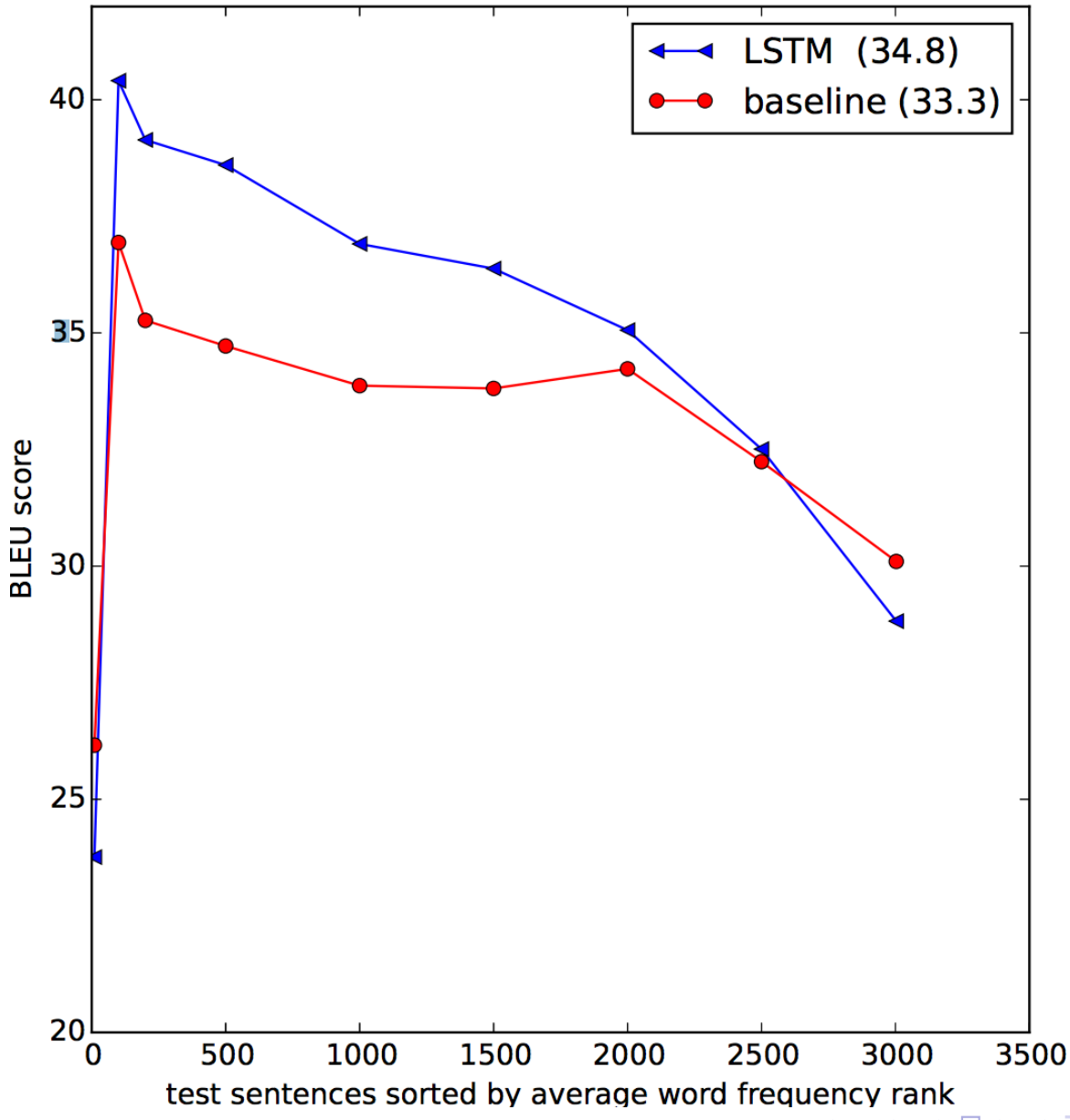
# Experiment

- Performance on longer sentence and less frequent words
- LSTM performs well on longer sentences
- LSTM performs better than SMT on rare words

# Test sentences sorted by length

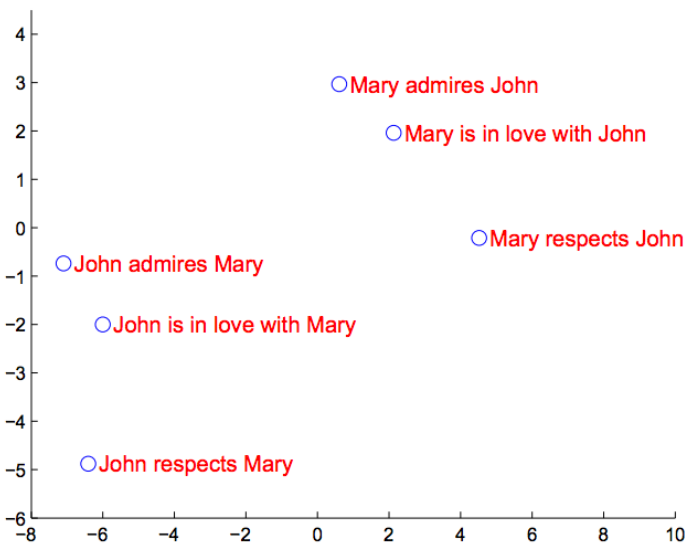


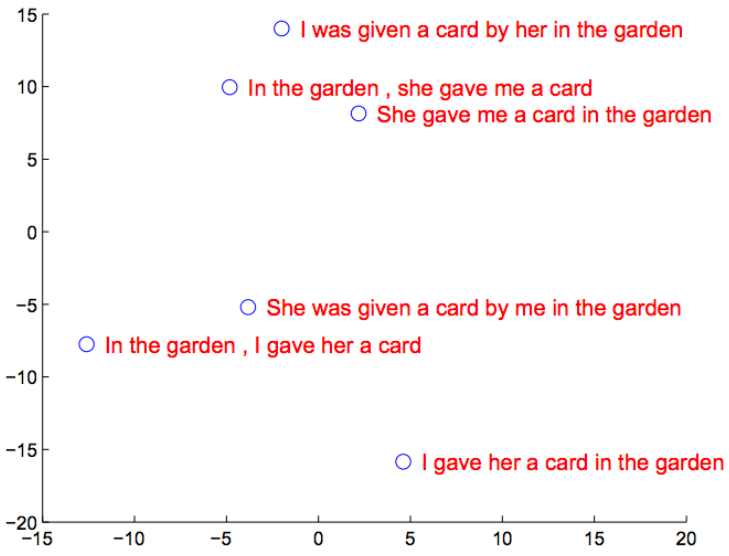
# Test sentences sorted by average word frequency rank



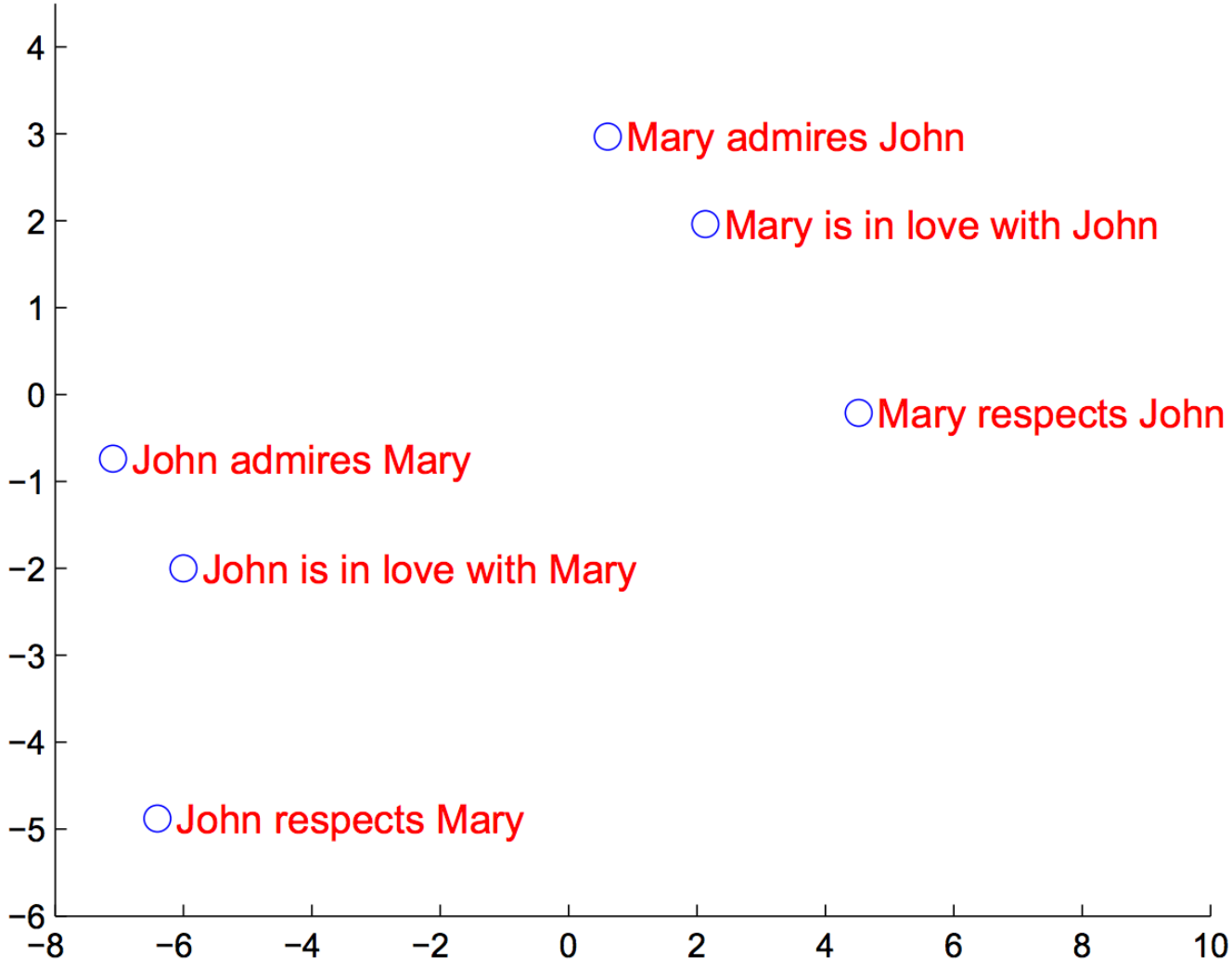
# Experiment

- PCA on vectored sentences
- Distinguish same word with different orders
- Recognize passive voice

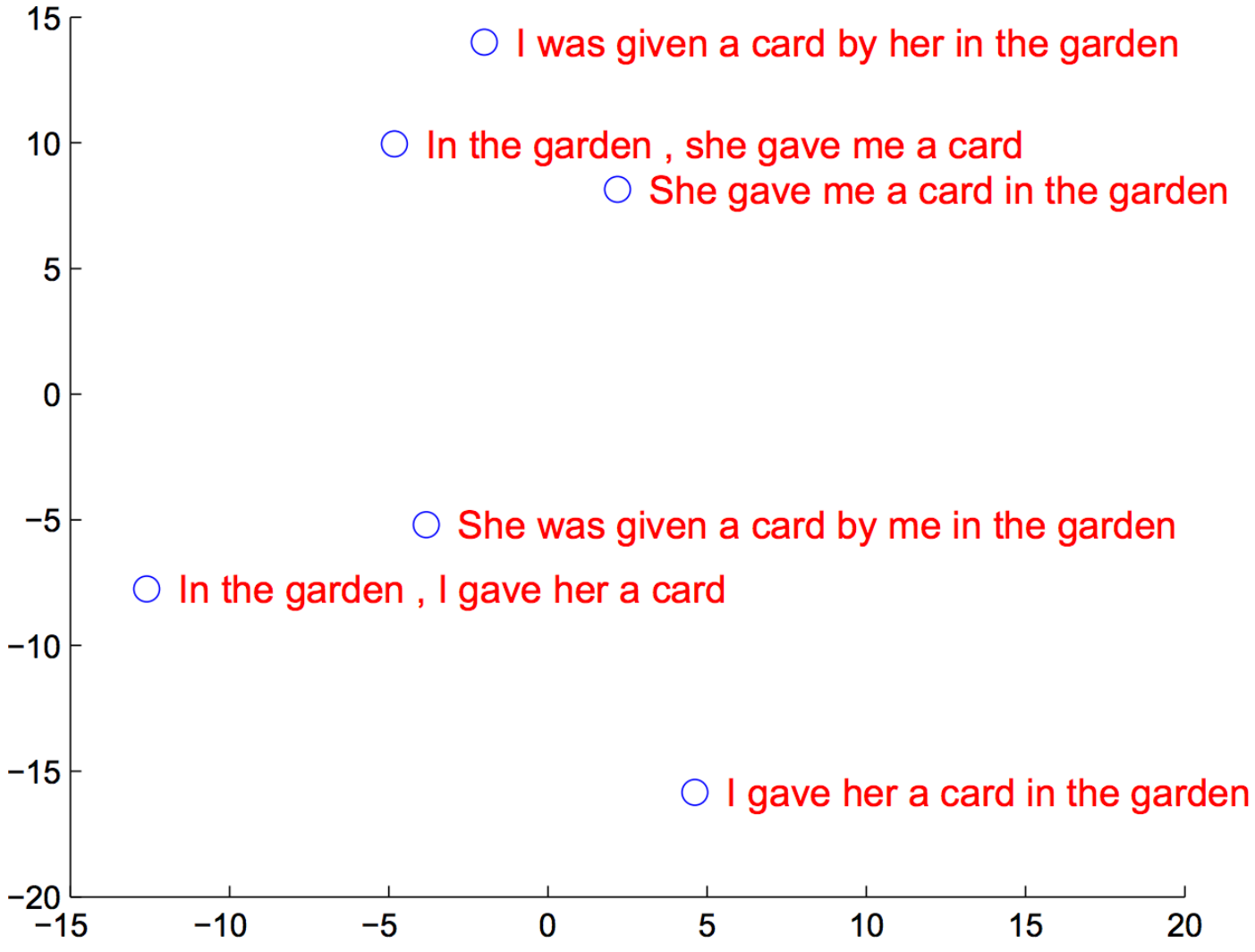


# Distinguish same word with different orders



# Recognize passive voice



# Conclusion

- NMT can do as well as SMT
- Three techniques: LSTM, Deeper, Reverse
- Suffer long sentences and rare words
- Good properties

# Thanks