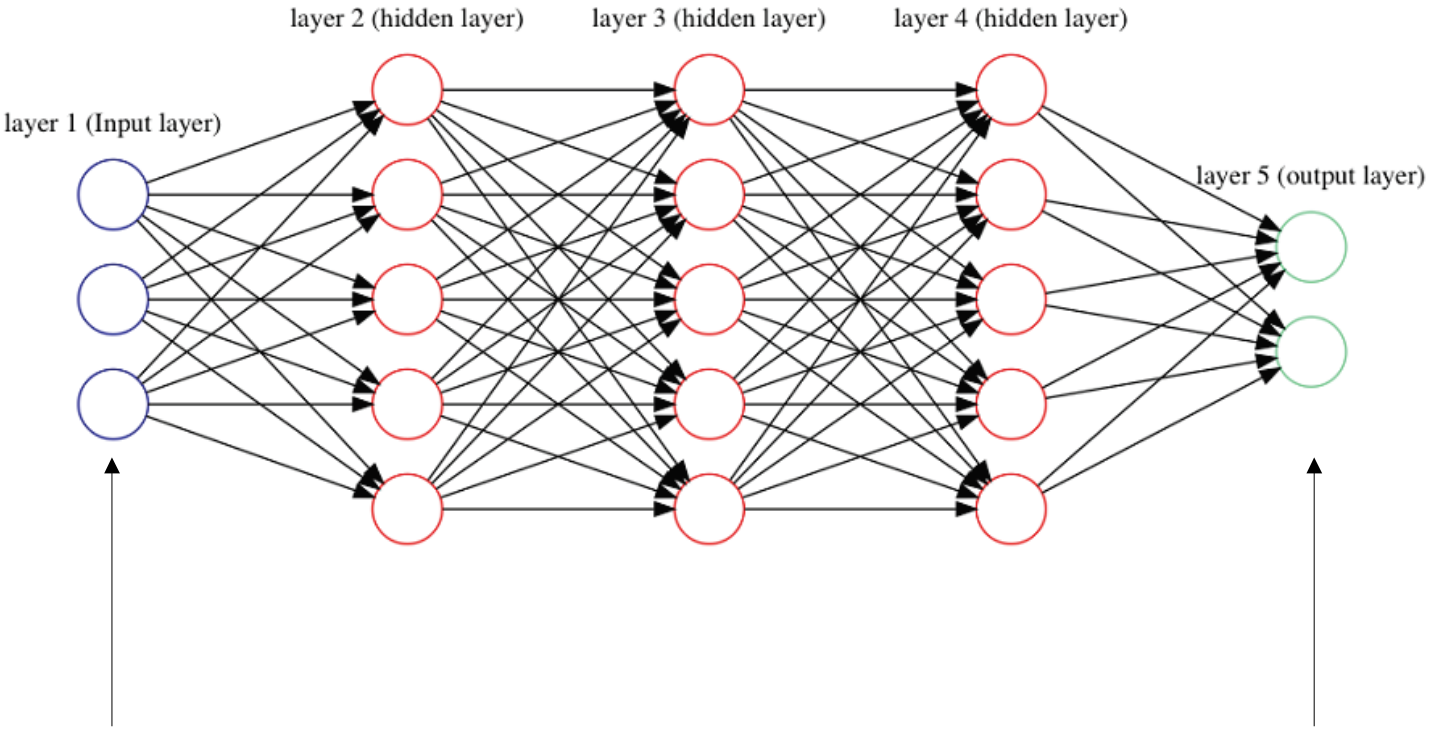
# Sequence to Sequence Learning with Neural Networks

Author: Ilya Sutskever, Oriol Vinyals, Quoc V. Le

# Introduction

- Problem: DNN cannot map sequences to sequences

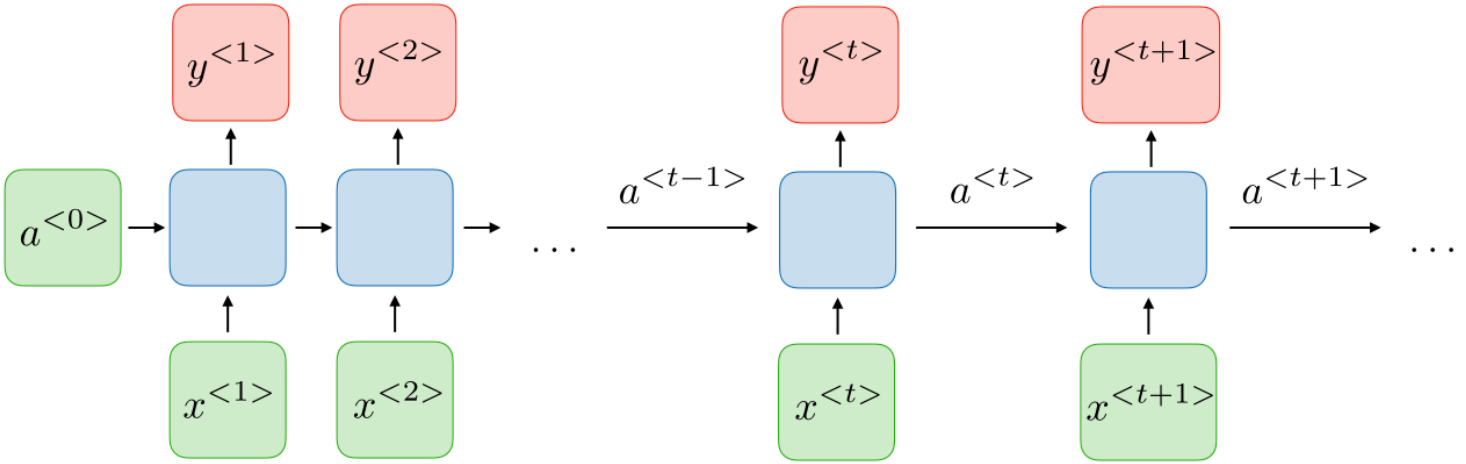



Fixed input & output, not suitable for problems where length of sequence are unknown.

**Source:** https://freecontent.manning.com/what-does-deep-learning-contribute-to-search/

# Introduction

- How about RNN?




- Can easily map if alignment is known beforehand
- Have one-to-one correspondence, problem with different lengths
- Trouble with long-range dependencies

**Source:** https://stanford.edu/~shervine/teaching/cs-230/cheatsheet-recurrent-neural-networks

# Introduction

- Goal of paper: Solve problem of mapping sequence to sequence
- Sequence to sequence learning is important for:
  - Machine Translation
  - Image caption generation
  - Question Answering
  - Text Summarization
  - Many other tasks

# Related Work

- Recurrent Continuous Translation Models (N. Kalchbrenner & P. Blunsom) → CNN for encoder, RNN for decoder
- Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation (Kyunghyun Cho et al.) → RNN encoder, RNN decoder for rescoring phrase pairs

# The Model

- Idea:
  - Composed of encoder-decoder
  - Map input sequence to a fixed-dimensional vector representation using one LSTM (encoder)
  - Map vector representation to target sequence using another LSTM (decoder)
  - Decoder LSTM is essentially language model conditioned on input sequence

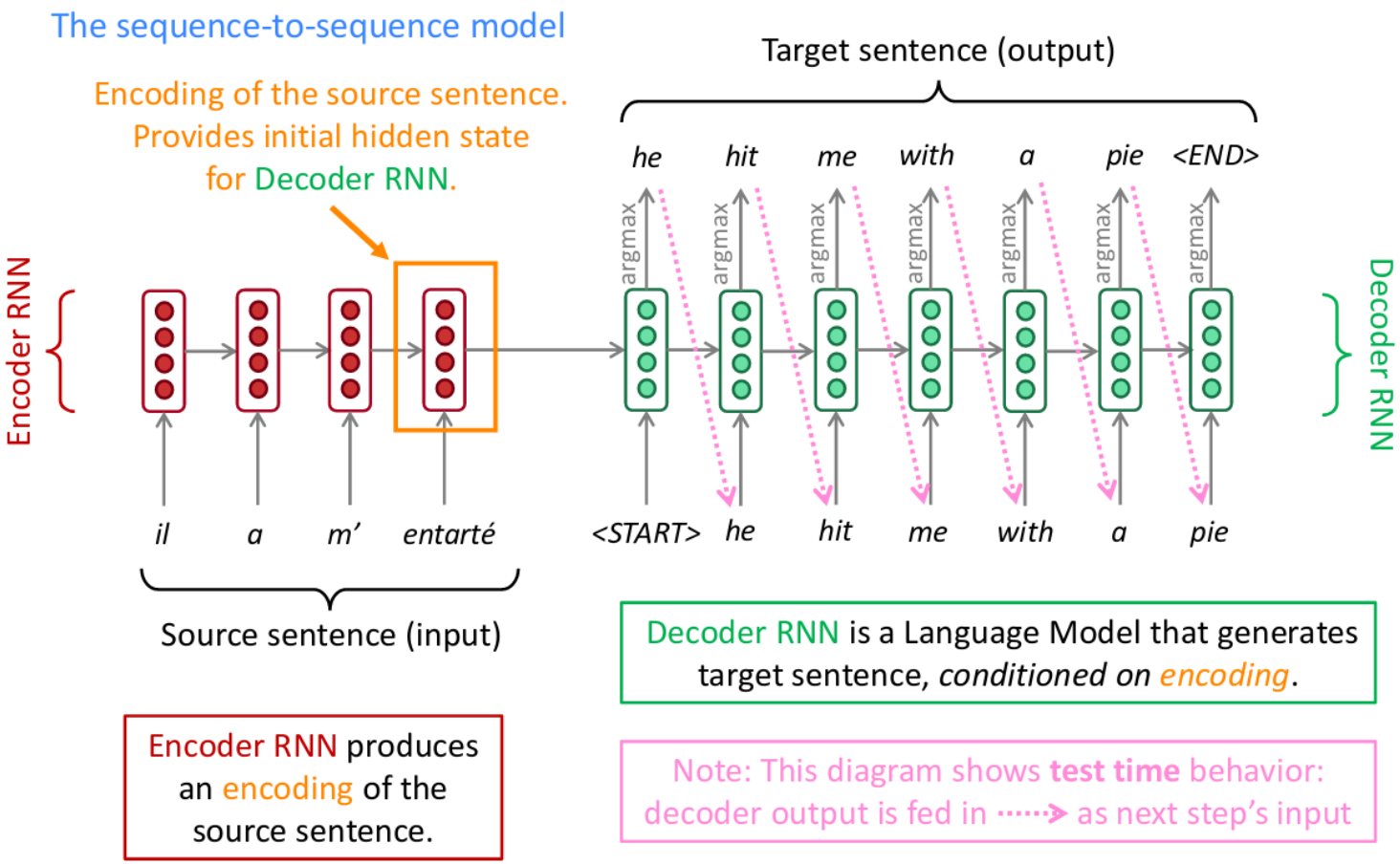
# The Model

- Goal: Estimate conditional probability of output sequence given input sequence

$$p(y_1, \ldots, y_{T'} | x_1, \ldots, x_T) = \prod_{t=1}^{T'} p(y_t | v, y_1, \ldots, y_{t-1})$$

- Distribution is represented with softmax over all words in the vocabulary

# The Model



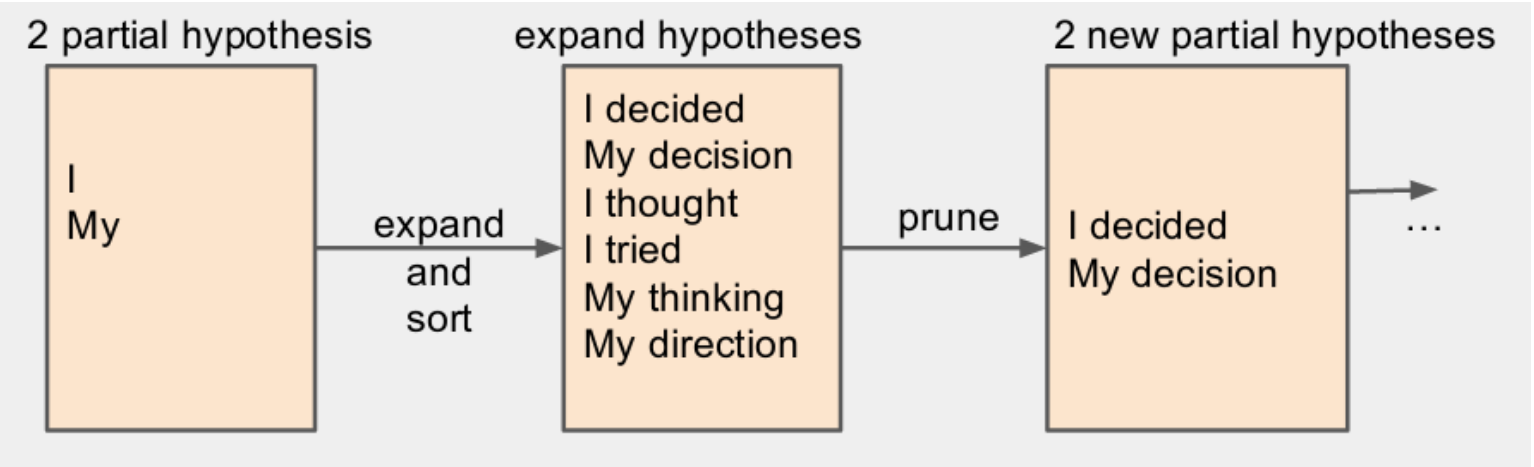

**Source:** https://web.stanford.edu/class/cs224n/slides/cs224n-2019-lecture08-nmt.pdf

# The Model

- Training objective:

$$1/|\mathcal{S}| \sum_{(T,S)\in\mathcal{S}} \log p(T|S)$$

- Search for most likely translation using left-to-right beam search decoder




**Source:** http://www.cips-cl.org/static/CCL2017/slides/T1_part2.pdf

# The Model

- Additional Details:
  - Seq2Seq is optimized as a single system
  - Input sequences are reversed to reduce “minimal time lag”
  - In experiment, deep LSTMs with 4 layers, 1000 cells and 1000 dimensional word embeddings are used.
  - Each layer of LSTM was executed on a different GPU, 4 remaining GPUs were used to parallelize softmax

# Experiments

- Dataset: WMT' 14 English to French MT
  - Directly translate input sentence w/o reference SMT system
  - Rescore n-best lists of an SMT baseline
- Vocabularies of 160k and 80k most frequent words for source and target language respectively
- OOV is replaced with "UNK" token

# Results

- BLEU score was used to evaluate quality of translations

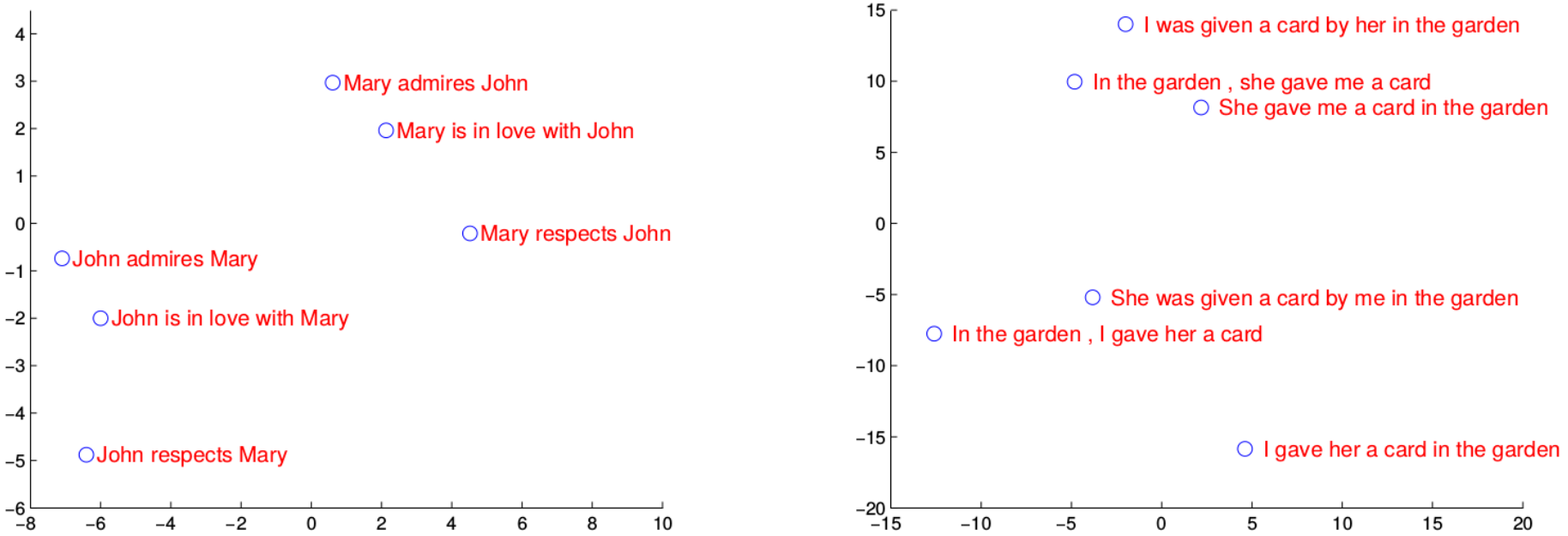
| Method | test BLEU score (ntst14) |
|---|---|
| Bahdanau et al. [2] | 28.45 |
| Baseline System [29] | 33.30 |
| Single forward LSTM, beam size 12 | 26.17 |
| Single reversed LSTM, beam size 12 | 30.59 |
| Ensemble of 5 reversed LSTMs, beam size 1 | 33.00 |
| Ensemble of 2 reversed LSTMs, beam size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

**LSTM**

| 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 |

**LSTM + SMT system**

# Results

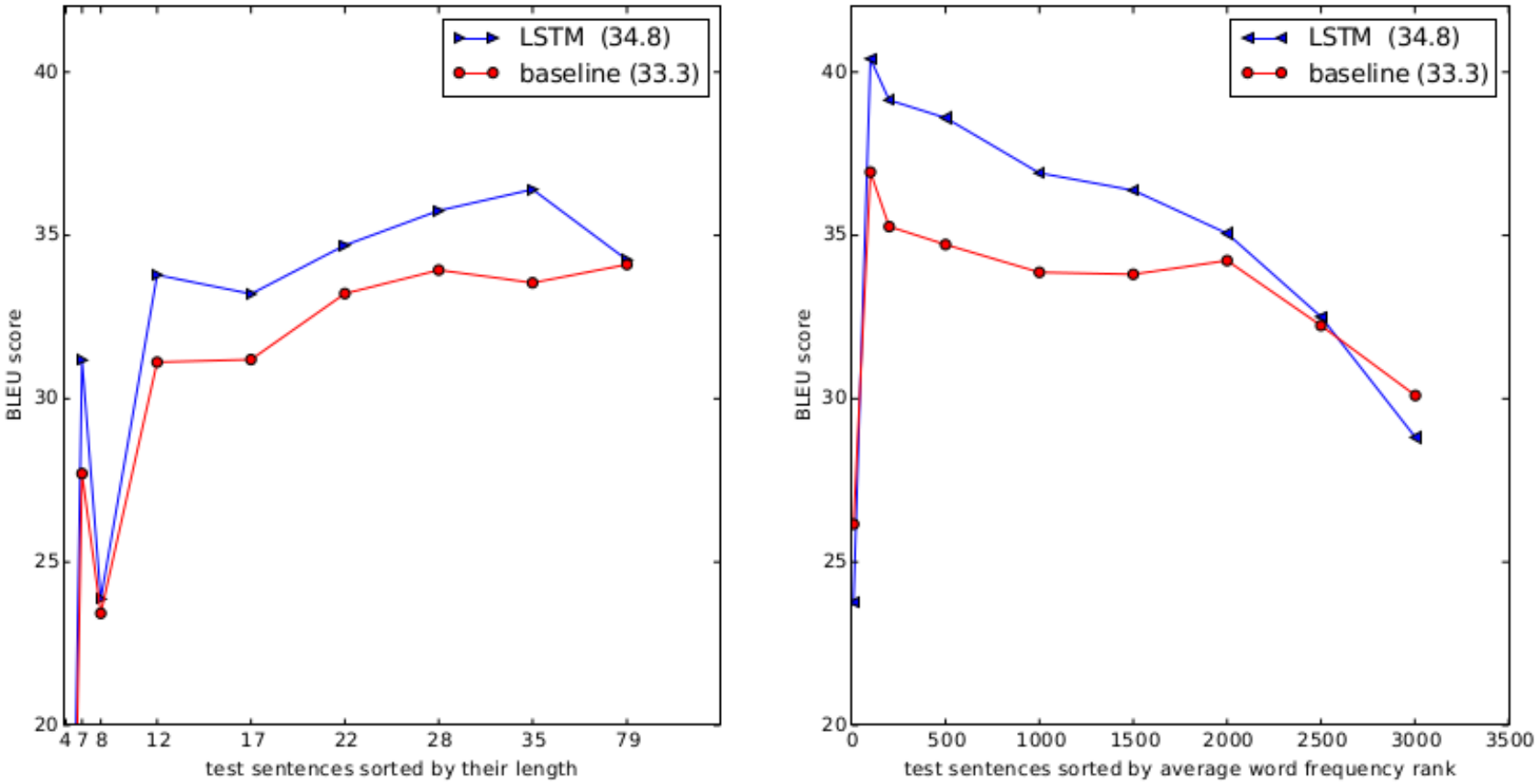



**2D PCA projection of LSTM hidden states after processing the phrase in the figures**

# Results

| 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 " . |

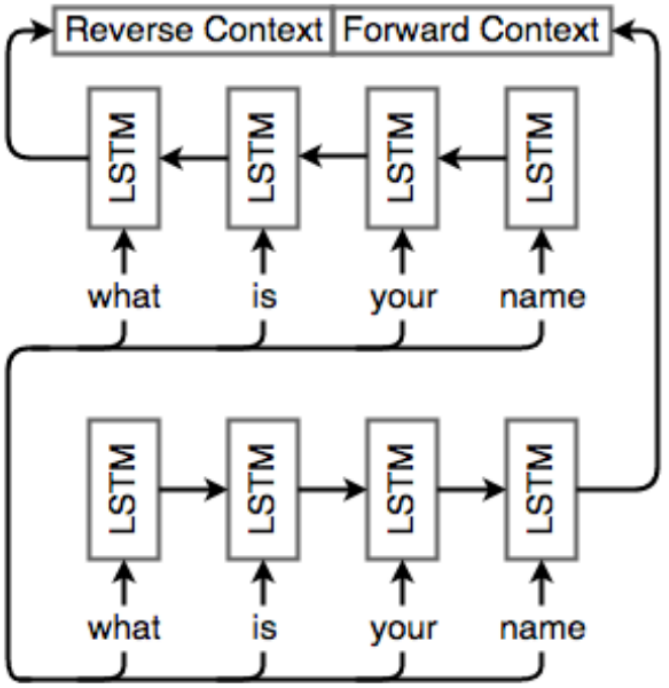
# Results




Left: LSTM performs quite well even on long sentences
Right: LSTM's performance degrade with more rare words

# Limitation with Seq2Seq

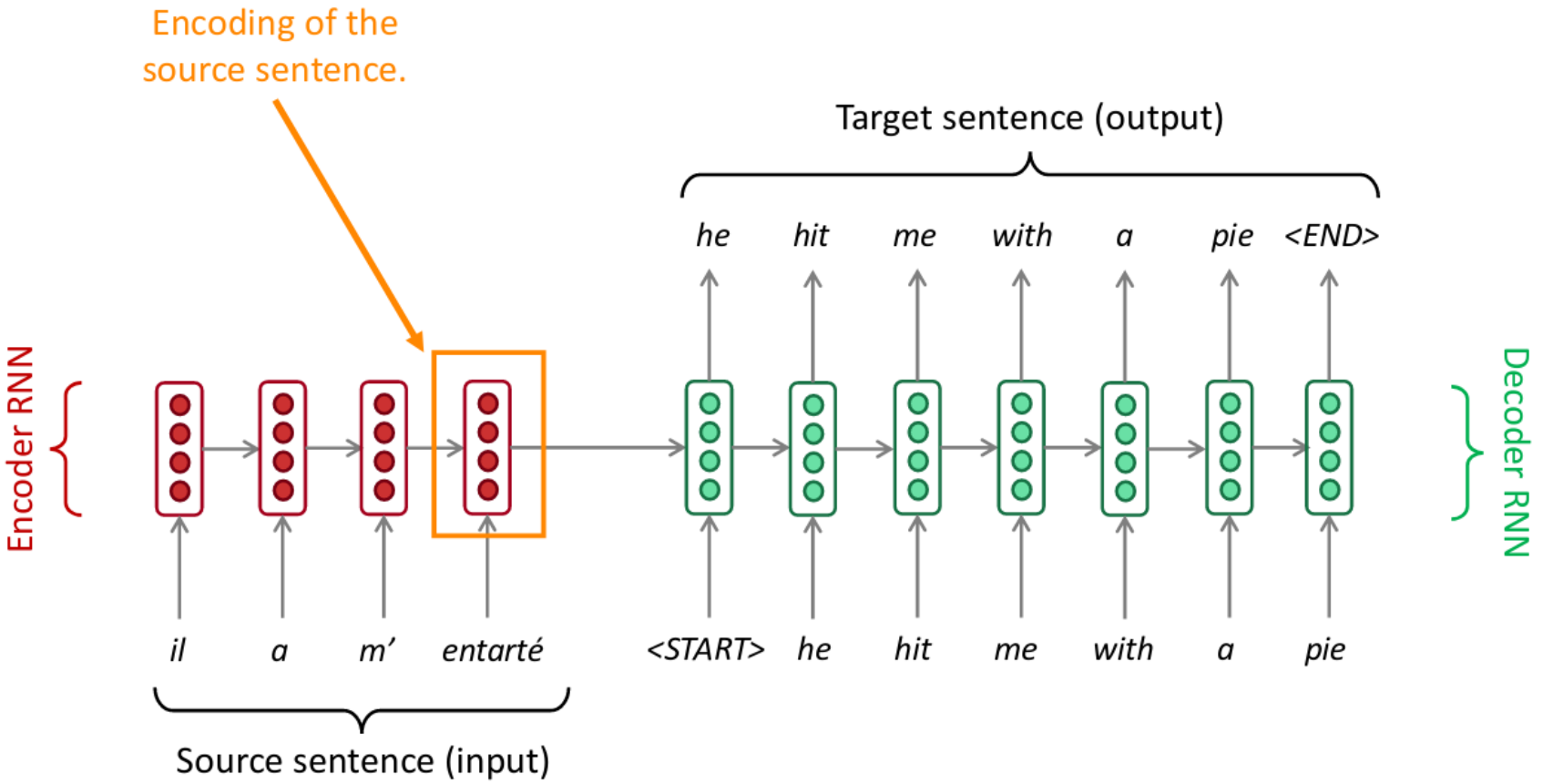
- Out of Vocabulary (OOV) problems
  - Addressing the Rare Word Problem in Neural Machine Translation (Minh-Thang Luong et. al) → follow-up work
- Information from words after current word?




**Source:** https://web.stanford.edu/class/cs224n/readings/cs224n-2019-notes06-NMT_seq2seq_attention.pdf

# Limitation with Seq2Seq

- Information bottleneck




Problems with this architecture?

**Source:** https://web.stanford.edu/class/cs224n/slides/cs224n-2019-lecture08-nmt.pdf