Learning Phrase Representations using RNN Encoder-decoder for Statistical Machine Translation

Cho et al. (2014)

Ana Valeria Gonzalez
Postdoc at University of Copenhagen
AI Research Scientist at Halfspace

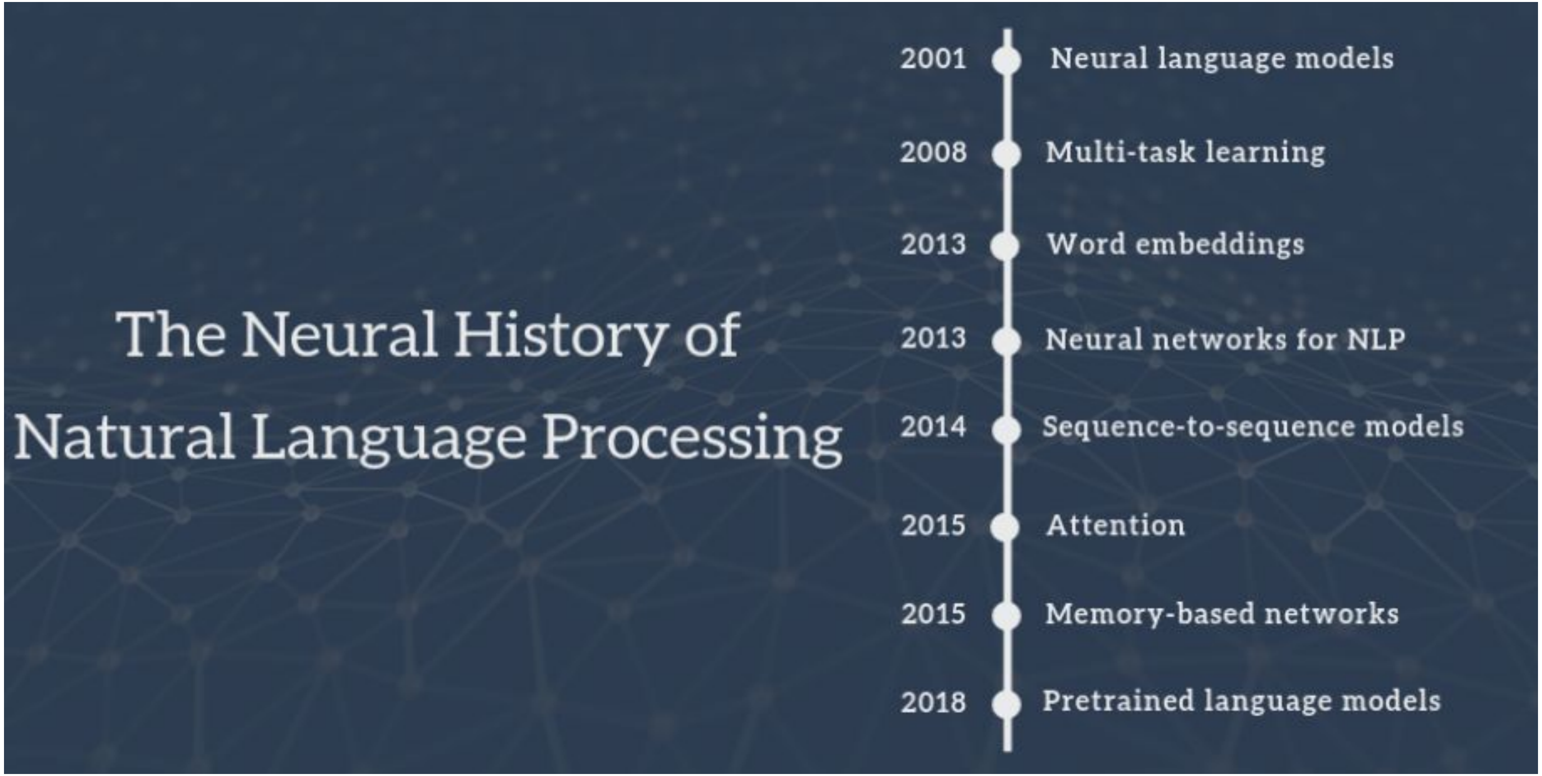
Outline

1. Motivation for choice
2. Preliminary on RNNs
3. GRU vs LSTM
4. Cho's experiments
5. Discussion

Outline

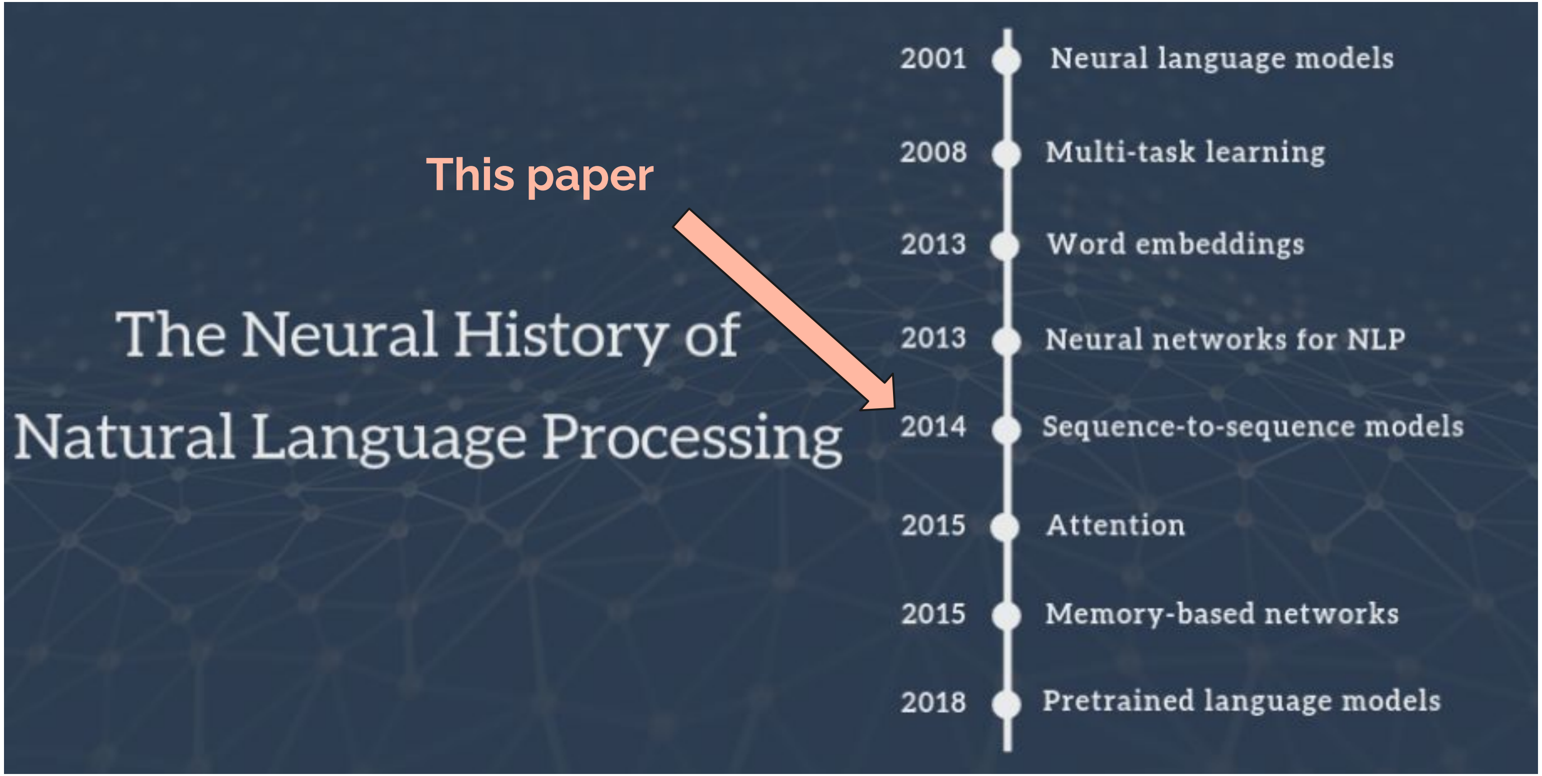
1. **Motivation for choice**
2. Preliminary on RNNs
3. GRU vs LSTM
4. Cho's experiments
5. Discussion

Why this paper?



Transformer-based pretraining

Credits: Sebastian Ruder



Transformer-based pretraining

Credits: Sebastian Ruder

2014- Sequence to sequence

- Sequence-to-sequence → Natural Language Generation (NLG) problems
- DNN architectures and their limitations:
  - Assumes the dimensionality of inputs and outputs are known and fixed

Motivation of Cho et al. (2014)

- Improving Phrase-Based SMT
- RNN allows to model variable-length sequences → as long as length of the sequences is known
    - Sequence to sequence with a neural net called **GRU**
- Concurrent work to Sutskever et al. (2014)

Outline

1. Motivation for choice
2. **Preliminary on RNNs**
3. GRU vs LSTM
4. Cho’s experiments
5. Discussion

Preliminary: RNNs

Generalization of feedforward NNs to sequences

We have a variable length sequence $\mathrm{x} = (x_1, \ldots, x_T)$. At each time step t, the hidden state $\mathrm{h}_{\langle t \rangle}$ of the RNN is updated by:

$$\mathrm{h}_{\langle t \rangle} = f\left(\mathrm{h}_{\langle t-1 \rangle}, x_t\right)$$

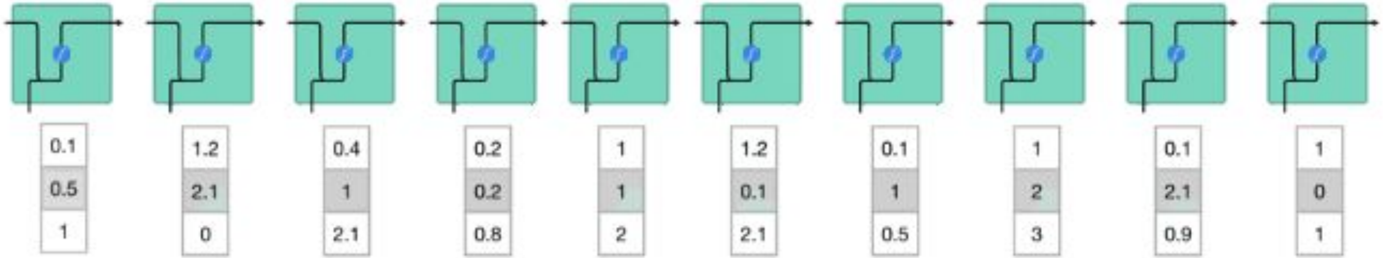
Where f is any non-linear activation function.

Preliminary: RNNs

RNNs can learn a probability distribution over a sequence, by learning to predict the next item in the sequence. Output at timestep t, is a conditional distribution:

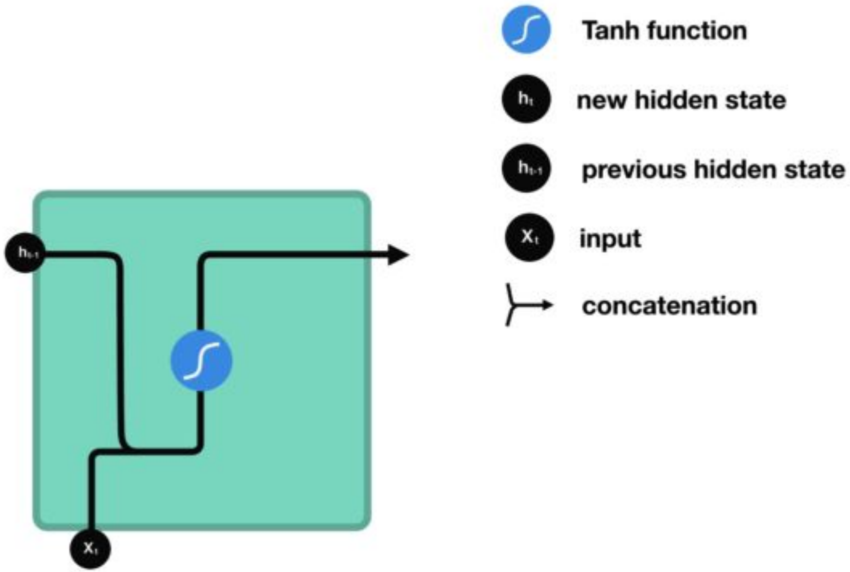
$$p(x_t | x_{t-1}, \ldots, x_1)$$

Preliminary RNNs



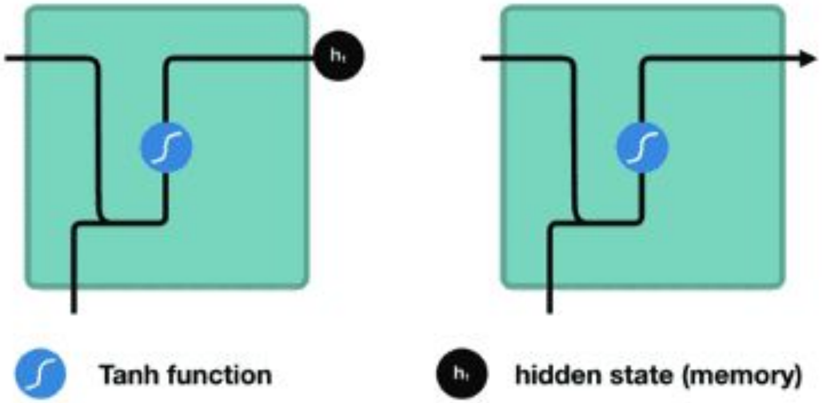
Animation credits: Michael Phi

Preliminary RNNs



Animation credits: Michael Phi

Preliminary RNNs



Animation credits: Michael Phi

Preliminary: RNNs

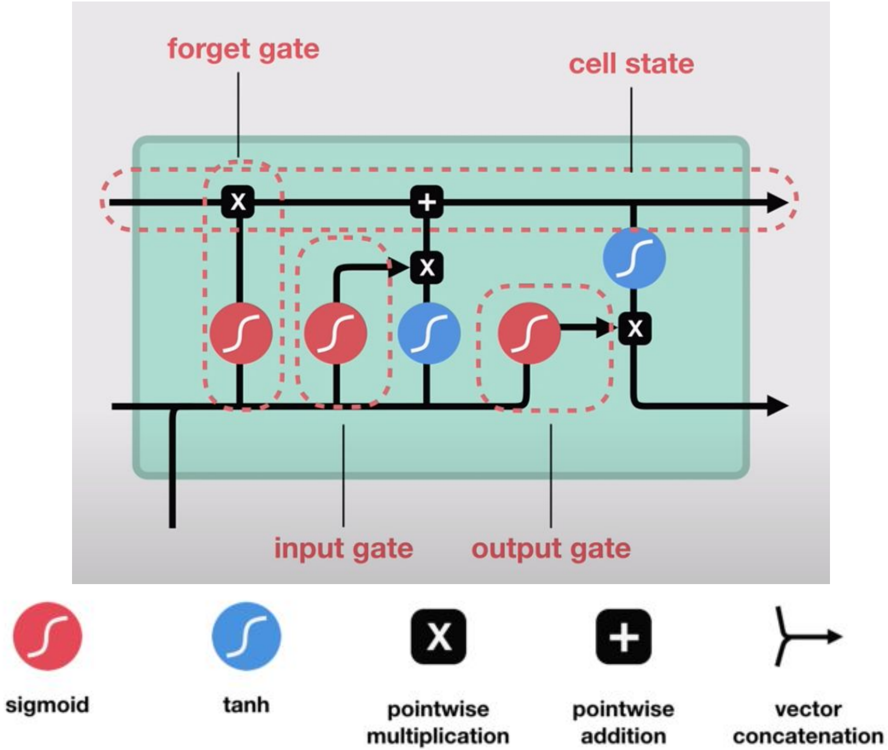
- Long-term dependencies
- Vanishing/exploding gradients
- Long Short Term Memory (LSTM) networks and Gated Recurrent Units (GRU) help deal with these problems

Outline

1. Motivation for choice
2. Preliminary on RNNs
3. **LSTM vs GRU**
4. Cho’s experiments
5. Discussion

LSTMs

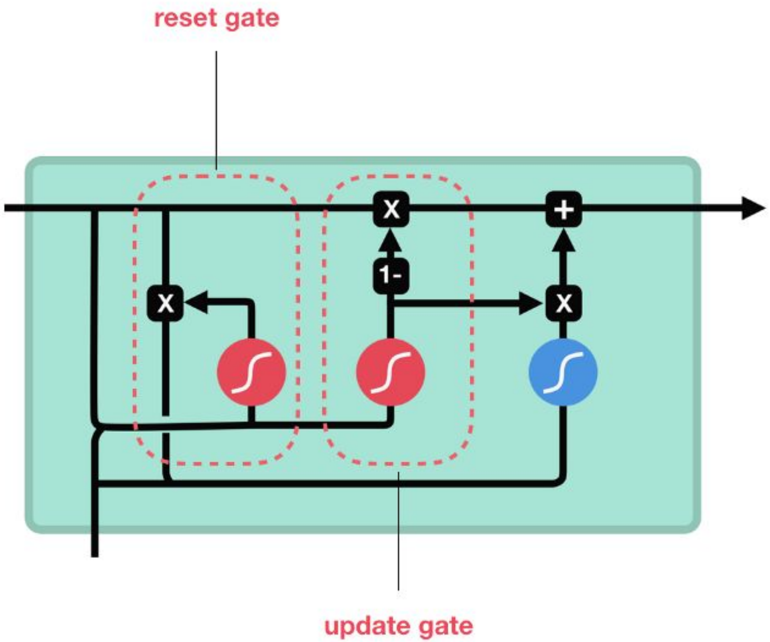
- Gates which can add or remove information
- Forget gate layer
- Input gate layer
- Output layer
- Cell



Animation credits: Michael Phi

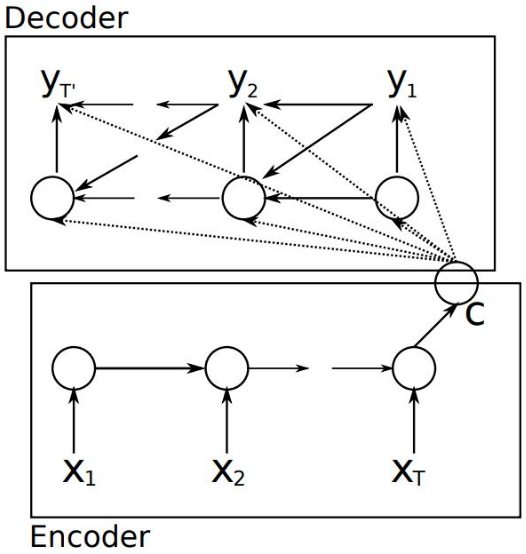
GRU

- Cho et al. (2014) introduced GRUs → inspired by LSTMs but computationally less expensive
- Reset gate
- Update gate



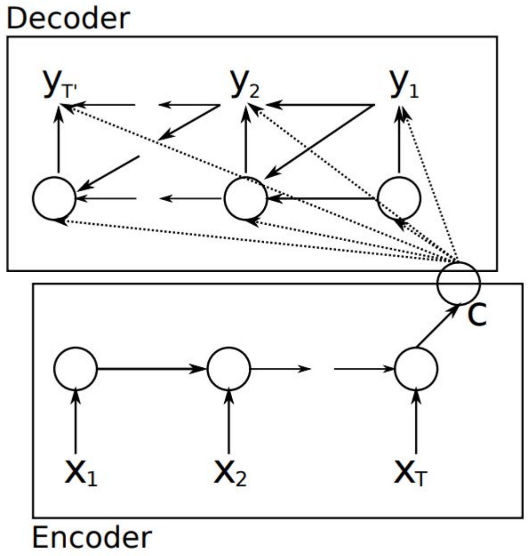
Animation credits: Michael Phi

RNN Encoder-Decoder



One RNN encodes a sequence into a fixed-length vector, another RNN decodes the vector into symbols.

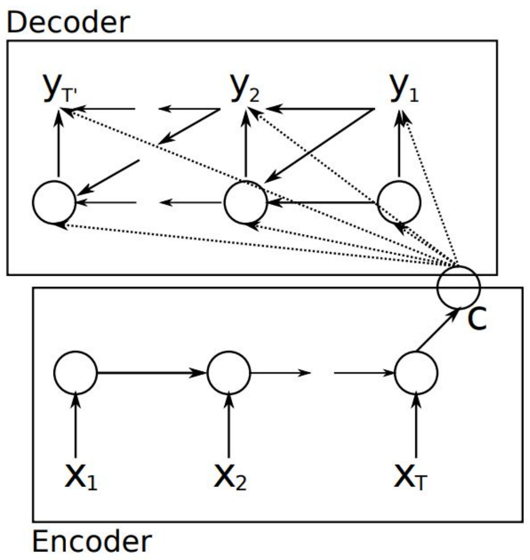
RNN Encoder-Decoder



One RNN encodes a sequence into a fixed-length vector, another RNN decodes the vector into symbols.

The internal states of the encoder are fed into the decoder. This is the initial state of the decoder.

RNN Encoder-Decoder



One RNN encodes a sequence into a fixed-length vector, another RNN decodes the vector into symbols.

The internal states of the encoder are fed into the decoder. This is the initial state of the decoder.

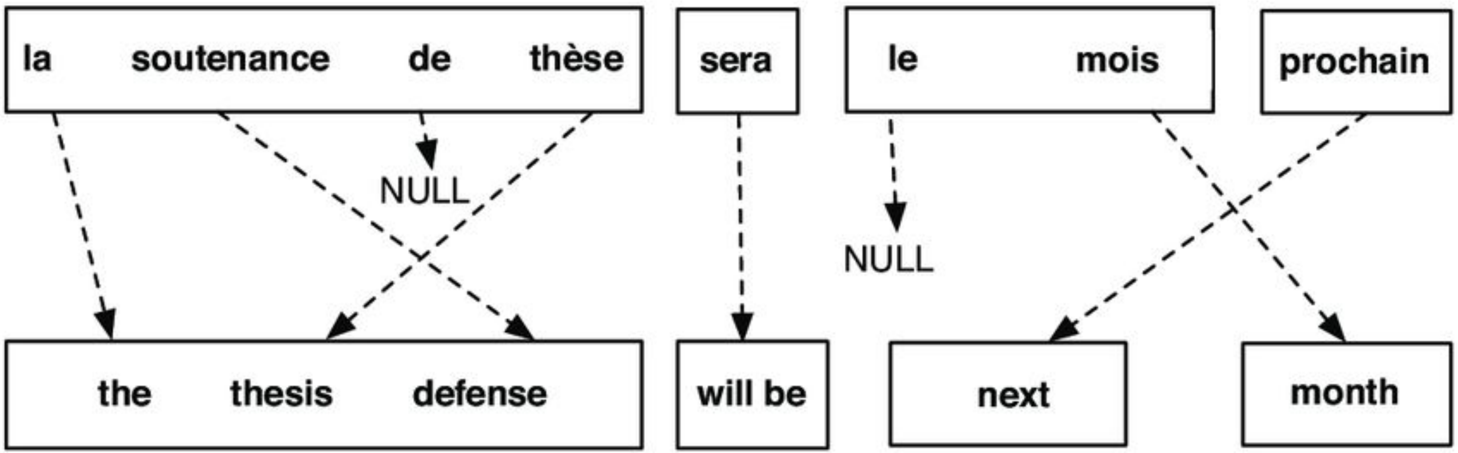
Sutskever et al. (2014) uses LSTMs…

Cho et al. (2014) introduces the GRU...

Outline

1. Motivation for choice
2. Preliminary on RNNs
3. GRU vs LSTM
4. **Cho's experiments**
5. Discussion

Experiments: Phrase-based Statistical Machine Translation



Experiments: Phrase-based Statistical Machine Translation

English	$\phi(\bar{e}\|\bar{f})$	English	$\phi(\bar{e}\|\bar{f})$
the proposal	0.6227	the suggestions	0.0114
's proposal	0.1068	the proposed	0.0114
a proposal	0.0341	the motion	0.0091
the idea	0.0250	the idea of	0.0091
this proposal	0.0227	the proposal ,	0.0068
proposal	0.0205	its proposal	0.0068
of the proposal	0.0159	it	0.0068
the proposals	0.0159	...	...

Rescoring Phrase Pairs

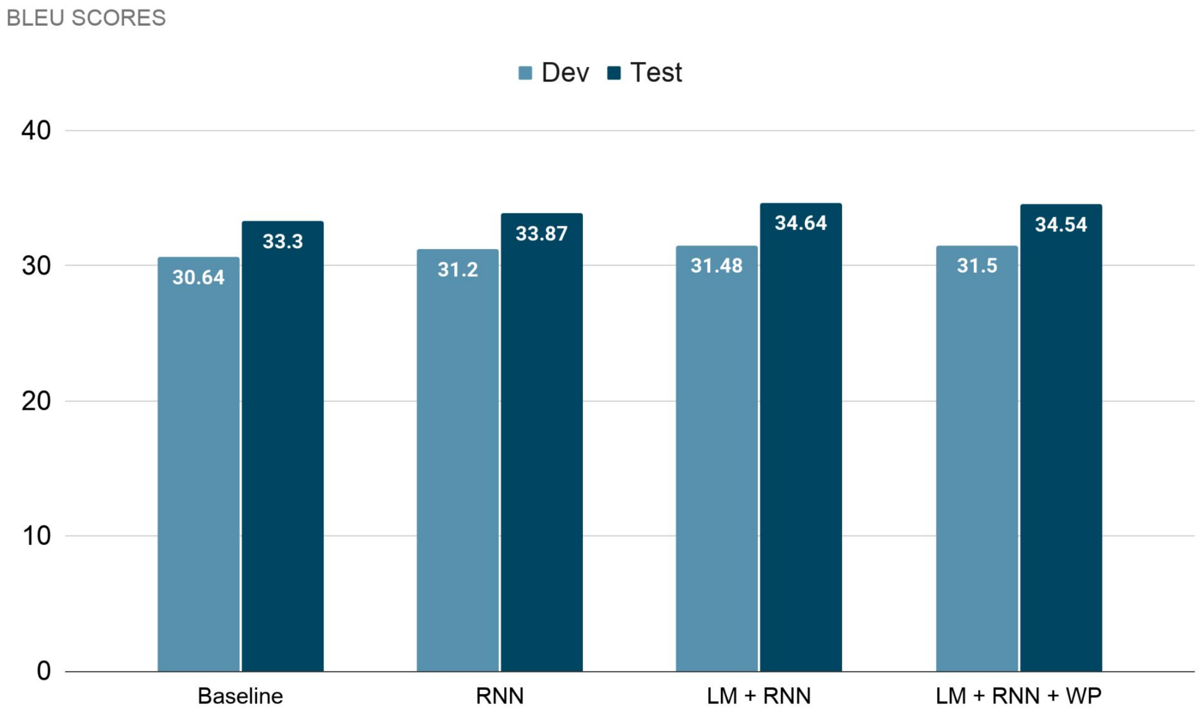
1. Train RNN encoder-decoder on a table of phrase pairs

2. Scores are used as input for SMT model
   a. Possible to replace the phrase table with the RNN encoder-decoder
   b. Cho et al. do not try this for computational reasons

3. Use the WMT'14 data

4. Compare RNN encoder-decoder phrase scoring, with SMT + LM
   a. Baseline configuration
   b. Baseline + RNN
   c. Baseline + CSLM + RNN
   d. Baseline + CSLM + RNN + Word penalty

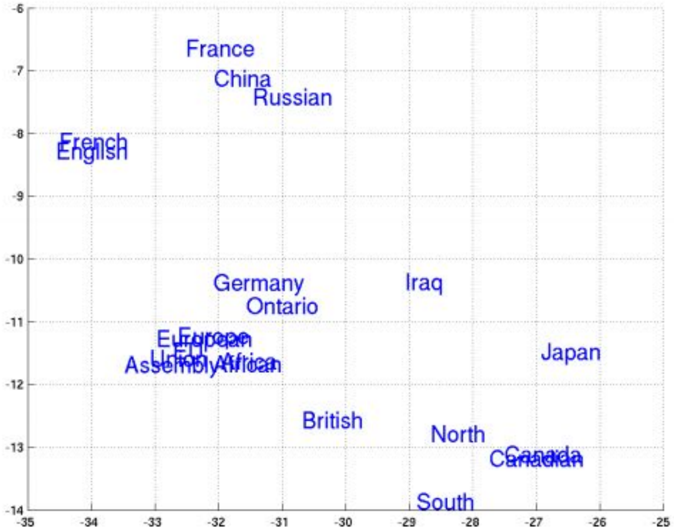
Results


BLEU SCORES
Dev
Test
40
30
20
10
0
30.64
33.3
31.2
33.87
31.48
34.64
31.5
34.54
Baseline
RNN
LM + RNN
LM + RNN + WP

Interesting properties of trained model

- Well known by now:
  - Naturally create a continuous-space representation of a phrase
  - Semantically similar words are clustered together
  - Encoder-decoder captures both semantic and syntactic properties

Extensions: [On the properties of NMT: Encoder-Decoder Approaches]()



Outline

1. Motivation for choice
2. Preliminary on RNNs
3. GRU vs LSTM
4. Cho’s experiments
5. **Discussion**

Relation to newer methods

- Attention mechanisms
- RNNs are no longer the standard architecture in NLP → Transformer-based models are now the *de-facto* architecture in NLP
- Transformer-based models can process input sequences of variable length without recurrence → highly parallelizable

References and group discussion

Sequence to Sequence Learning with Neural Networks. Sutskever et al. (2014)

On the properties of NMT: Encoder-Decoder Approaches

Attention is all you need

Papers

Neural Language modeling:
A Neural Probabilistic Language Model (2001)

Multitask learning:
A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning (2008)

Word Embeddings:
Natural language processing (almost) from scratch (2011)
Distributed Representations of Words and Phrases and their Compositionality (2013)
Glove: Global vectors for word representation (2014)

Attention:
Neural Machine Translation by Jointly Learning to Align and Translate (2016)