Attention, please

7 May, 2020

Alexey Zaytsev, head of Laboratory LARSS, PhD

Foundations of Data Science

Machine translation

Translate a sentence from one language to another

source language

target language

$\boldsymbol{x} \rightarrow \boldsymbol{y}$

A la guerre comme a la guerre

На войне как на войне

Early machine translation, 50s

Cold war child: Russian to English IBM 701 Translator

Doctor Dostert predicted that *"five, perhaps three years hence, interlingual meaning conversion by electronic process in important functional areas of several languages may well be an accomplished fact. "* (1954)

Mostly rule-based approach, uses a dictionary to map Russian words to English

https://www.ibm.com/ibm/history/exhibits/701/701_translator.html
https://youtu.be/8ZtdVUB007A

Statistical machine translation, 90s-2010s

Learn probabilistic model from data

$p(\boldsymbol{y}|\boldsymbol{x})$

To translate: for input English sequence $\boldsymbol{x}$ find the most probable Russian sentence $\boldsymbol{y}$

$$p(\boldsymbol{y}|\boldsymbol{x}) \to \max_{\boldsymbol{y}}$$

Bayesian perspective:

$$p(\boldsymbol{y}|\boldsymbol{x}) \sim p(\boldsymbol{x}|\boldsymbol{y})p(\boldsymbol{y})$$

Translation model: learn from parallel corpus

Language model: learn from monolingual corpus

Learning translation model

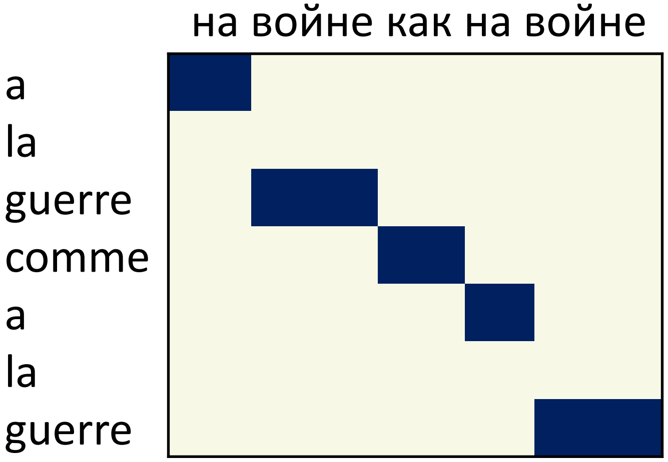
Learn translation model from data

$p(\boldsymbol{x}|\boldsymbol{y})$

- Large amount of parallel data
- Alignment

$$p(\boldsymbol{x}, \boldsymbol{a}|\boldsymbol{y})$$

correspondence between words in different languages



Alignment types:
- one-to-one
- spurious words
- one-to-many
- Many-to-one

Many problems on the way

Decoding

The optimization problem is hard

$$p(\boldsymbol{y}|\boldsymbol{x}) \to \max_{\boldsymbol{y}}$$

- Full search is not possible
- Heuristic search algorithm to search for the best translation: look through a tree of possible options

The best SMT systems are very complex

- Language itself is very complex
- Many details we don’t even mention
- Separately designed subcomponents
- Tricky feature engineering
- Extra information
- The language changes – we need to maintain the system

All difficulties we saw about pre-Neural approaches to sequence processing multiplied x100!

Neural machine translation

Our goal is to do machine translation with one Neural network

It works with two RNNs

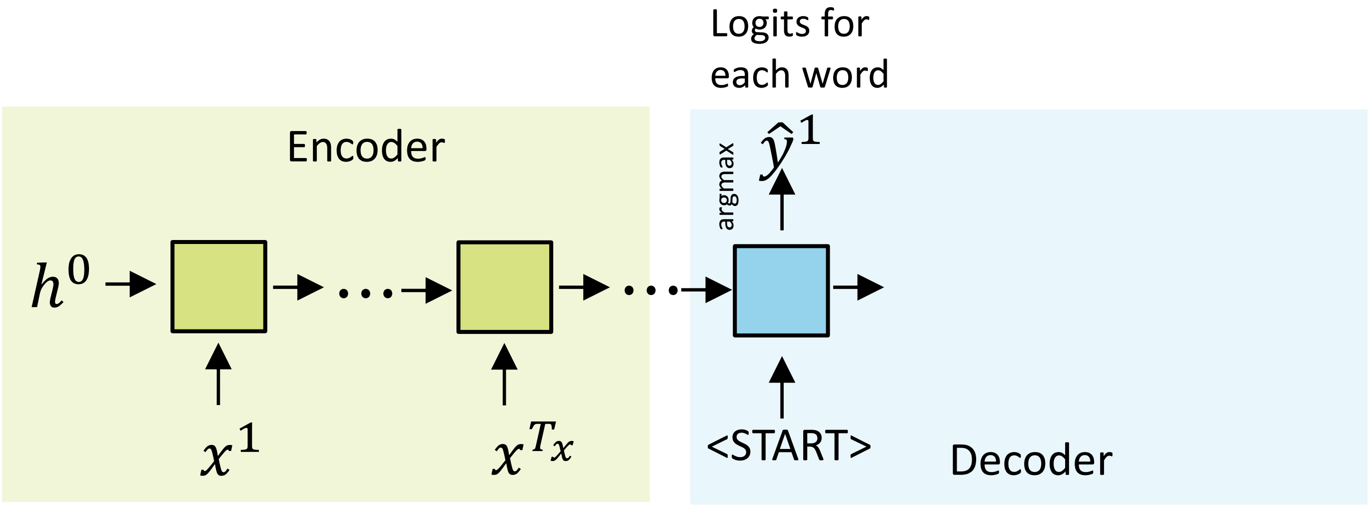
Sequence to sequence **seq2seq** architecture

Neural machine translation

Our goal is to do machine translation with one Neural network

It works with two RNNs

Sequence to sequence **seq2seq** architecture

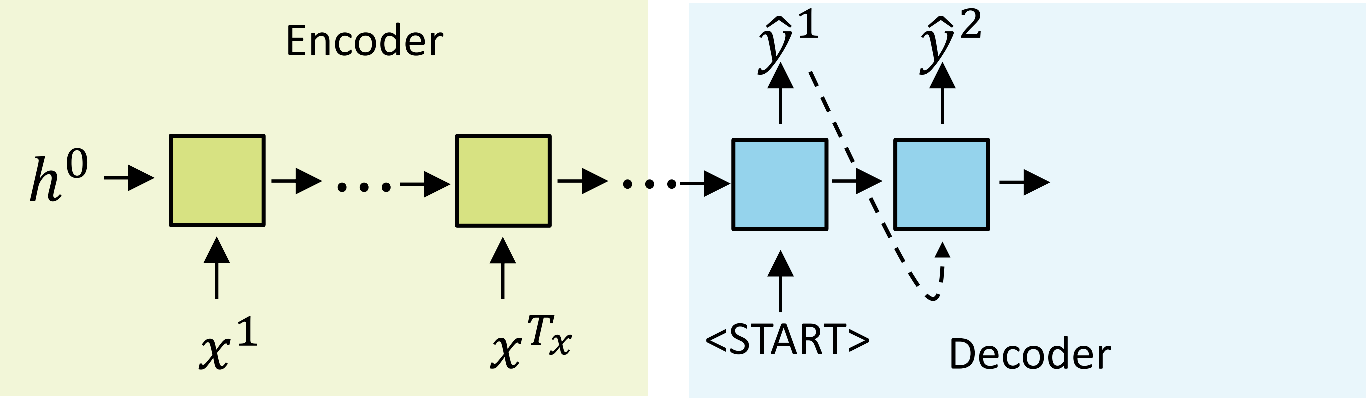


Neural machine translation

Our goal is to do machine translation with one Neural network

It works with two RNNs

Sequence to sequence **seq2seq** architecture

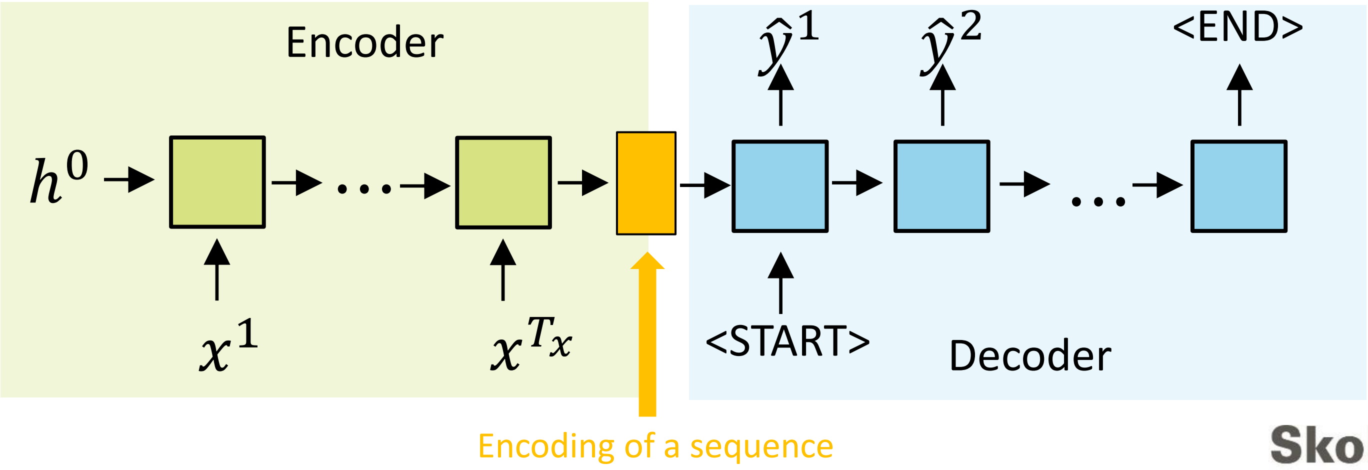


Neural machine translation

Our goal is to do machine translation with one Neural network

It works with two RNNs

Sequence to sequence **seq2seq** architecture

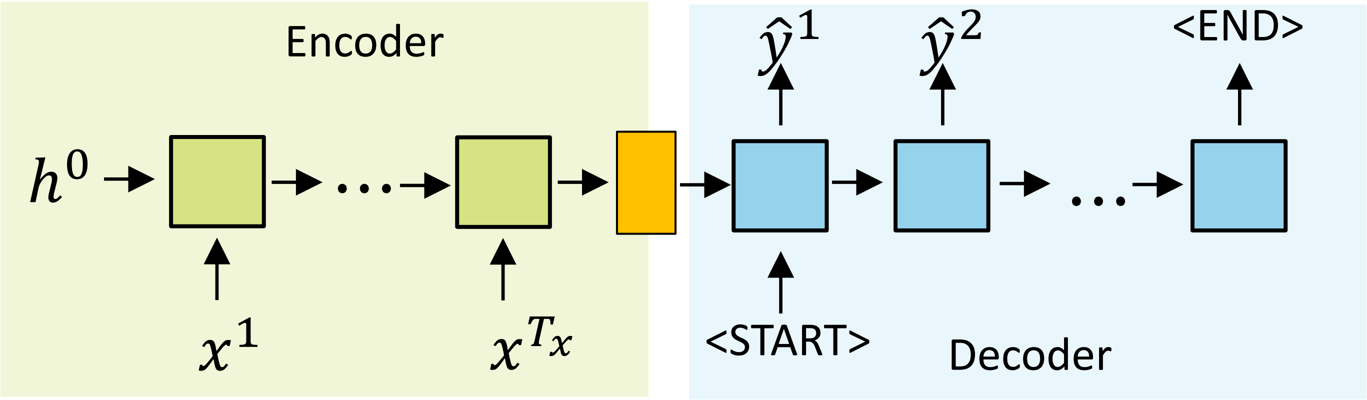


Neural machine translation

Our goal is to do machine translation with one Neural network

It works with two RNNs

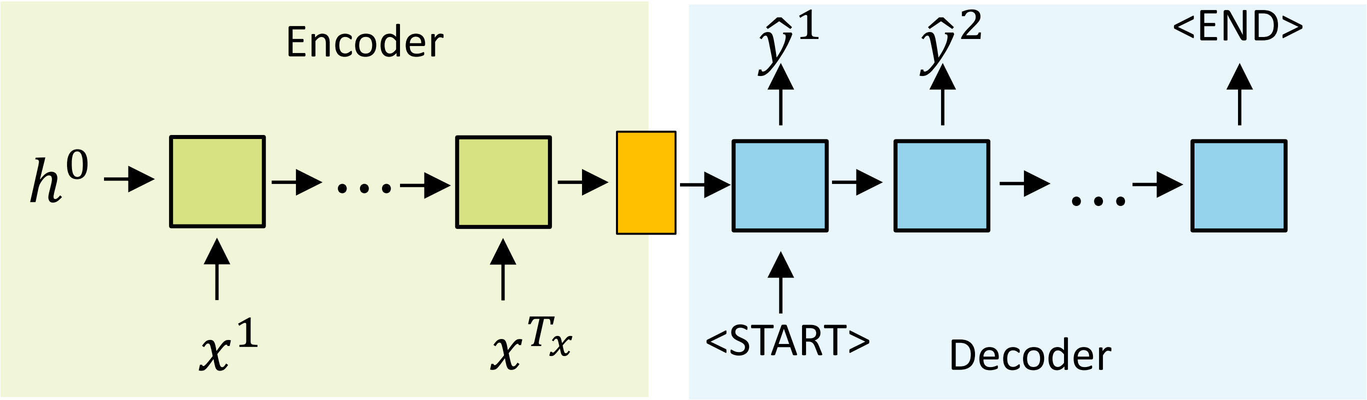
Sequence to sequence **seq2seq** architecture



Language model

New world of seq2seq models

Summarization: long text – text summary
Dialogue: one phrase – another phrase
Parsing: input – output parse as a sequence
Code generation: task description – python code



Language model

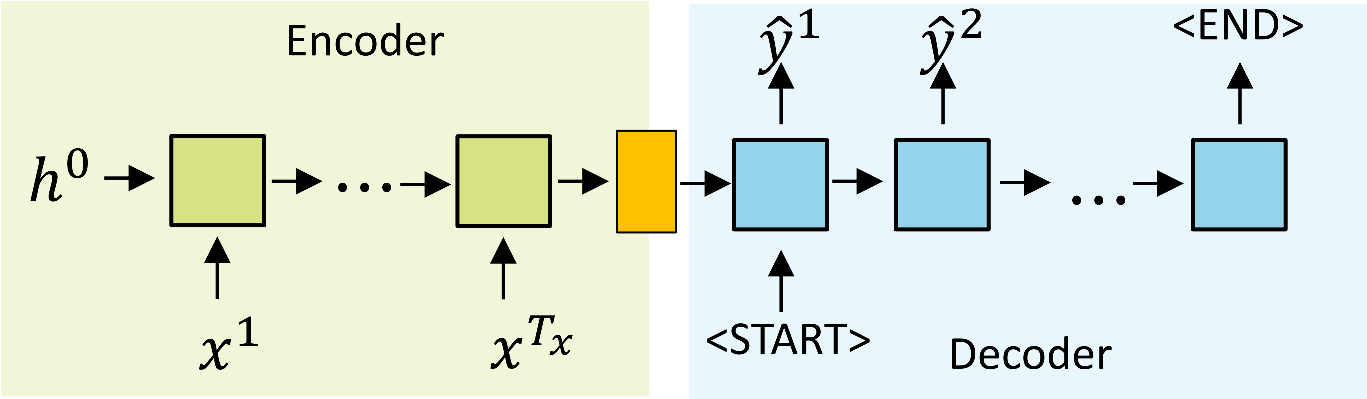
NMT directly models the conditional language model

Learn probabilistic model from data

$$p(\boldsymbol{y}|\boldsymbol{x}) = p(y_T|y_1, y_2, \dots, y_{T-1}, \boldsymbol{x})\, p(y_{T-1}|y_1, y_2, \dots, y_{T-2}, \boldsymbol{x}) \dots p(y_1|\boldsymbol{x})$$

Each term is an RNN block

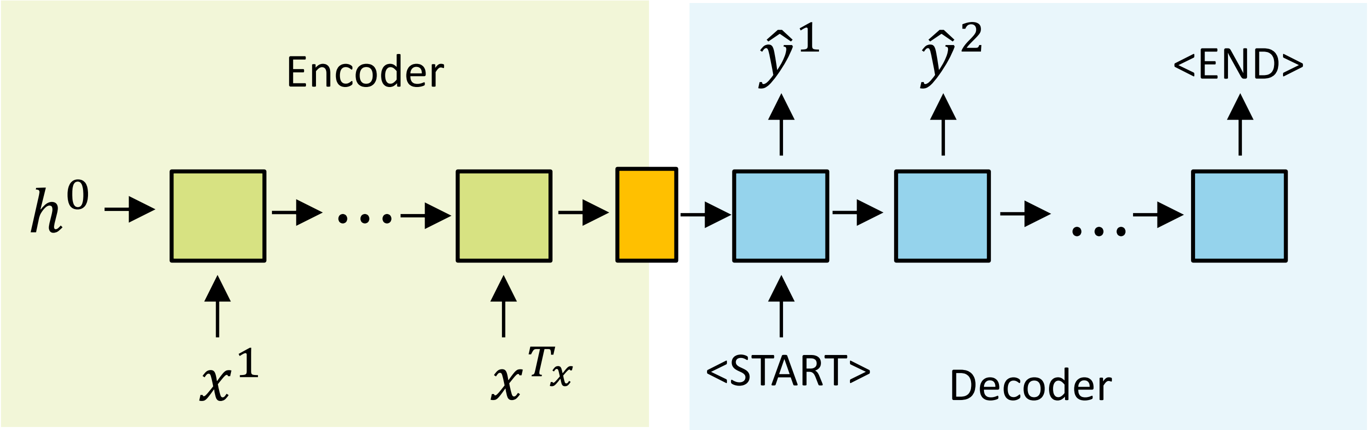
Loss function: compare logits to true words, now we have a logloss if we have a parallel corpus



Beam search decoding

Problem:
We generate tokens one by one from the very beginning
Wrong decision at some step leads to wrong translation as a whole

Solution:
“Beam search”: Keep a population of solutions and
select the most probable sequences at each step



Skoltech
Skolkovo Institute of Science and Technology

Beam search decoding: practical implementation

Stopping criteria:
- Generated <END> token
- Reached maximum sequence length

We have $\{\boldsymbol{y}_1, \boldsymbol{y}_2, \dots, \boldsymbol{y}_m\}$ after beam search

Selection criterion:

Compare normalized $\tilde{p}(\boldsymbol{y}_i|\boldsymbol{x})$ instead of $p(\boldsymbol{y}_i|\boldsymbol{x})$

$$\tilde{p}(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{|\boldsymbol{y}|} p(\boldsymbol{y}|\boldsymbol{x})$$

Advantages of NMT

- Better performance compared to SLT

- Single end2end neural network

- Much less human engineering effort
  - No feature engineering
  - Same method for all language pairs

Evaluations of Machine Translation: BLEU

"the closer a machine translation is to a professional human translation, the better it is“

Precision: share of words from $\hat{\boldsymbol{y}}$ that appear in $\boldsymbol{y}$

True $\boldsymbol{y}$	the	cat	is	on	the	mat
Candidate $\hat{\boldsymbol{y}}$	the	the	the	the	the	the

Unigram precision is 6/6 = 1

Limit with number of words from references: there are 2 “the” in $\boldsymbol{y}$

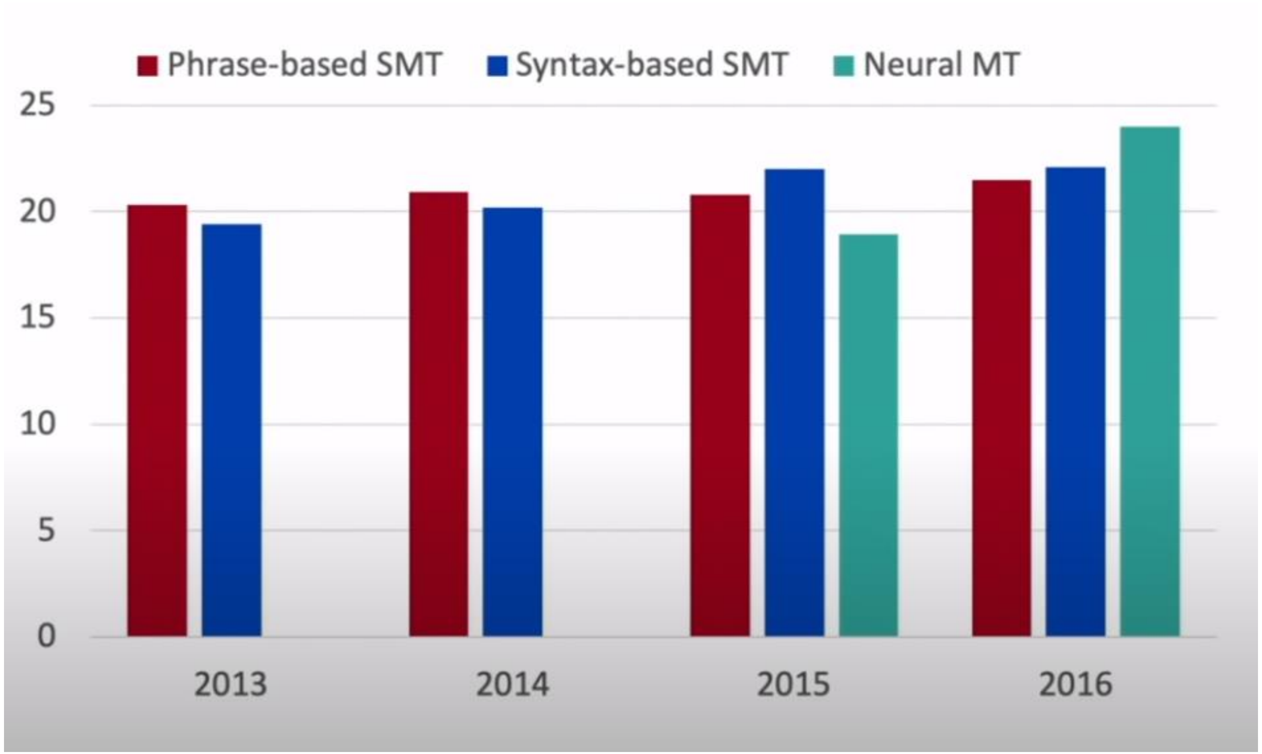
Unigram precision is 2/6 = 0.33

In practice we use n-gram modified precision with n = 4 that “best correlates to human judgement”.

Recall is also important

Papineni, K.; Roukos, S.; Ward, T.; Zhu, W. J. 2002. *BLEU: a method for automatic evaluation of machine translation*. ACL-2002: 40th Annual meeting of the Association for Computational Linguistics

Evaluations of Machine Translation



source: Stanford CS224N: NLP with Deep Learning | Winter 2019 | Lecture 8 – Translation, Seq2Seq, Attention

NMT: success story for deep learning

2014: first seq2seq paper published
2016: Google translate switches from SMT to NMT

SMT: hundreds of engineers for many years
NMT: handful of engineers in a few months

Disadvantages of NMT

- NMT is less interpretable
  - no alignment

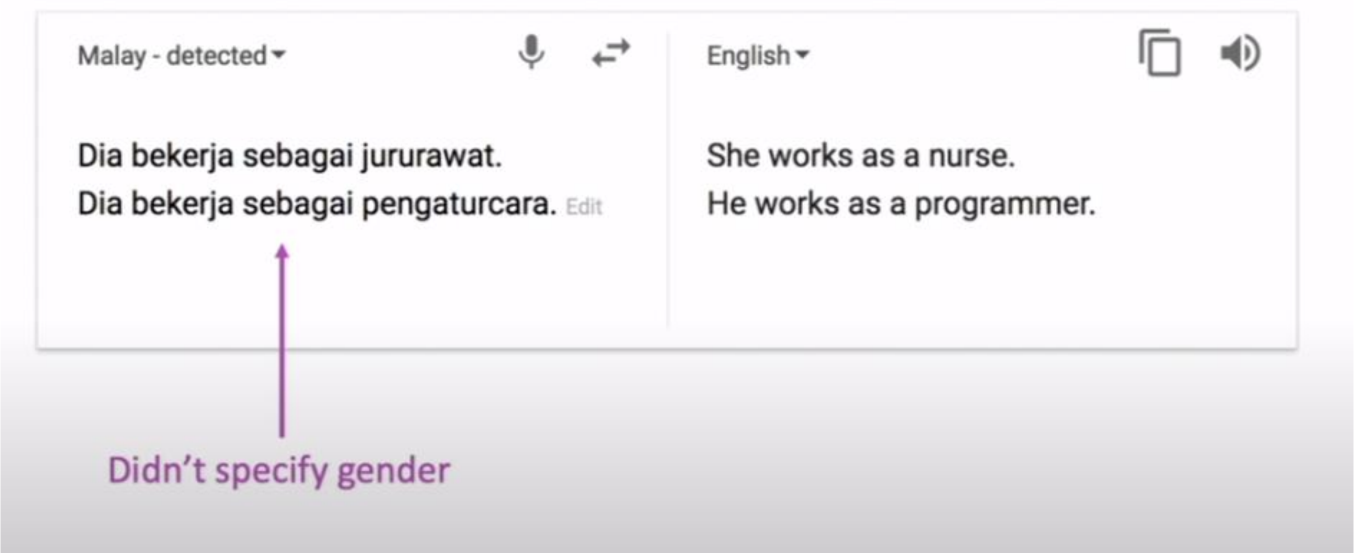
- Difficult to control
  - hard to enforce specific translation rules
  - No guidelines or rules for translation
  - Safety concerns

Any other problems with NMT?

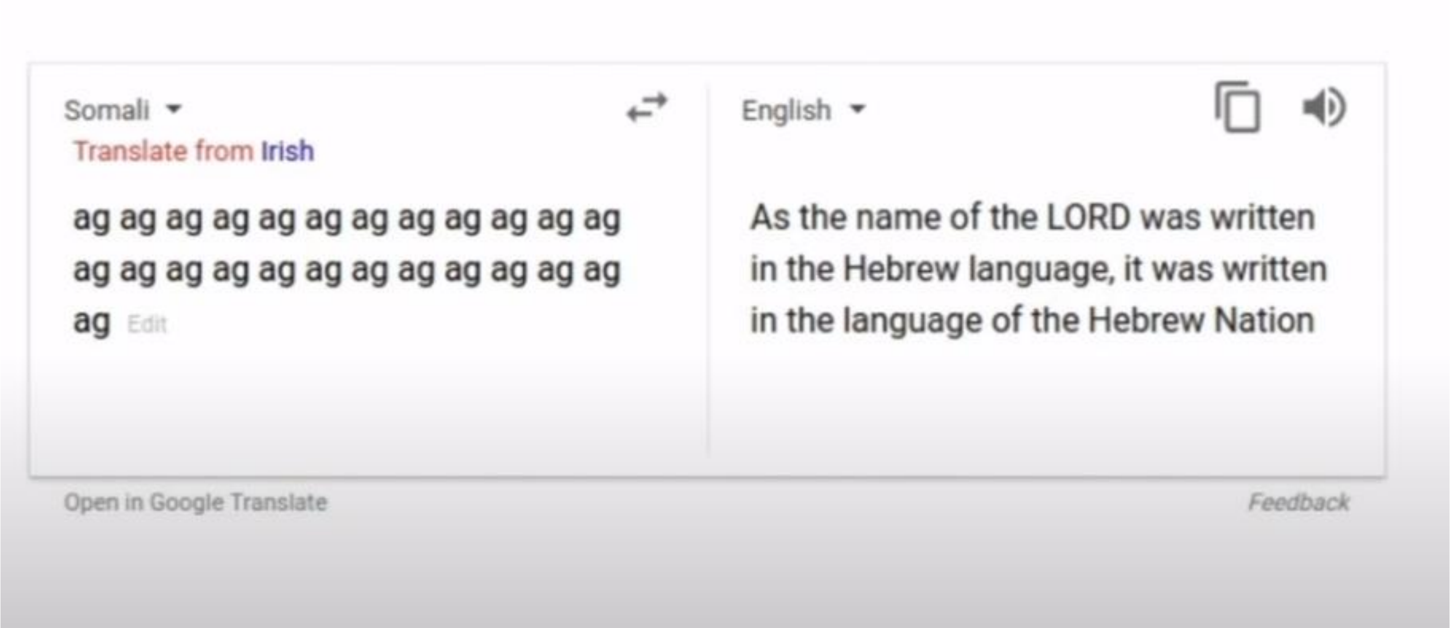
- Out-of-vocabulary words
- Domain mismatch: training and test data
- Context over longer texts
- Low-resource language pairs

Common sense

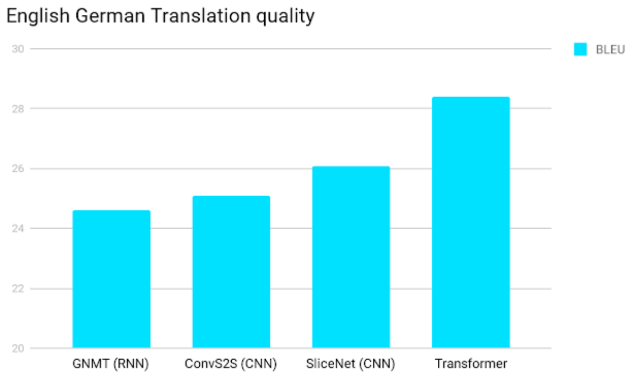
NMT picks up biases in training data



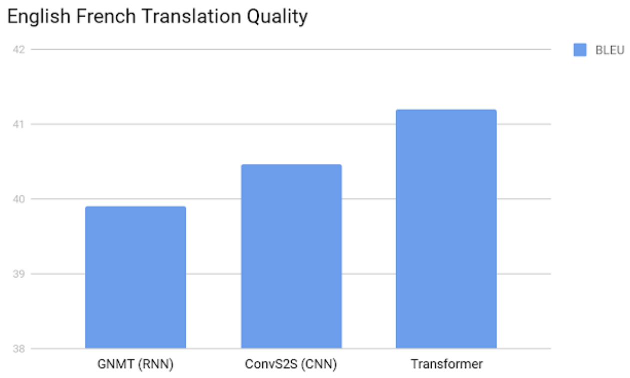
Strange things



New state of the art: **attention** is all we need



BLEU scores (higher is better) of single models on the standard WMT newstest2014 English to German translation benchmark.

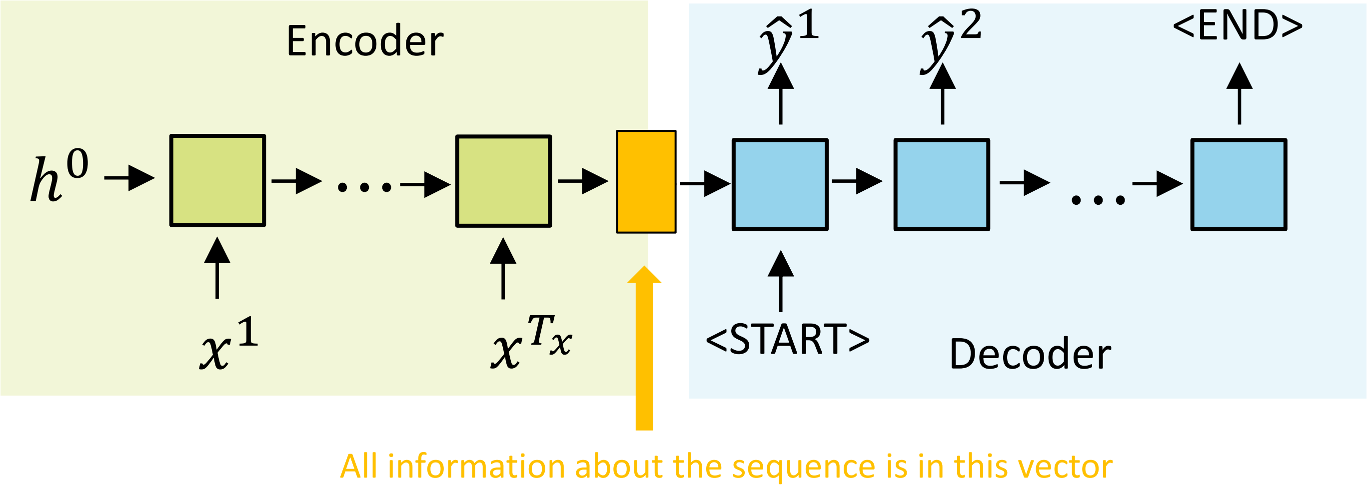


BLEU scores (higher is better) of single models on the standard WMT newstest2014 English to French translation benchmark.

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

Skoltech
Skolkovo Institute of Science and Technology

Bottleneck in seq2seq models



Attention

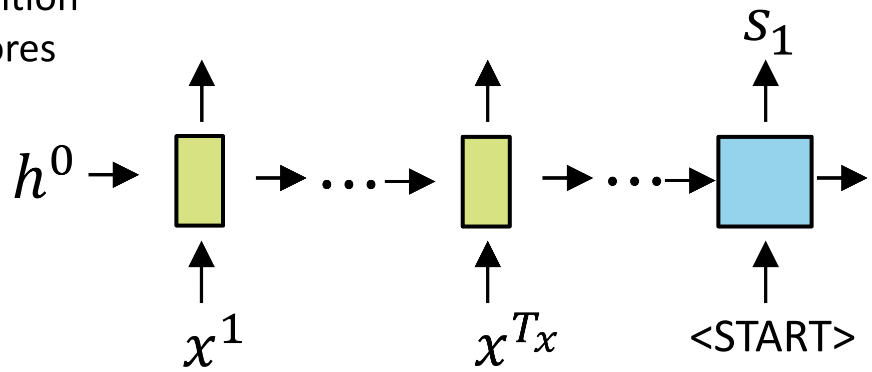
- Solution to the bottleneck problem
- Direction connection between parts of input and output sequence

Sequence 2 sequence with attention (no formulas now)

Attention output

Attention distribution

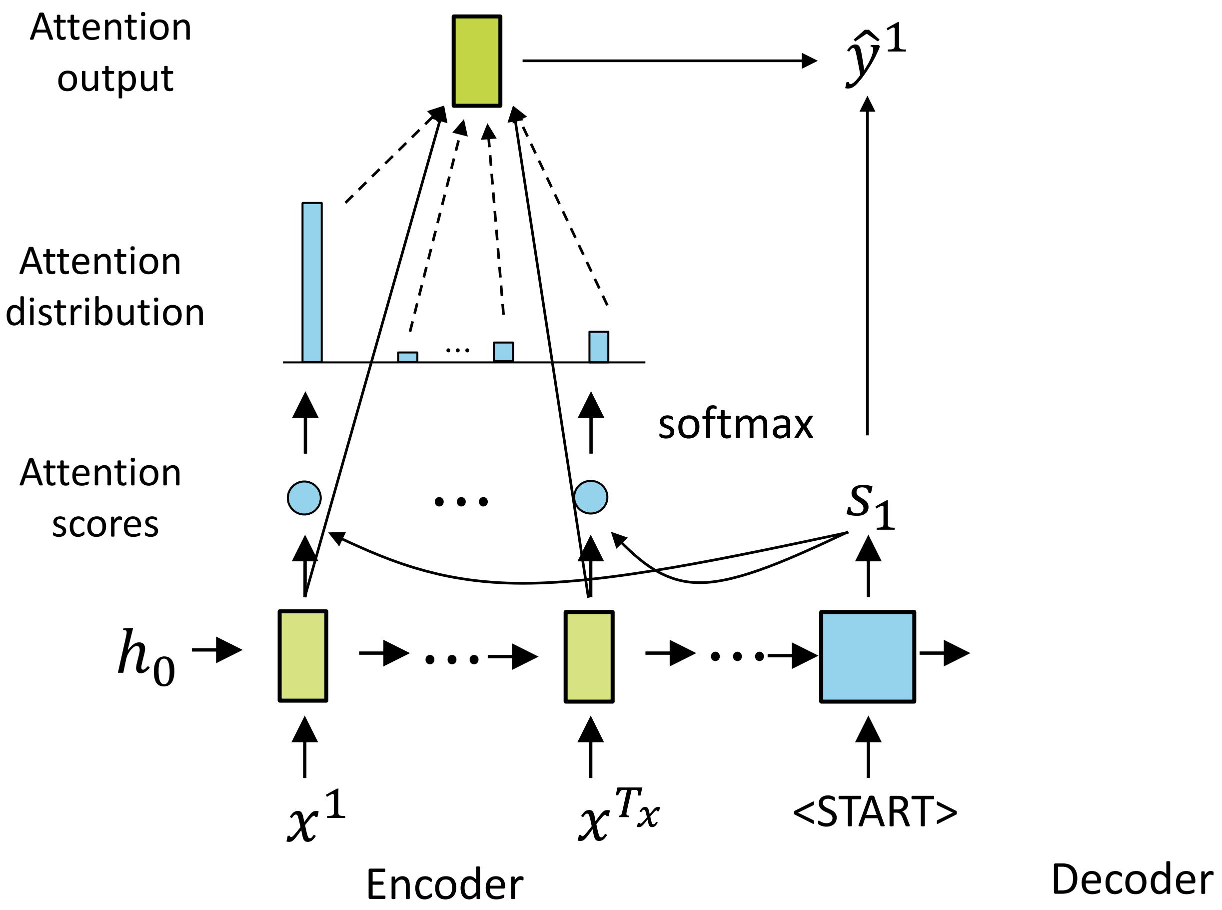
Attention scores



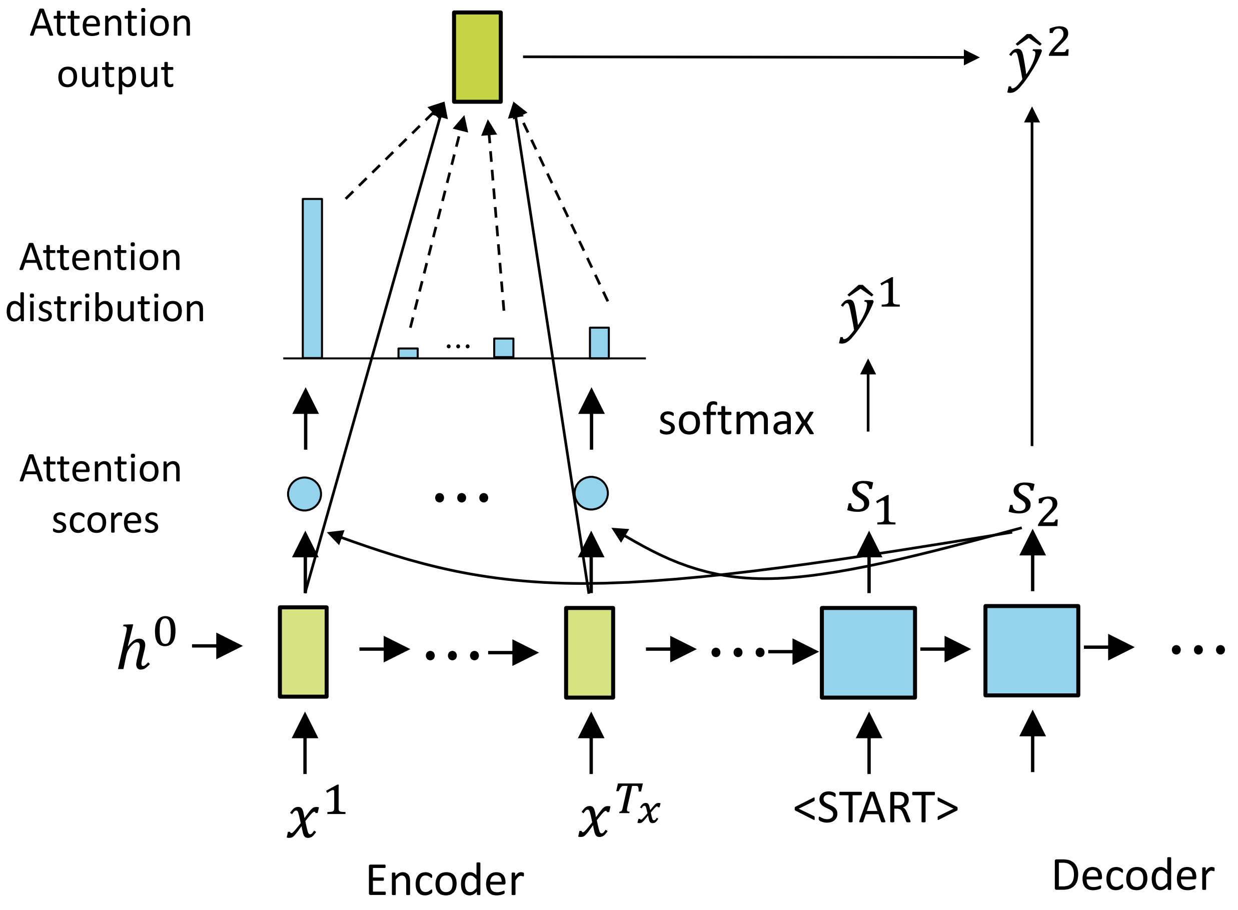
Encoder

Decoder

Sequence 2 sequence with attention



Sequence 2 sequence with attention



Attention: formulas

- First RNN produces encoder hidden states $\boldsymbol{h}_1, \dots, \boldsymbol{h}_{T_x} \in \mathbb{R}^h$
- Decoder hidden state $\boldsymbol{s}_t \in \mathbb{R}^h$ at time step t
- Attention scores for step t:
  $\boldsymbol{e^t} = [\boldsymbol{s}_t^{\boldsymbol{T}}\boldsymbol{h}_1, \dots, \boldsymbol{s}_t^{\boldsymbol{T}}\boldsymbol{h}_{T_x}] \in \mathbb{R}^{T_x}$
- Softmax to get attention distribution: all values are positive, sum of all values is 1:
  $\boldsymbol{\alpha^t} = \mathrm{softmax}(\boldsymbol{e^t}) \in \mathbb{R}^{T_x}$
- Attention output $\boldsymbol{a}_t$ is a weighted sum of hidden states:

$$\boldsymbol{a}_t = \sum_{i=1}^{T_x} \alpha_i^t \boldsymbol{h}_i \in \mathbb{R}^h$$

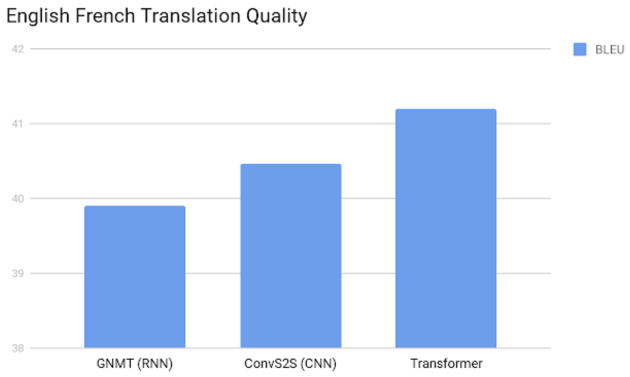
- We concatenate the attention output $\boldsymbol{a}_t$ with the decoder hidden state $\boldsymbol{s}_t$ and proceed to the non-attention part of our seq2seq model

$$[\boldsymbol{a}_t, \boldsymbol{s}_t] \in \mathbb{R}^{2h}$$

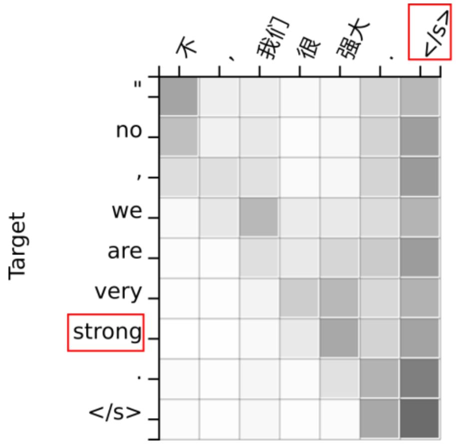
Attention is just great

- Significantly improves performance of NMT
- Solves the bottleneck problem
  - All encoder tokens are connected to all decoder tokens
- No more vanishing gradients
  - All to All connection
- Provides some interpretability
  - see alignment figure

- Similar to RNN seq2seq, but greater!



BLEU scores (higher is better) of single models on the standard WMT newstest2014 English to French translation benchmark.



Attention is a general deep learning idea

We can use attention in many architectures and many tasks

- Other NLP problems
- Graph Neural networks

Key value interpretation:

s_i - query to a database	Hidden state of the decoder
k_i - keys in the database	Hidden state of the encoder
h_i - values in the database	Hidden state of the encoder

We calculate correspondence $e(s_i, k_i)$

Then we extract information as weighted sum of values $\sum_{i=1}^{T_x} \alpha_i^t \boldsymbol{h}_i$

Sources

- Stanford CS224N: NLP with Deep Learning | Winter 2019 | Lecture 8 – Translation, Seq2Seq, Attention

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

Skoltech
Skolkovo Institute of Science and Technology