A World Leading SFI Research Centre

Science Foundation Ireland For what's next

Engaging Content
Engaging People

Encoder-decoder, Machine Translation and more

Dimitar Shterionov

Post-doctoral researcher, DCU

"One naturally wonders if the problem of translation could conceivably be treated as a problem in cryptography. When I look at an article in Russian, I say: 'This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.' "
- Warren Weaver, 1947

Encoding $\leftrightarrow$ Decoding

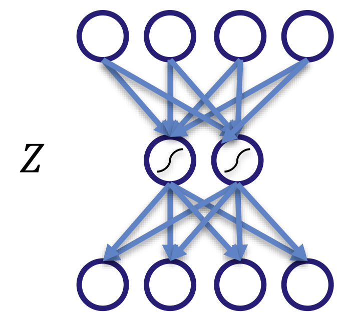
Autoencoders

- Suppose we have a set of **multi-dimensional** data points $X = \{x^1, x^2, \dots, x^m\}$.
- Is there a general way to map $X \to Z = \{\mathrm{z}^1, \mathrm{z}^2, \dots, \mathrm{z}^m\}$, where z's have **lower dimensionality** than x's and
- Z can faithfully **reconstruct** X: $Z \to \tilde{X}$

$$z^i = W_1 x^i + b_1$$
$$\tilde{x}^i = W_2 z^i + b_2$$

$$J(W_1, b_1, W_2, b_2) = \sum_{i=1}^{m} \left(\tilde{x}^i - x^i\right)^2$$

$X = \{x^1, x^2, \dots, x^m\}$



$\tilde{X} = \{\tilde{x}^1, \tilde{x}^2, \dots, \tilde{x}^m\}$

- Use stochastic gradient descent to minimize
- Autoencoders are unsupervised

[Quoc V. Le, A Tutorial on Deep Learning Part 2: Autoencoders, Convolutional Neural Networks and Recurrent Neural Networks]

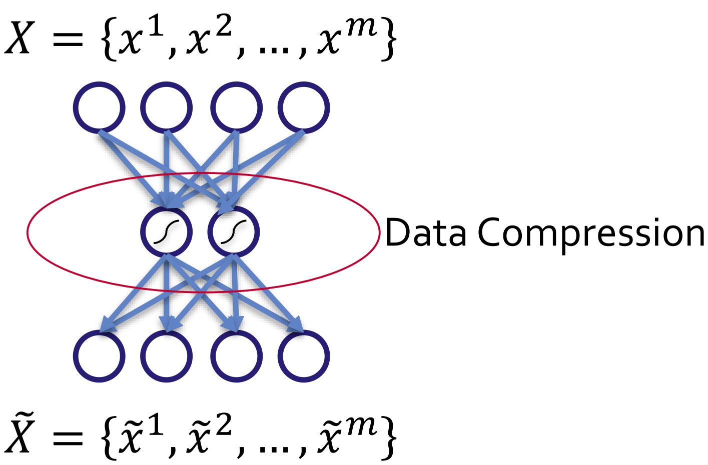
Encoding ⟷ Decoding

Autoencoders

- Suppose we have a set of **multi-dimensional** data points $X = \{x^1, x^2, \dots, x^m\}$.
- Is there a general way to map $X \to Z = \{\mathrm{z}^1, \mathrm{z}^2, \dots, \mathrm{z}^\mathrm{m}\}$, where z's have **lower dimensionality** than x's and
- Z can faithfully **reconstruct** X: $Z \to \tilde{X}$

$$z^i = W_1 x^i + b_1$$
$$\tilde{x}^i = W_2 z^i + b_2$$

$$J(W_1, b_1, W_2, b_2) = \sum_{i=1}^{m} \left(\tilde{x}^i - x^i\right)^2$$



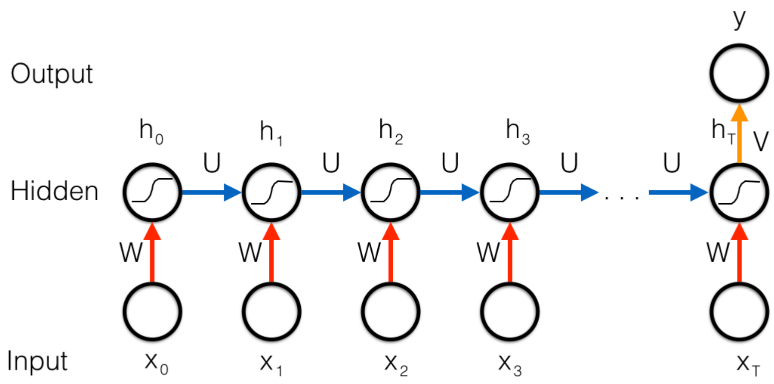
- Use stochastic gradient descent to minimize
- Autoencoders are unsupervised

[Quoc V. Le, A Tutorial on Deep Learning Part 2: Autoencoders, Convolutional Neural Networks and Recurrent Neural Networks]

Encoding ↔ Decoding

Sequences

- N→1
  Language modelling: $X = \{x^1, x^2, \ldots, x^{T-1}\}, y = x^T$,
  x^i is the words i, T is current word.



- N→M
  Translation: $X = \{x^1, x^2, \ldots, x^T\}, Y = \{y^1, y^2, \ldots, y^{T'}\}$,
  X is a sentence in the source language and Y is the sentence in the target language

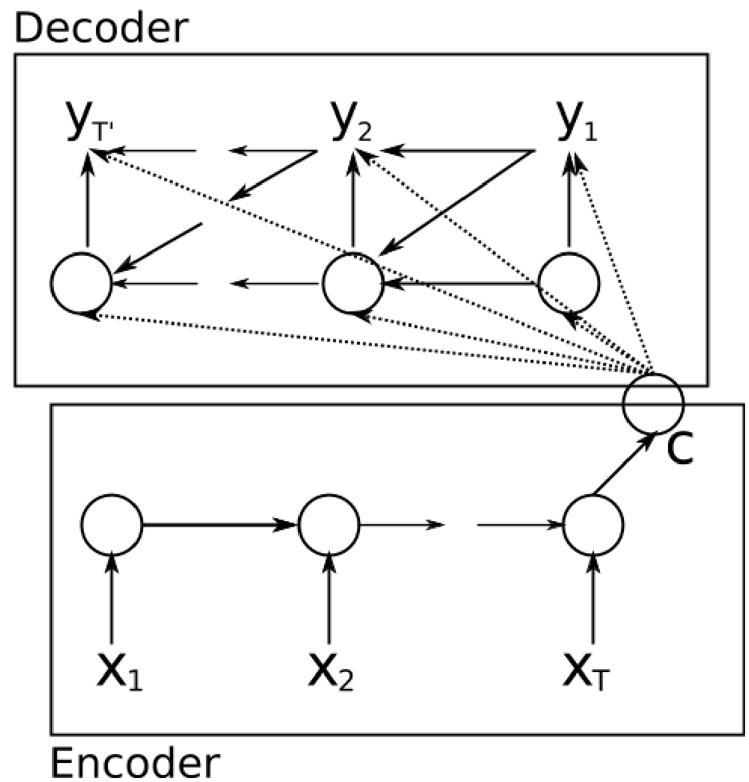
[Quoc V. Le, A Tutorial on Deep Learning Part 2: Autoencoders, Convolutional Neural Networks and Recurrent Neural Networks]

Encoding ↔ Decoding

Sequences

- N→1
  Language modelling: $X = \{x^1, x^2, \dots, x^{T-1}\}, y = x^T$,
  x^i is the words i, T is current word.

- N→M
  Translation: $X = \{x^1, x^2, \dots, x^T\}, Y = \{y^1, y^2, \dots, y^{T'}\}$,
  X is a sentence in the source language and Y is the sentence in the target language



[Cho et al, 2014 Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation]

Encoding ↔ Decoding

Sequences

- N→1

  Language modelling: $X = \{x^1, x^2, \ldots, x^{T-1}\}, y$

  x^i is the words i, T is current word.

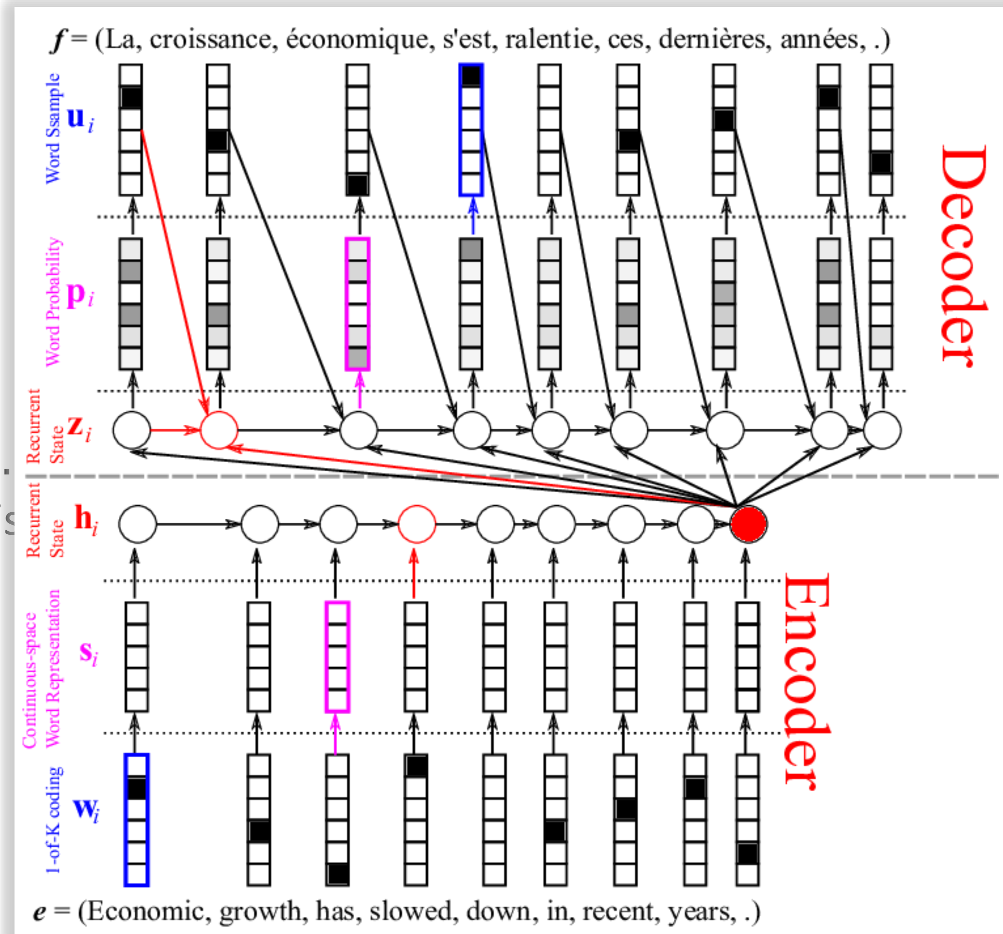
- N→M

  Translation: $X = \{x^1, x^2, \ldots, x^T\}, Y = \{y^1, y^2, \ldots$

  X is a sentence in the source language and Y i

  sentence in the target language

$$p(y^i | y^1, y^2, \ldots, y^{i-1}, h^T)$$



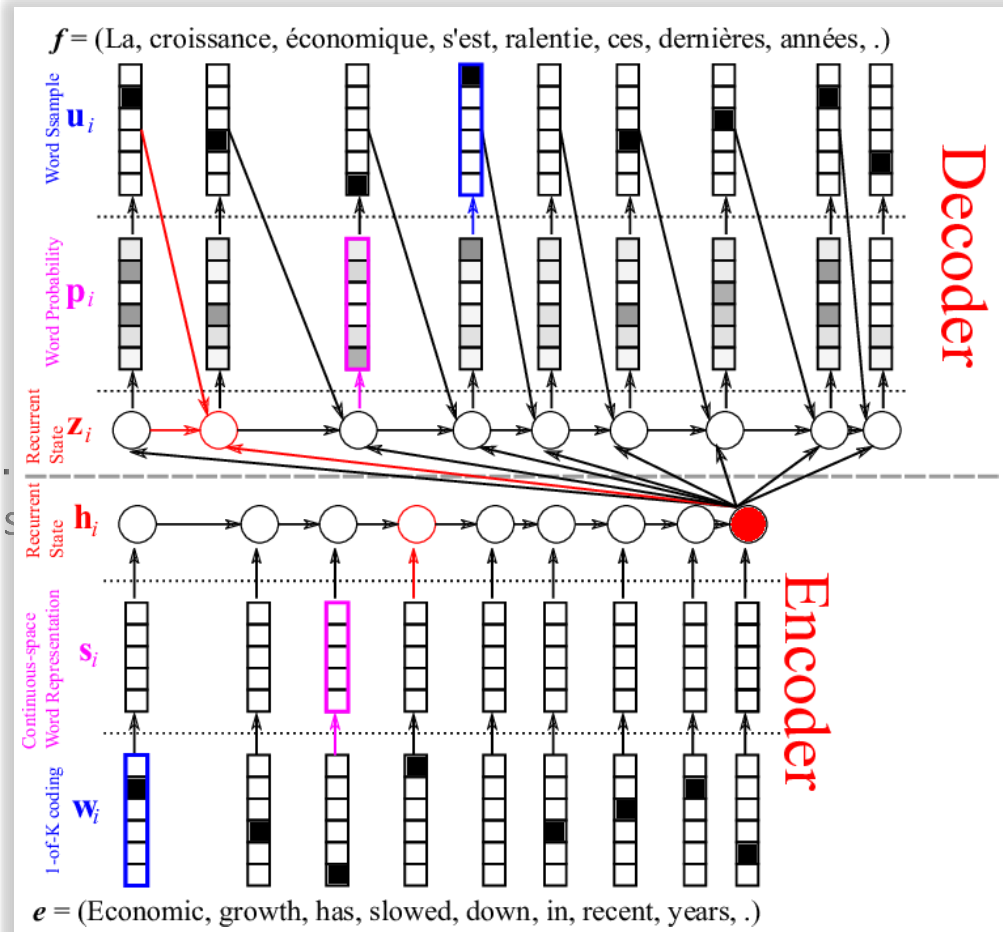
[Cho et al, 2014 Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation]

Encoding ↔ Decoding

Sequences

- N→1
  Language modelling: $X = \{x^1, x^2, \ldots, x^{T-1}\}, y$
  x^i is the words i, T is current word.

- N→M
  Translation: $X = \{x^1, x^2, \ldots, x^T\}, Y = \{y^1, y^2, ..$
  X is a sentence in the source language and Y i
  sentence in the target language



$$p(y^i | y^1, y^2, \ldots, y^{i-1}, h^T)$$

$$p(Y^n | X^n, \theta)$$

[Cho et al, 2014 Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation]

Sequence to sequence

Machine Translation

- Bilingual: $\mathrm{p}(Y^n|X^n, \theta)$
- Multilingual: $p(Y^n|X^n, L^k, \theta)$

Automatic Post-editing: $p(Z^n|X^n, Y^n, \theta)$

- Single source/encoder
- Multi-source

Quality estimation: $p(Y|X^n, \theta), Y \in [0, 1]$

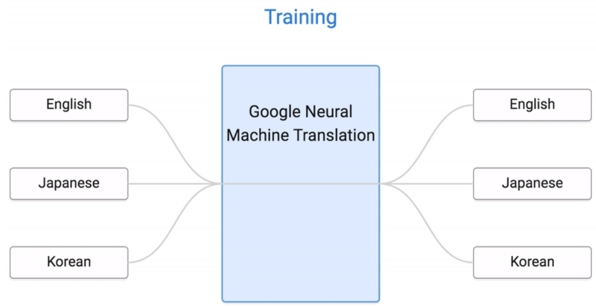
- Equivalent encoders
- Different encoders

Cross lingual text entailment: $p(Y|X^n, \theta)$,
$Y \in \{entails, contradicts, none\}$

Zero Shot Translation

Google

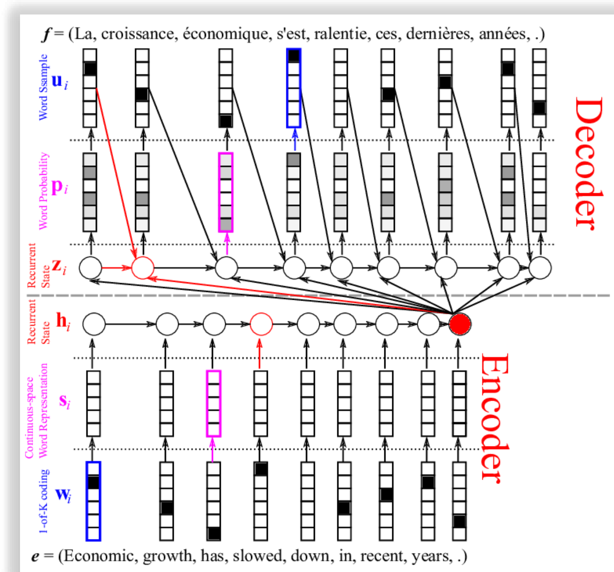
- Multilingual NMT with no parallel data
- Indicate target language $< 2ko >$



KantanMT

- Multilingual NMT with and without parallel data
- Low resource scenarios
- Indicate target language $< 2ko >$
- Indicate source language $< 2hi >$

Engine:	BLEU*	F-Measure*
ZST_2	0.21	3.26
ZST_3	9.78	26.40
one-to-one$_1$	8.20	22.16
$Pivot_3$ + $Pivot_4$	0.16	16.94



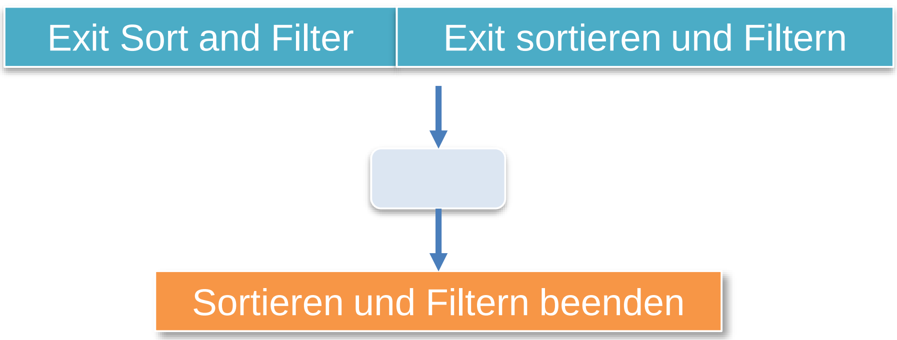
[https://ai.googleblog.com/2016/11/zero-shot-translation-with-googles.html]
[Mattoni et al, Zero-Shot Translation for Indian Languages with Sparse Data, MT Summit 2017]

Automatic post editing (APE or NPE)

Automatic post editing

- Given source and MT output generate improved translation

Single encoder

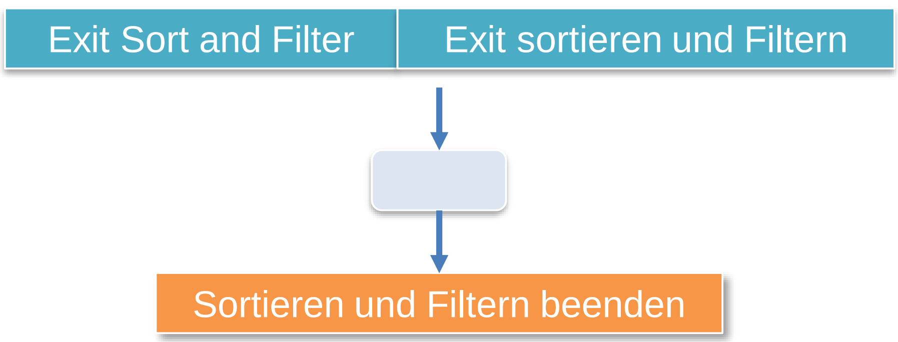


Automatic post editing (APE or NPE)

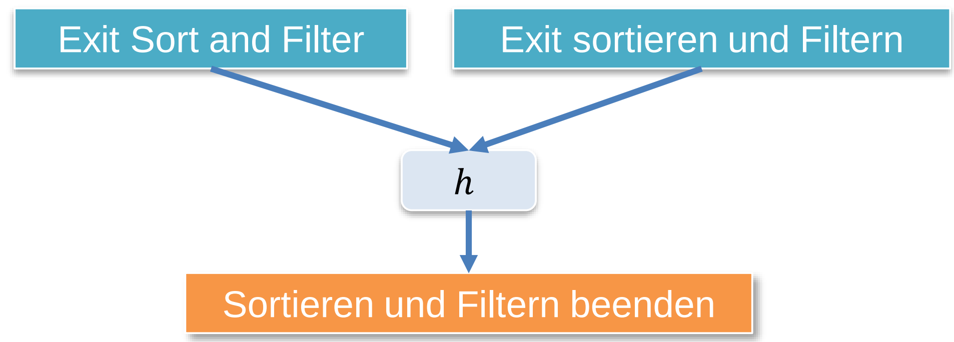
Automatic post editing

- Given source and MT output generate improved translation

Single encoder



Multiple encoders



$$h = \tanh(W_c, \left[\frac{\sum_{i=1}^{T^1} h_i^1}{T^1}; \frac{\sum_{i=1}^{T^2} h_i^2}{T^2}\right])$$

[Barret Zoph, Kevin Knight, Multi-Source Neural Translation]

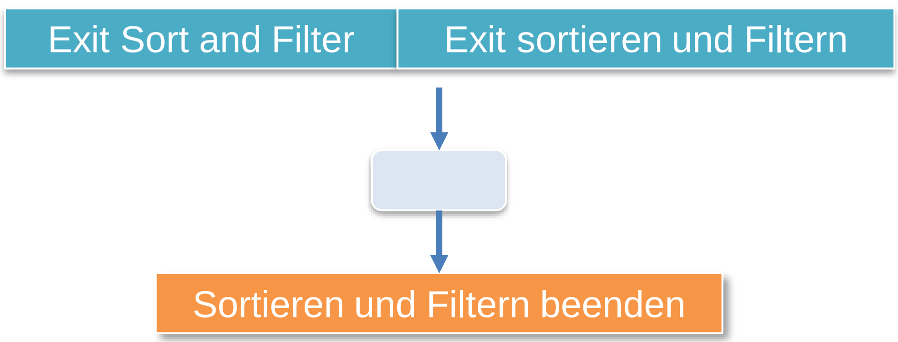
[Marcin Junczys-Dowmunt, Roman Grundkiewicz, An Exploration of Neural Sequence-to-Sequence Architectures for Automatic Post-Editing]

Automatic post editing (APE or NPE)

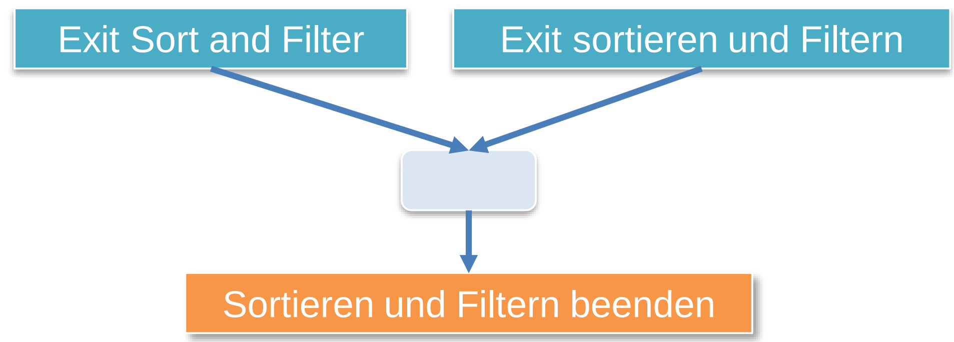
Automatic post editing

- Given source and MT output generate improved translation

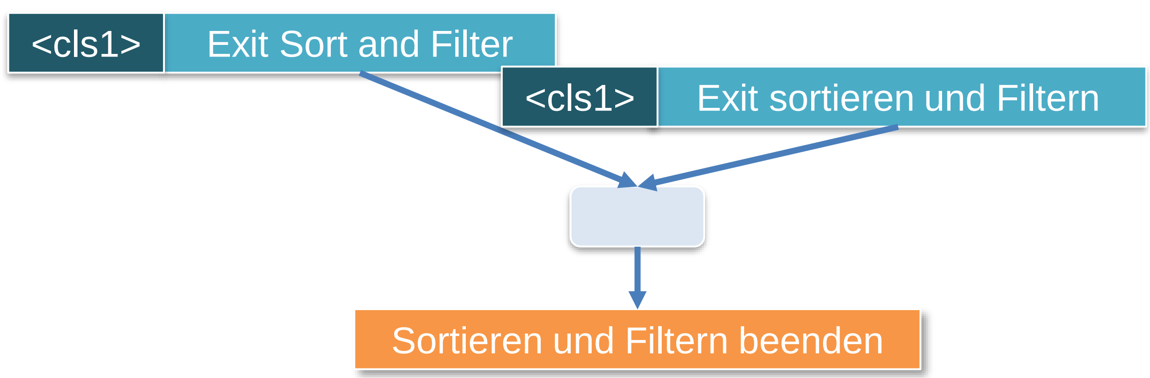
Single encoder



Multiple encoders



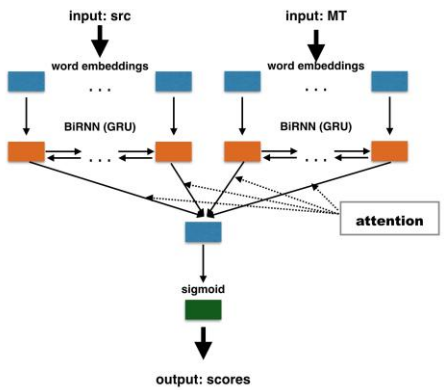
Multiple encoders with extra information

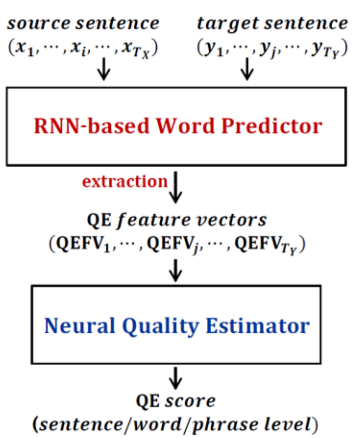


Quality Estimation and Cross lingual textual entailment

Quality estimation

- Given the source and MT output generate a quality score (TER)





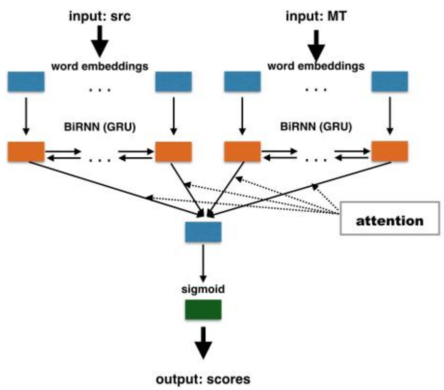
[Ive et al, deepQuest: A Framework for Neural-based Quality Estimation]

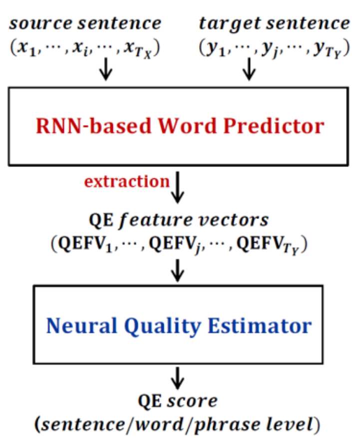
[Kim et al, Predictor-Estimator: Neural Quality Estimation Based on Target Word Prediction for Machine Translation]

Quality Estimation and Cross lingual textual entailment

Quality estimation

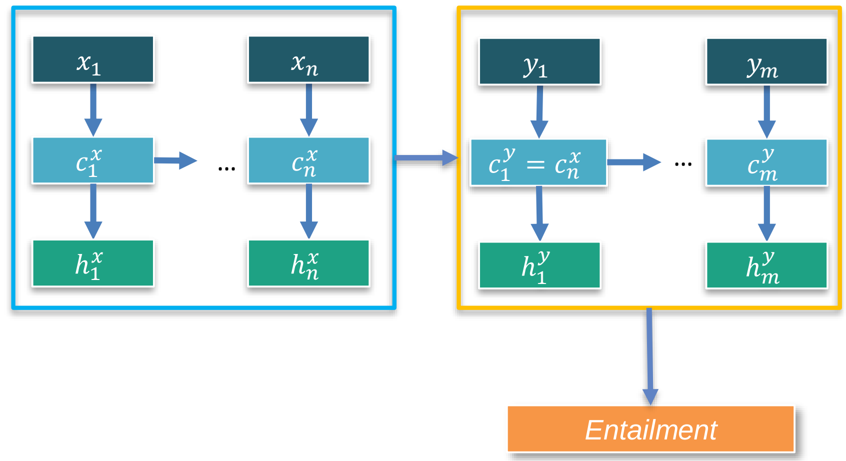
- Given the source and MT output generate a quality score (TER)





Cross lingual textual entailment

- Given two sentences (one in language L1 another in language L2) predict entailment



[Rocktäschel et al, Reasoning about entailment with Neural Attention]

[Ive et al, deepQuest: A Framework for Neural-based Quality Estimation]

[Kim et al, Predictor-Estimator: Neural Quality Estimation Based on Target Word Prediction for Machine Translation]

Takeaway

- Encoder – decoder architectures provide solutions for a large set of NLP (and others) problems.
- Model reusability is a bonus.

- Parallel data is not always necessary to do MT, but always helpful.