



**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

## Autoencoders

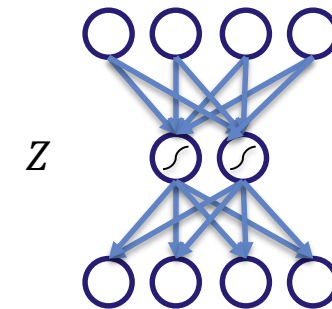
- Suppose we have a set of **multi-dimensional** data points  $X = \{x^1, x^2, \dots, x^m\}$ .
- Is there a general way to map  $X \rightarrow Z = \{z^1, z^2, \dots, z^m\}$ , where  $z$ 's have **lower dimensionality** than  $x$ 's and
- $Z$  can faithfully **reconstruct**  $X: Z \rightarrow \tilde{X}$

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

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

- Use stochastic gradient descent to minimize
- Autoencoders are unsupervised

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



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

## Autoencoders

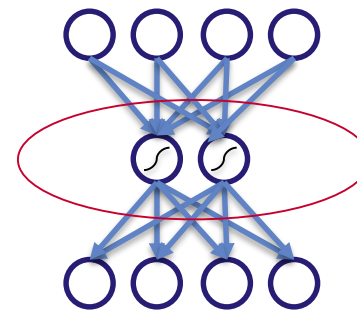
- Suppose we have a set of **multi-dimensional** data points  $X = \{x^1, x^2, \dots, x^m\}$ .
- Is there a general way to map  $X \rightarrow Z = \{z^1, z^2, \dots, z^m\}$ , where  $z$ 's have **lower dimensionality** than  $x$ 's and
- $Z$  can faithfully **reconstruct**  $X: Z \rightarrow \tilde{X}$

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

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

- Use stochastic gradient descent to minimize
- Autoencoders are unsupervised

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



Data Compression

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

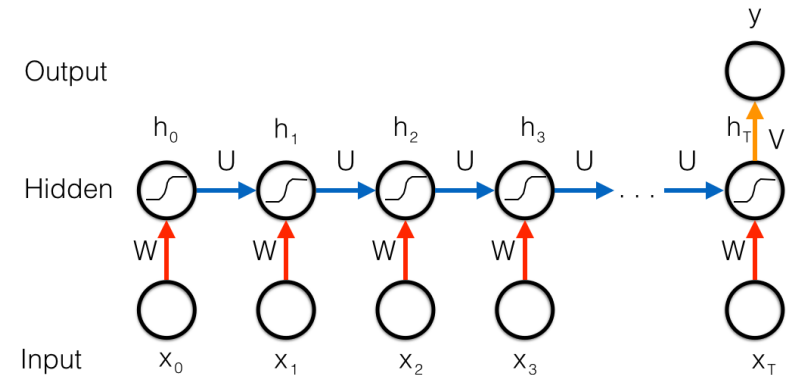
## Sequences

- $N \rightarrow 1$

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

- $N \rightarrow M$

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



## Sequences

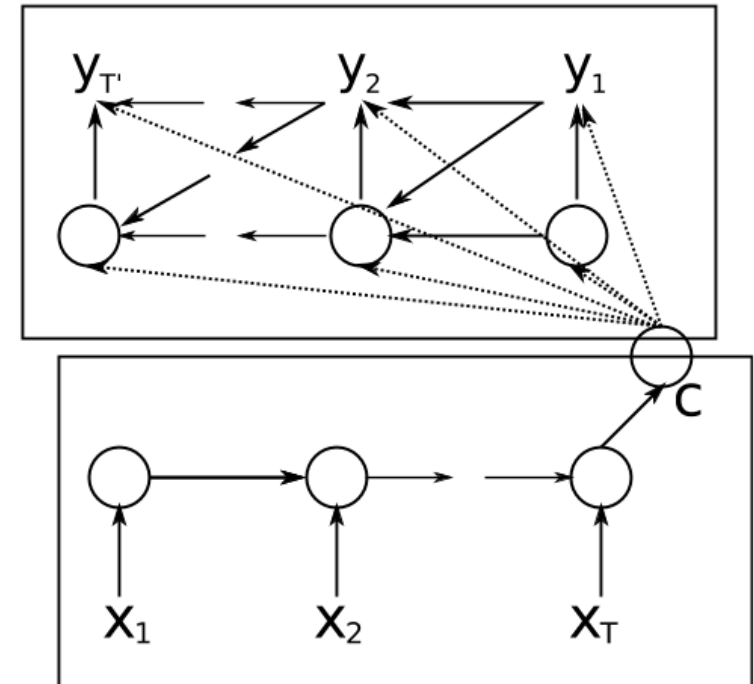
### - $N \rightarrow 1$

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

### - $N \rightarrow M$

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

## Decoder



## Encoder

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

## Sequences

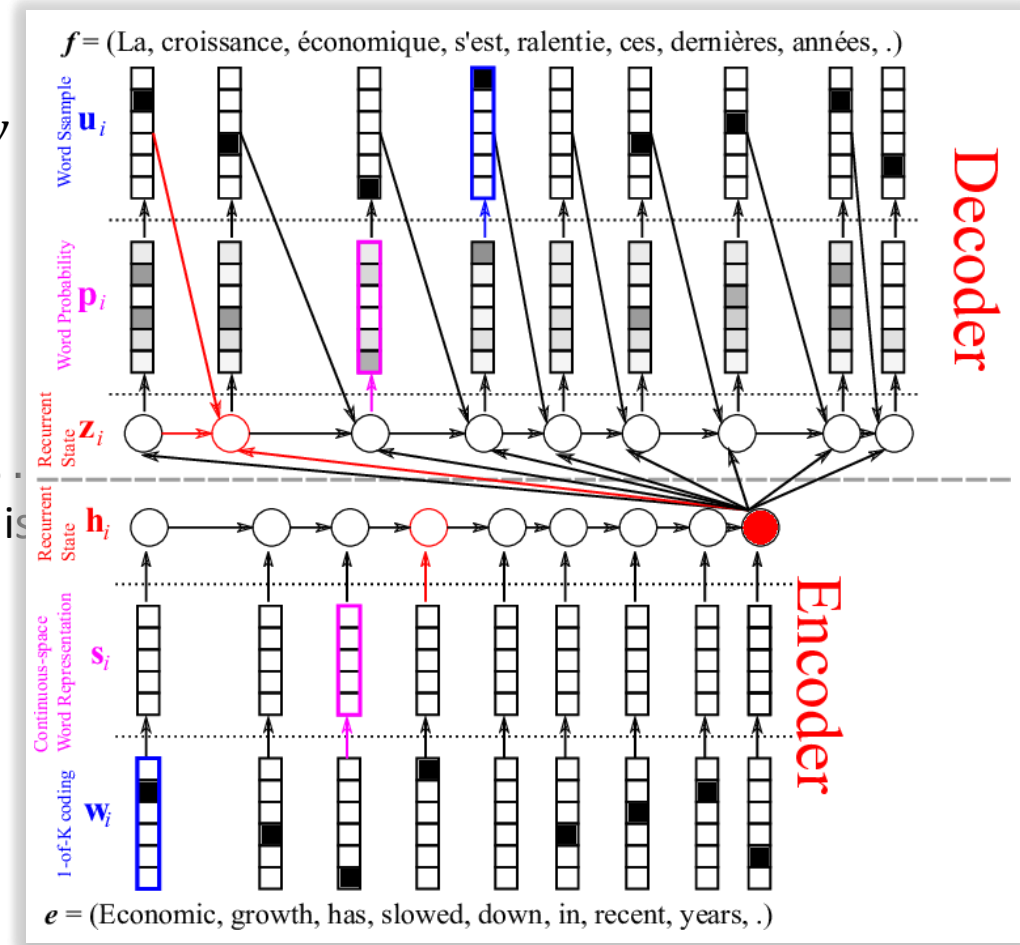
- N→1

Language modelling:  $X = \{x^1, x^2, \dots, x^{T-1}\}, y$   
 $x^i$  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 sentence in the target language

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



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



## Sequences

- N→1

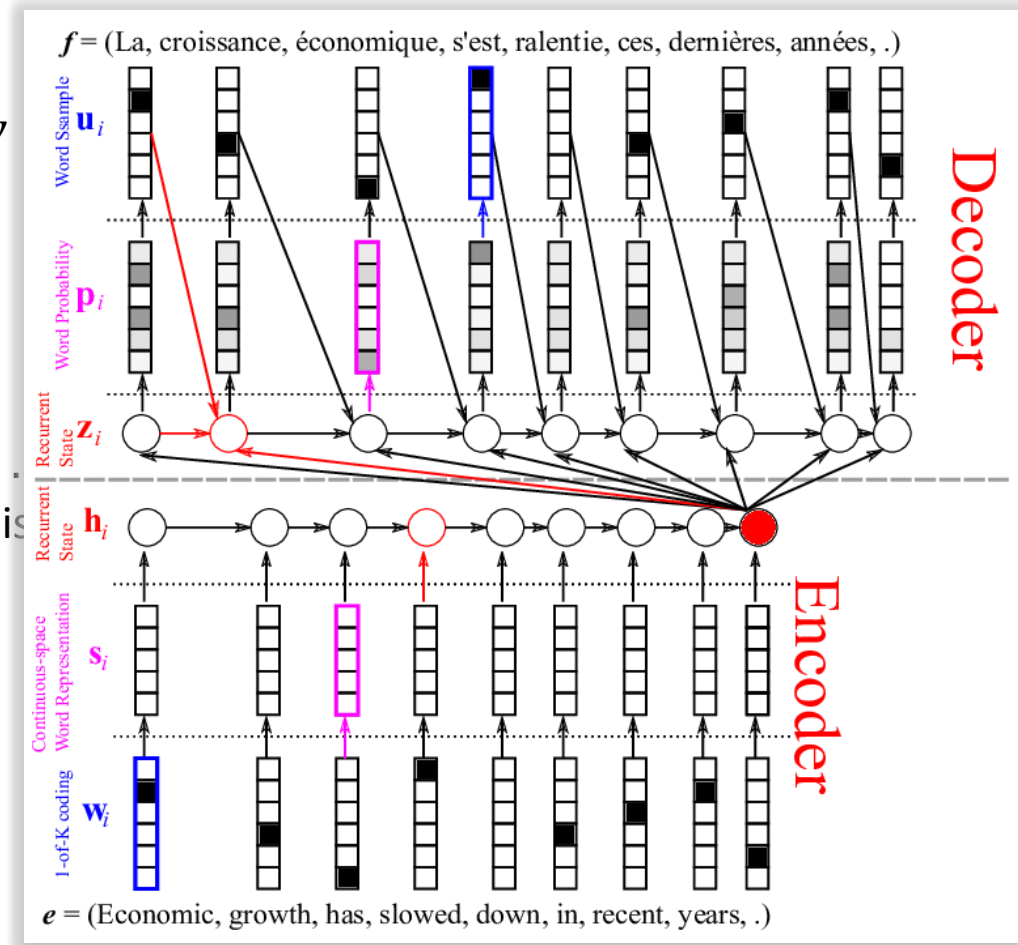
Language modelling:  $X = \{x^1, x^2, \dots, x^{T-1}\}, y$   
 $x^i$  is the words  $i$ ,  $T$  is current word.

- 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 sentence in the target language

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

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



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

# Sequence to sequence

## Machine Translation

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

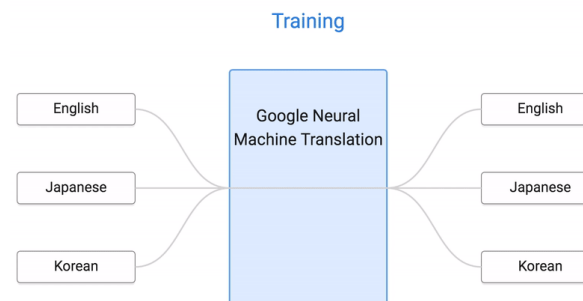
## Cross lingual text entailment: $p(Y|X^n, \theta),$ $Y \in \{entails, contradicts, none\}$



# Zero Shot Translation

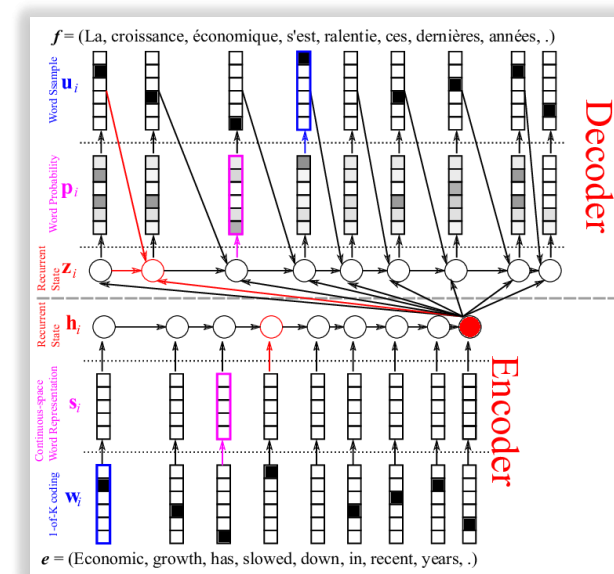
## Google

- Multilingual NMT with no parallel data
- Indicate target language  $\langle 2ko \rangle$



## KantanMT

- Multilingual NMT with and without parallel data
- Low resource scenarios
- Indicate target language  $\langle 2ko \rangle$
- Indicate source language  $\langle 2hi \rangle$



Engine:	BLEU*	F-Measure*
ZST <sub>2</sub>	0.21	3.26
ZST <sub>3</sub>	9.78	26.40
one-to-one <sub>1</sub>	8.20	22.16
Pivot <sub>3</sub> + Pivot <sub>4</sub>	0.16	16.94

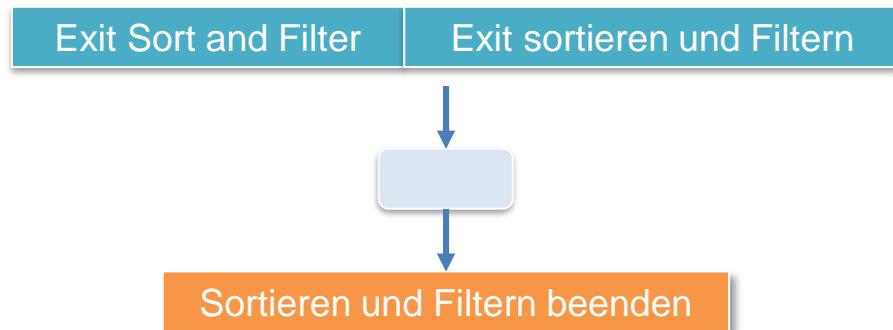
[<https://ai.googleblog.com/2016/11/zero-shot-translation-with-googles.html>]

[Mattoni et al, Zero-Shot Translation for Indian Languages with Sparse Data, MT Summit 2017]

## Automatic post editing

- Given source and MT output generate improved translation

Single encoder

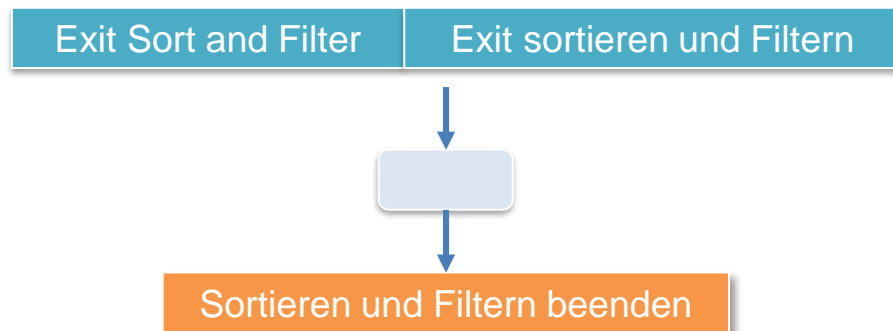


# Automatic post editing (APE or NPE)

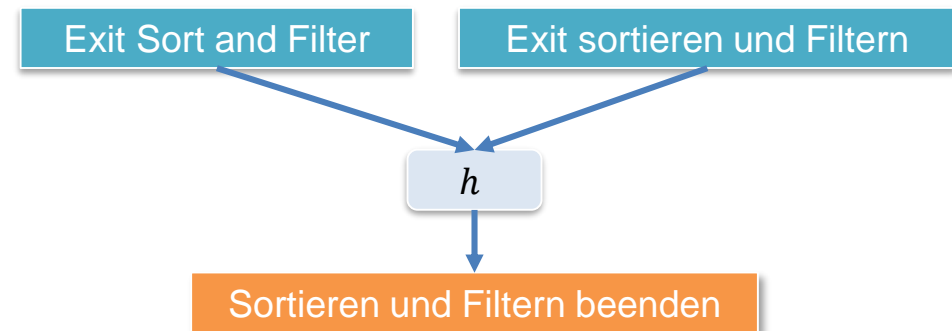
## Automatic post editing

- Given source and MT output generate improved translation

### Single encoder



### Multiple encoders



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

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

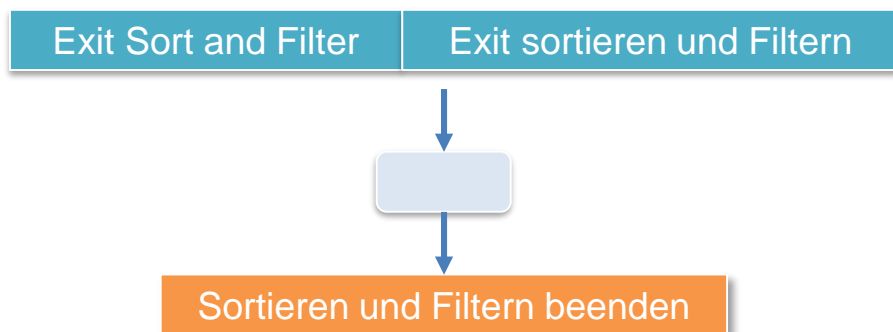
[Marcin Junczys-Dowmunt, Roman Grundkiewicz, An Exploration of Neural Sequence-to-Sequence Architectures for Automatic Post-Editing]

# Automatic post editing (APE or NPE)

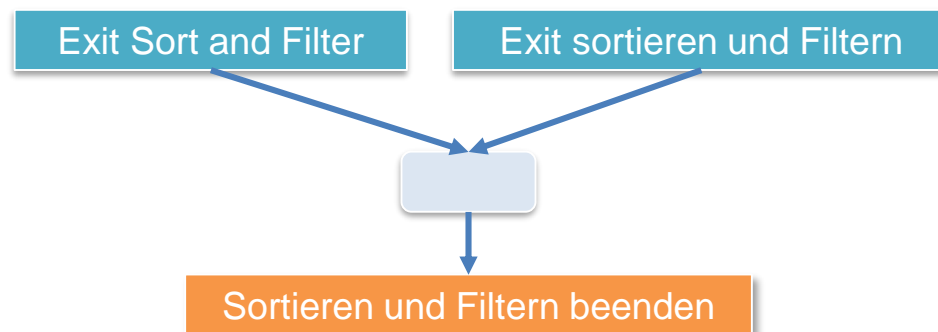
## Automatic post editing

- Given source and MT output generate improved translation

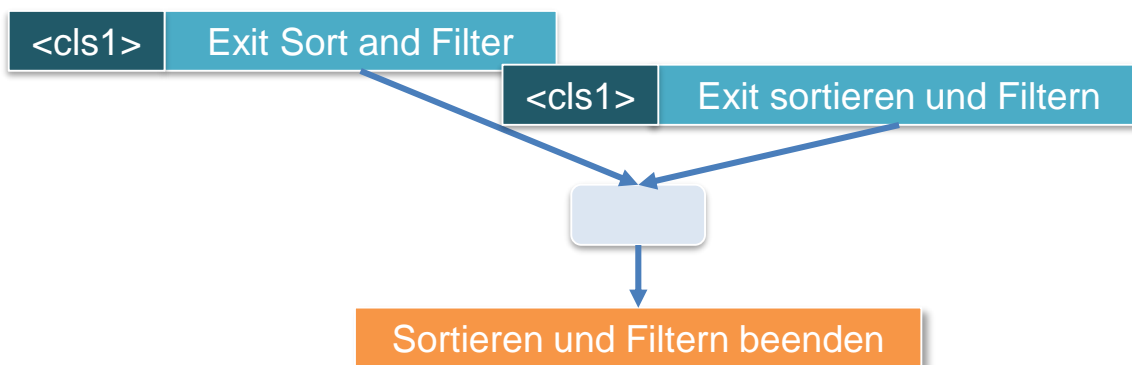
### Single encoder



### Multiple encoders

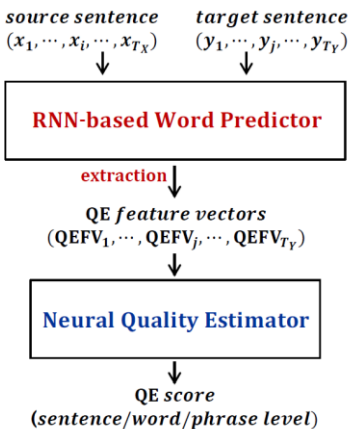
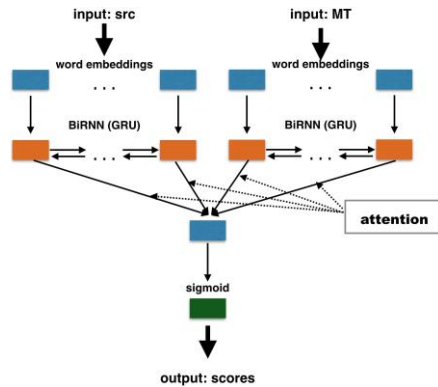


### Multiple encoders with extra information



## Quality estimation

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



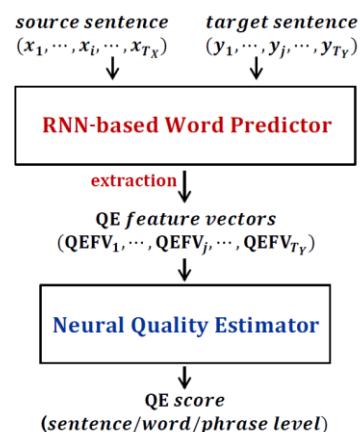
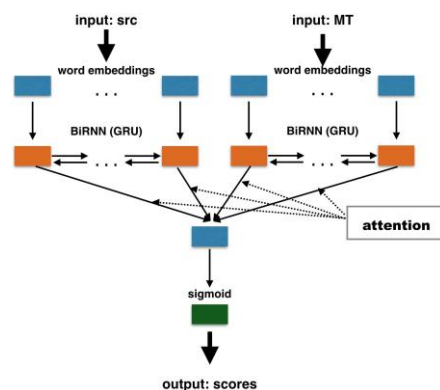
[Ive et al, deepQuest: A Framework for Neural-based Quality Estimation]

[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)



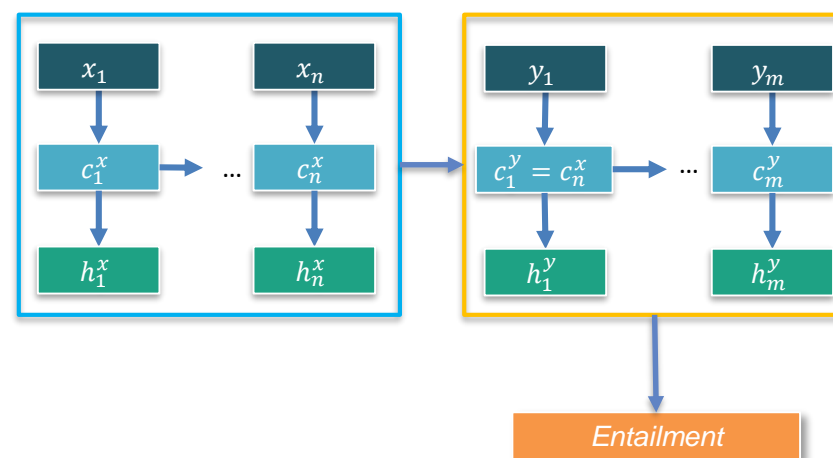
[Ive et al, deepQuest: A Framework for Neural-based Quality Estimation]

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

A World Leading SFI Research Centre

## 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]



- 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.

