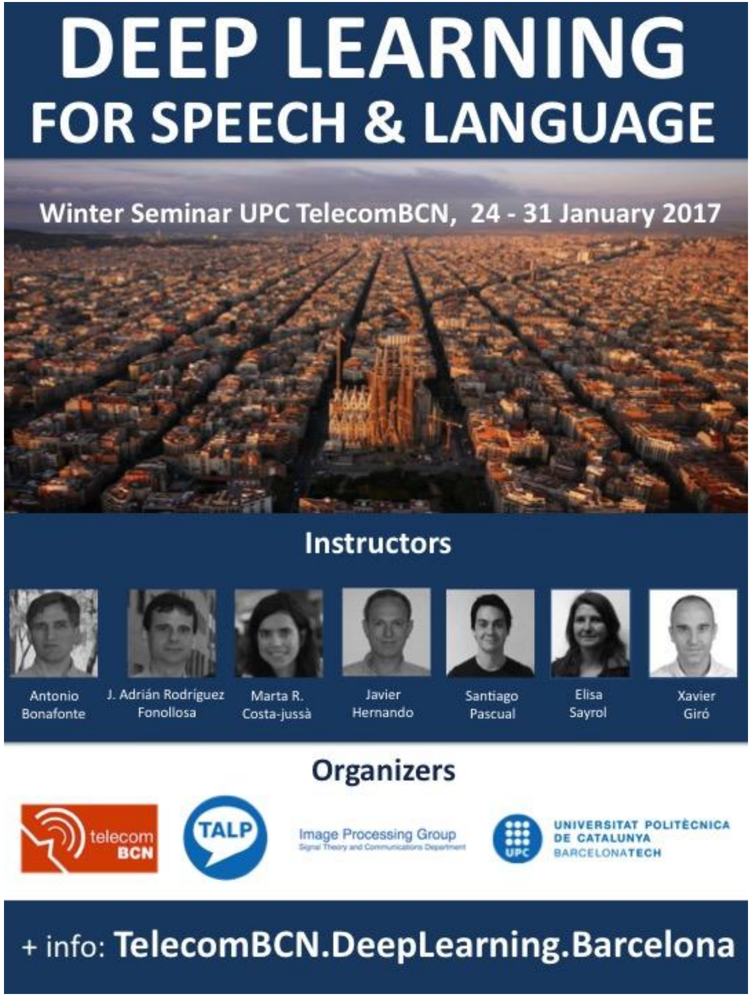


[course site]

Day 4 Lecture 2

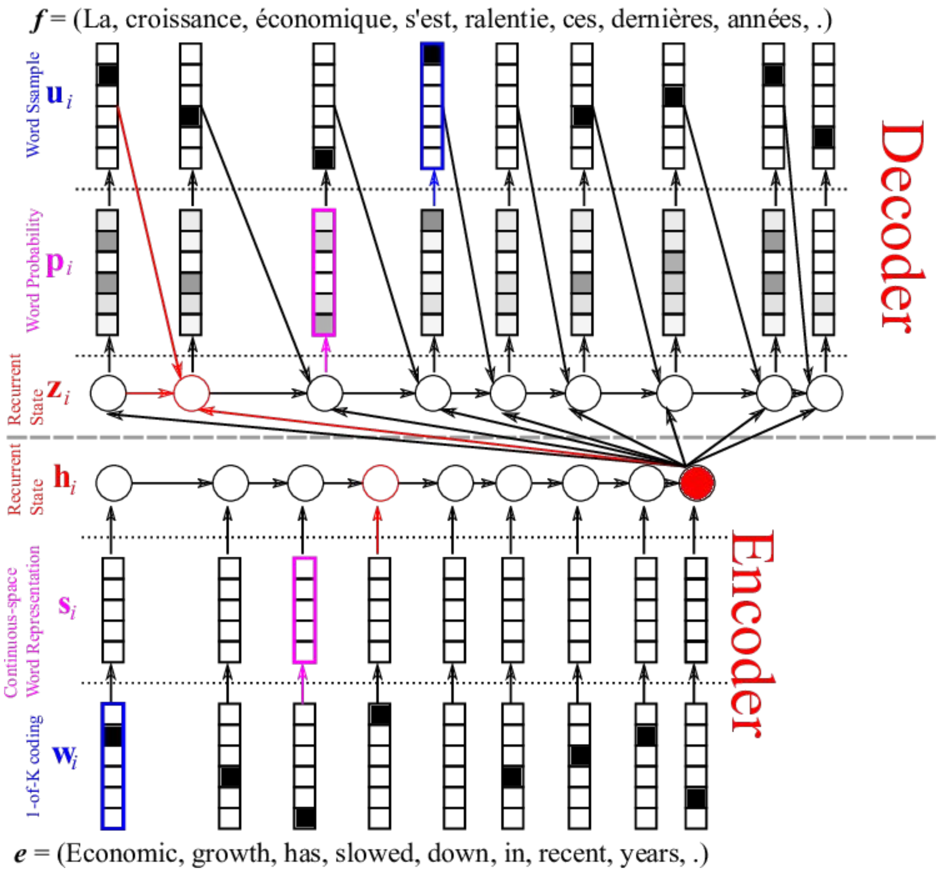
Advanced Neural Machine Translation

Marta R. Costa-jussà

Acknowledgments

Kyunghyun Cho, NVIDIA BLOGS:
https://devblogs.nvidia.com/parallelforall/introduction-neural-machine-translation-with-gpus/

From previous lecture...



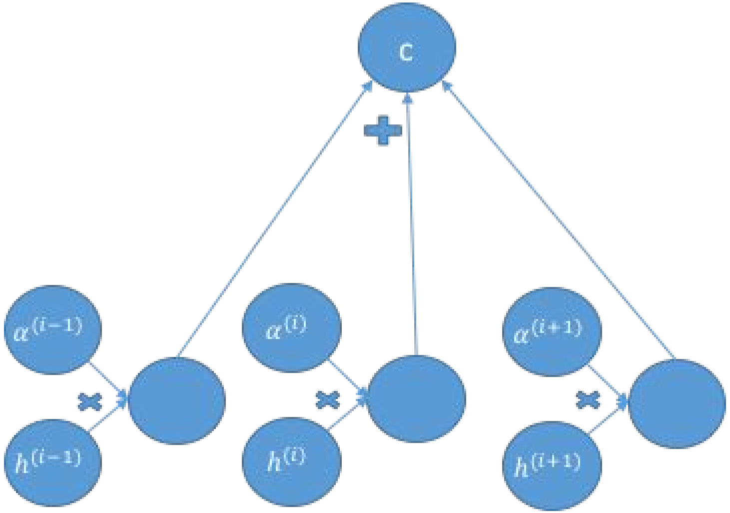
Kyunghyun Cho, [“Introduction to Neural Machine Translation with GPUs”](#) (2015)

Attention-based mechanism

Read the whole sentence, then produce the translated words one at a time, each time focusing on a different part of the input sentence

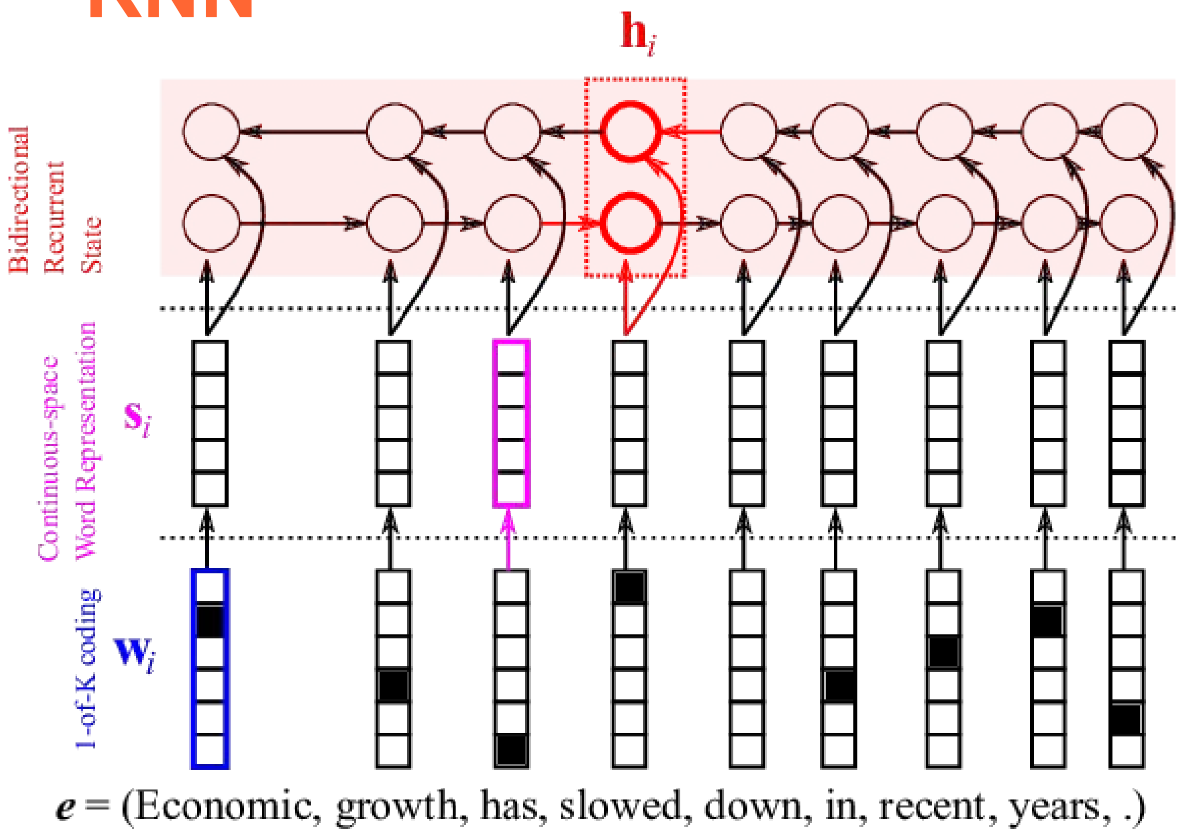
Encoder with attention: context vector

GOAL: Encode a source sentence into a set of context vectors



http://www.deeplearningbook.org/contents/applications.html

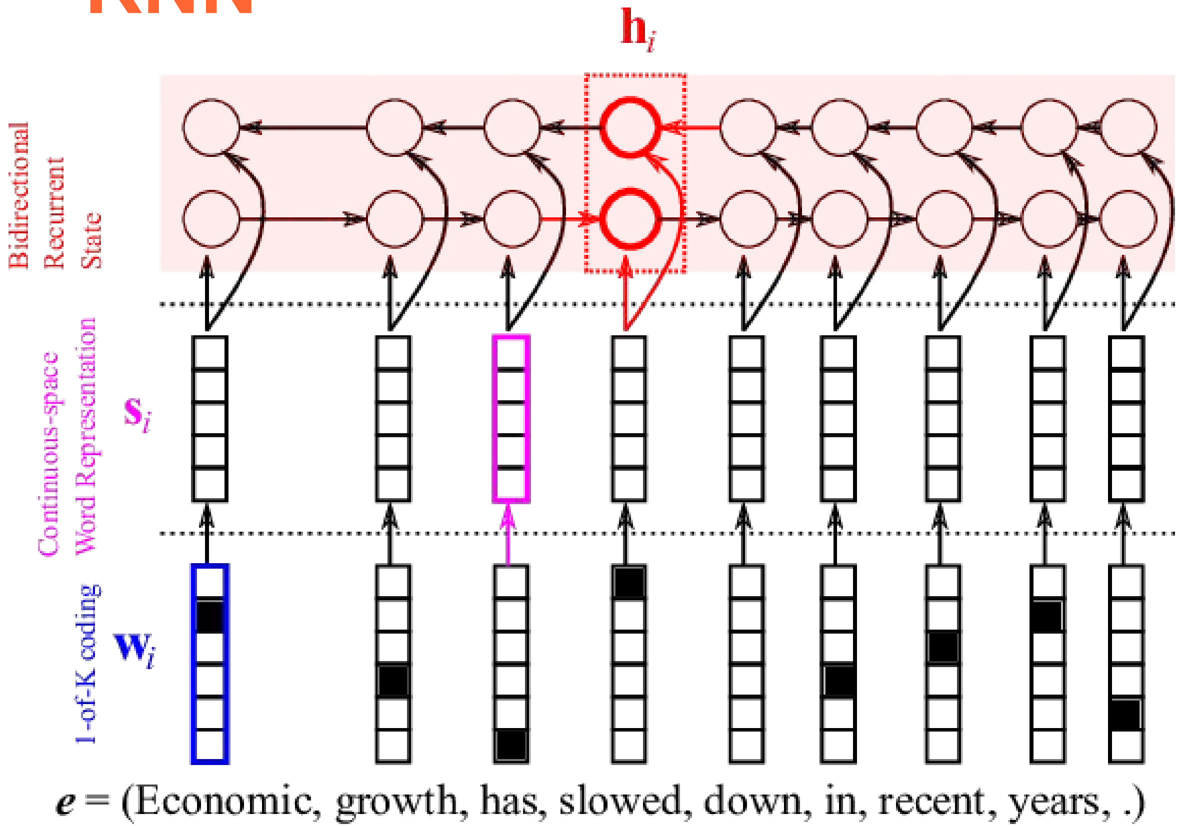
Composing the context vector: bidirectional RNN



$$\vec{h}_i = \phi_\theta(\vec{h}_{i-1}, s_i)$$

$$\{\vec{h}_1, \vec{h}_2, \ldots, \vec{h}_{T_x}\}$$

Composing the context vector: bidirectional RNN



$$\overleftarrow{h}_i = \phi_\theta(\overleftarrow{h}_{i-1}, s_i)$$

$$\{\overleftarrow{h}_1, \overleftarrow{h}_2, \ldots, \overleftarrow{h}_{T_x}\}$$

Decoder with attention

- The context vector now concatenates forward and reverse encoding vectors
- The decoder generates one symbol at a time based on this new context set

To compute the new decoder memory state, we must get one vector out of all context vectors.

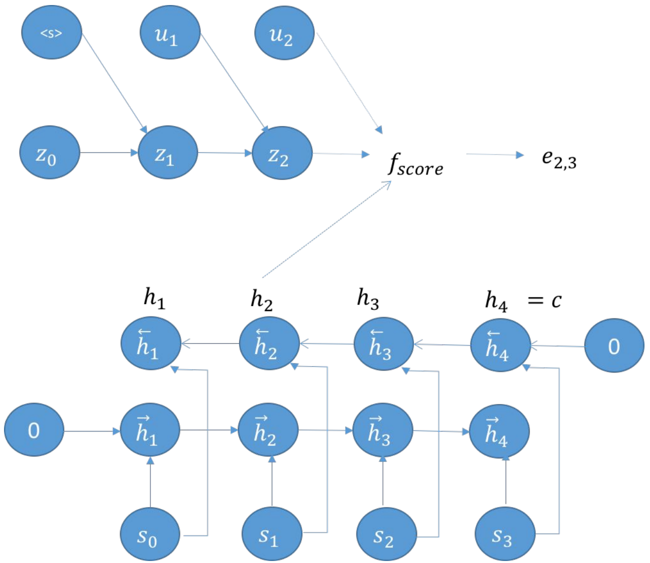
Compute the context vector

Each time step *t*, *ONE* vector context (c_i) is computed based on the (1) previous hidden state of the decoder (z_{i-1}), (2) previously decoded symbol (u_{i-1}), (3) whole context set (C)

Score each context vector based on how relevant it is for translating the next target word

This scoring (h_j, *j=1…T_x*) is based on the previous memory state, the previous generated target word and the j-th context vector

$$e_{j,i} = f_{score}(z_{i-1}, u_{i-1}, h_j)$$

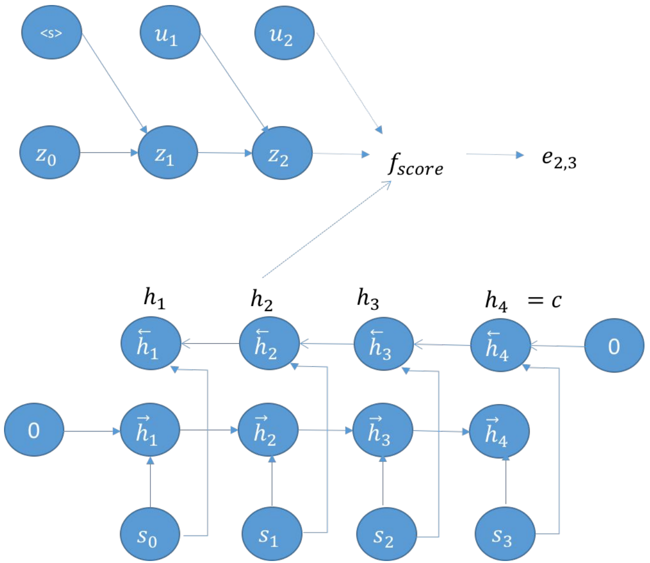


Score each context vector based on how relevant it is for translating the next target word

$$e_{j,i} = f_{score}(z_{i-1}, u_{i-1}, h_j)$$

fscore is usually a simple single-layer feedforward network

this relevance score measures how relevant the j-th context vector of the source sentence is in deciding the next symbol in the translation



Normalize relevance scores=attention weight

$$\alpha_{j,i} = \frac{\exp\left(e_{j,i}\right)}{\sum_{j'=1}^{T_x} \exp\left(e_{j,j'}\right)}$$

These attention weights correspond to how much the decoder attends to each of the context vectors.

Obtain the context vector c_i

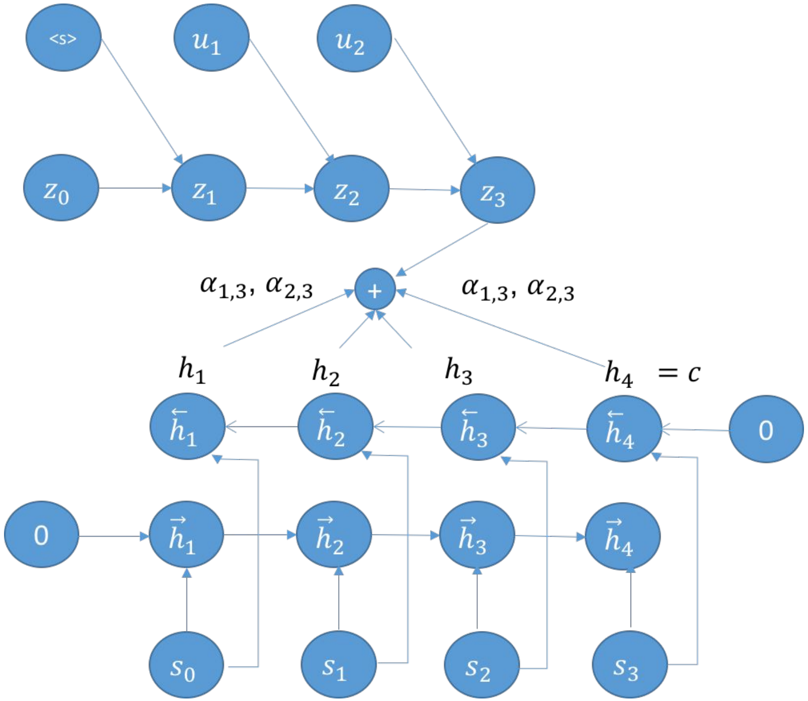
as the weighted sum of the context vectors with their weights being the attention weights

$$c_i = \sum_{j=1}^{T_x} \alpha_{i,j} h_j$$

Update the decoder's hidden state

$$z_i = \phi_{\theta'}(c_i, z_{i-1}, u_{i-1})$$

(The initial hidden state is initialized based on the last hidden state of the reverse RNN)



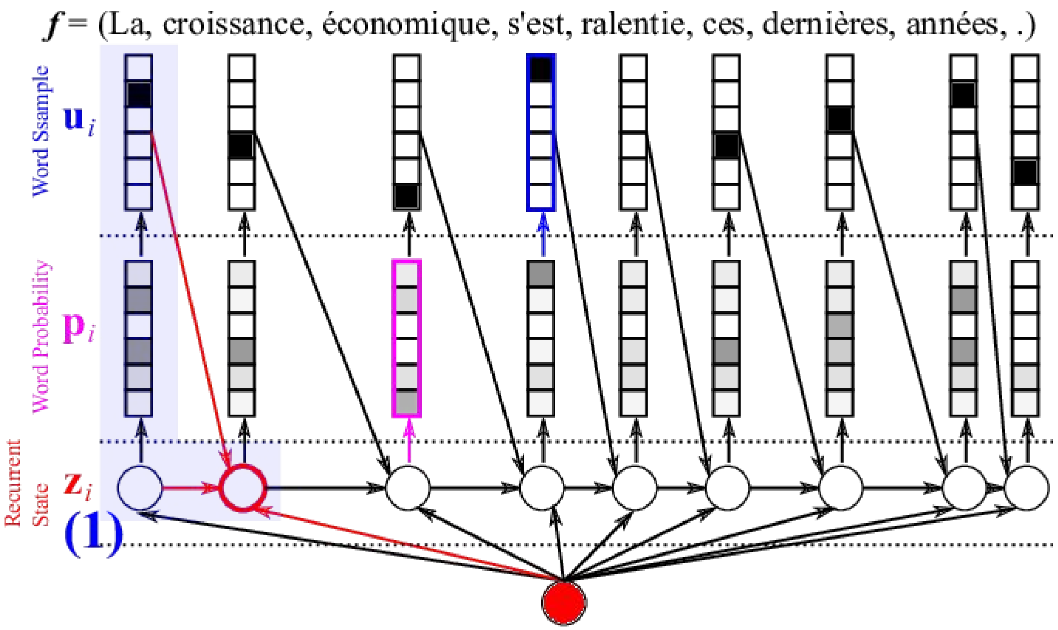
Decoder

From previous session

RNN's internal state z_i depends on: summary vector $h_{t'}$, previous output word u_{i-1} and previous internal state z_{i-1}.

NEW INTERNAL STATE

$$\boxed{z_i = \phi_{\theta'}(h_T, u_{i-1}, z_{i-1}).}$$



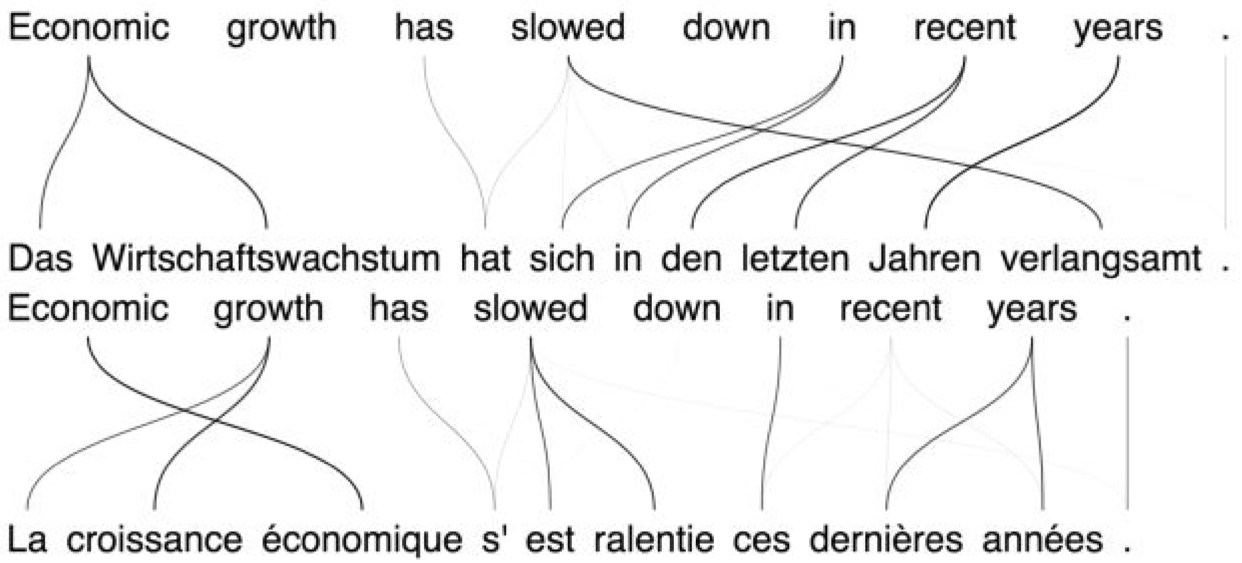
Kyunghyun Cho, "Introduction to Neural Machine Translation with GPUs" (2015)

Translation performances comparison

Model	BLEU
Simple Encoder-Decoder	17.82
+Attention-based	37.19
Phrase-based	37.03

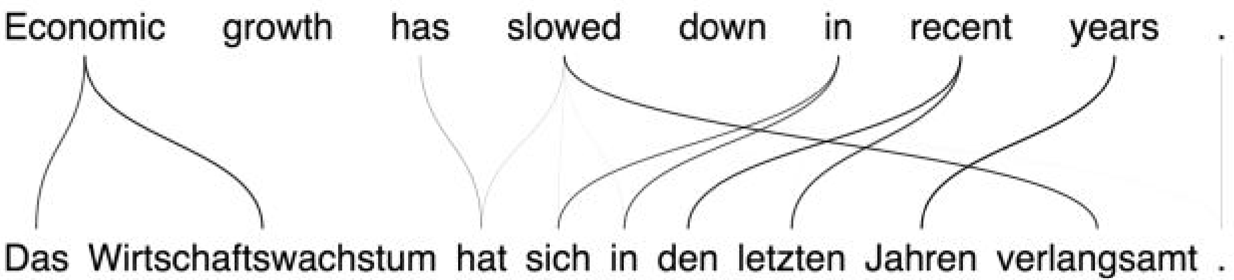
English-to-French WMT 2014 task

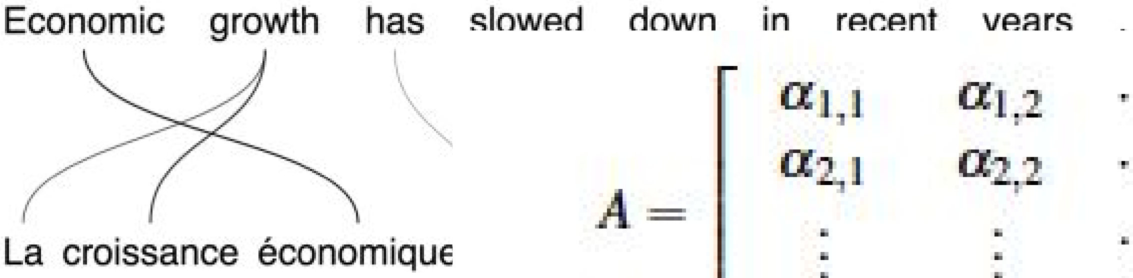
What attention learns… WORD ALIGNMENT



Kyunghyun Cho, "Introduction to Neural Machine Translation with GPUs" (2015)

What attention learns… WORD ALIGNMENT





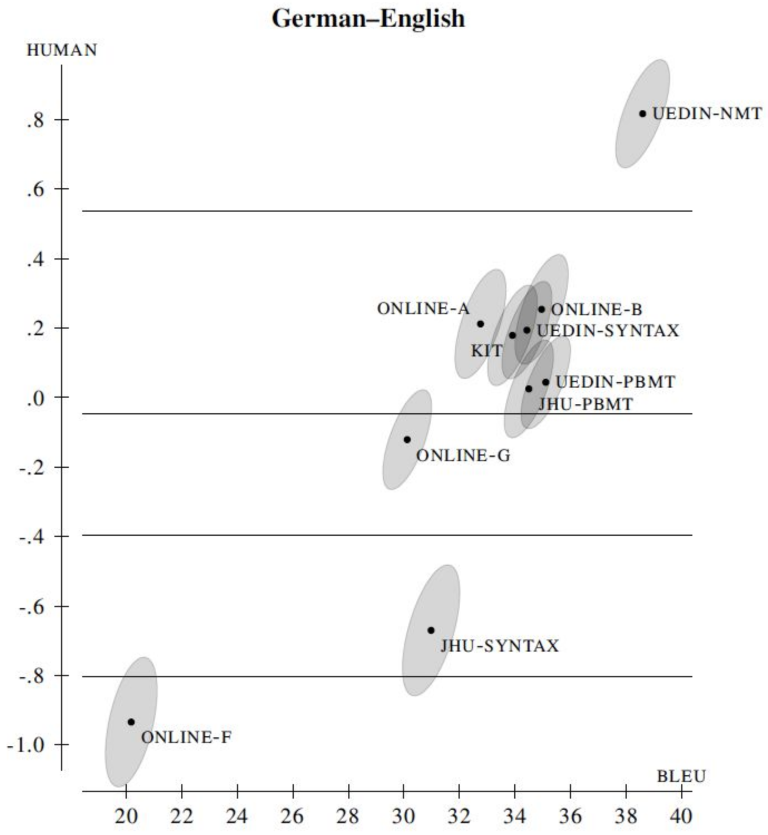
$$A = \begin{bmatrix} \alpha_{1,1} & \alpha_{1,2} & \cdots & \alpha_{1,|Y|} \\ \alpha_{2,1} & \alpha_{2,2} & \cdots & \alpha_{2,|Y|} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{|X|,1} & \alpha_{|X|,2} & \cdots & \alpha_{|X|,|Y|} \end{bmatrix},$$

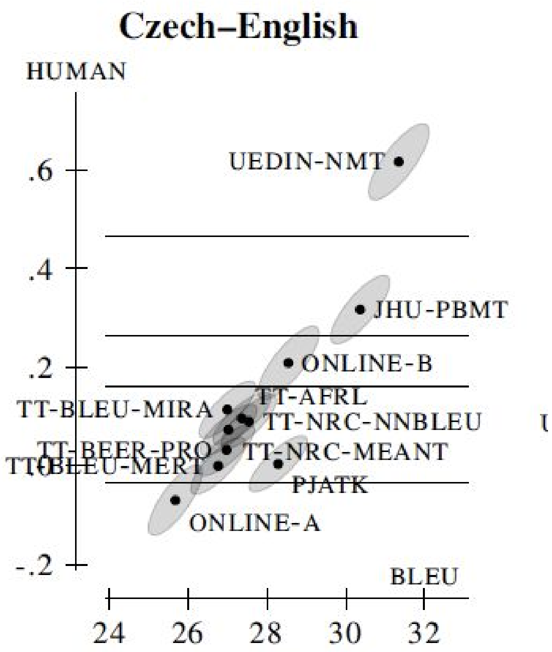
Kyunghyun Cho, [“Introduction to Neural Machine Translation with GPUs”] (2015)

Neural MT is better than phrase-based

[Neural Network for Machine Translation at Production Scale]

Results in WMT 2016 international evaluation



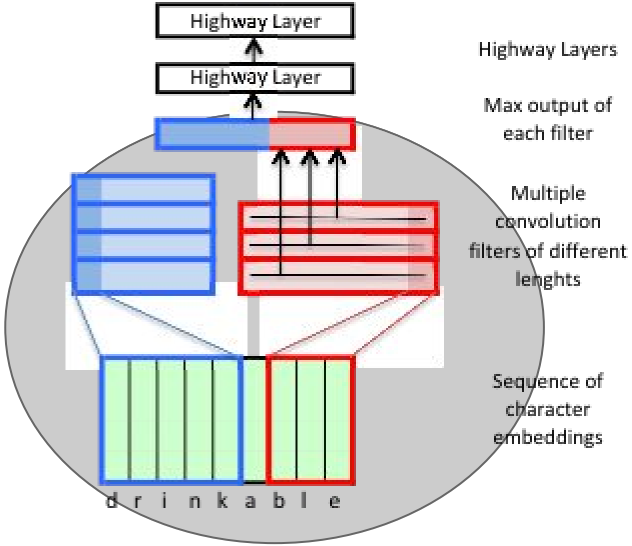


What Next?

Character-based Neural Machine Translation: Motivation

- Word embeddings have been shown to boost the performance in many NLP tasks, including machine translation.
- However, the standard look-up based embeddings are limited to a finite-size vocabulary for both computational and sparsity reasons.
- The orthographic representation of the words is completely ignored.
- The standard learning process is blind to the presence of stems, prefixes, suffixes and any other kind of affixes in words.

Character-based Neural MT: Proposal (Step 1)



Kim et al, 2015

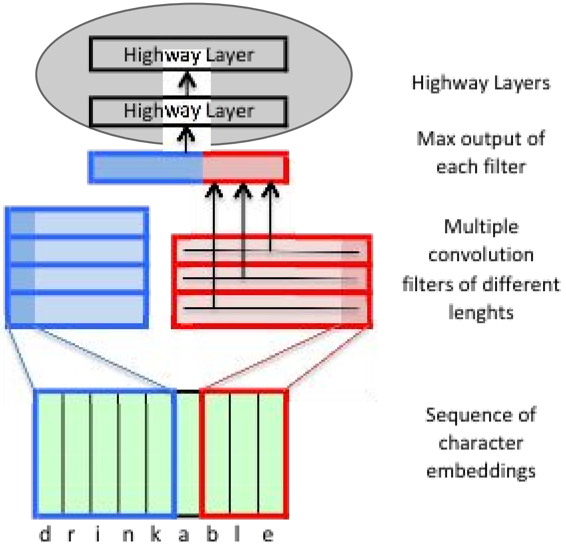
■The computation of the representation of each word starts with a character-based embedding layer that associates each word (sequence of characters) with a sequence of vectors.

■This sequence of vectors is then processed with a set of 1D convolution filters of different lengths followed with a max pooling layer.

■For each convolutional filter, we keep only the output with the maximum value. The concatenation of these max values already provides us with a representation of each word as a vector with a fixed length equal to the total number of convolutional kernels.

Character-based Neural MT: Proposal (Step 2)

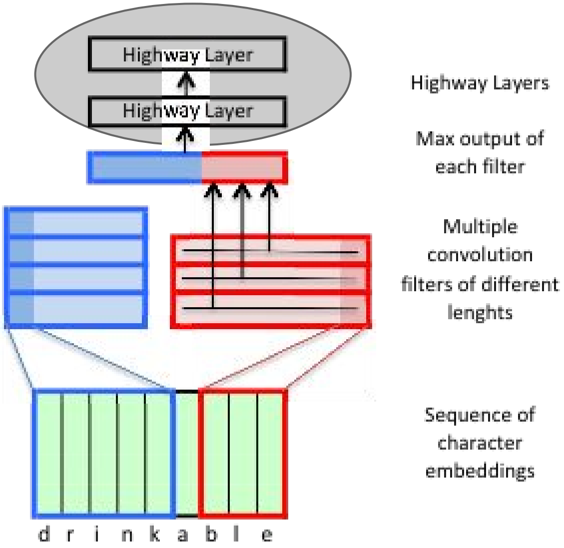
architecture designed to ease gradient-based training of deep networks



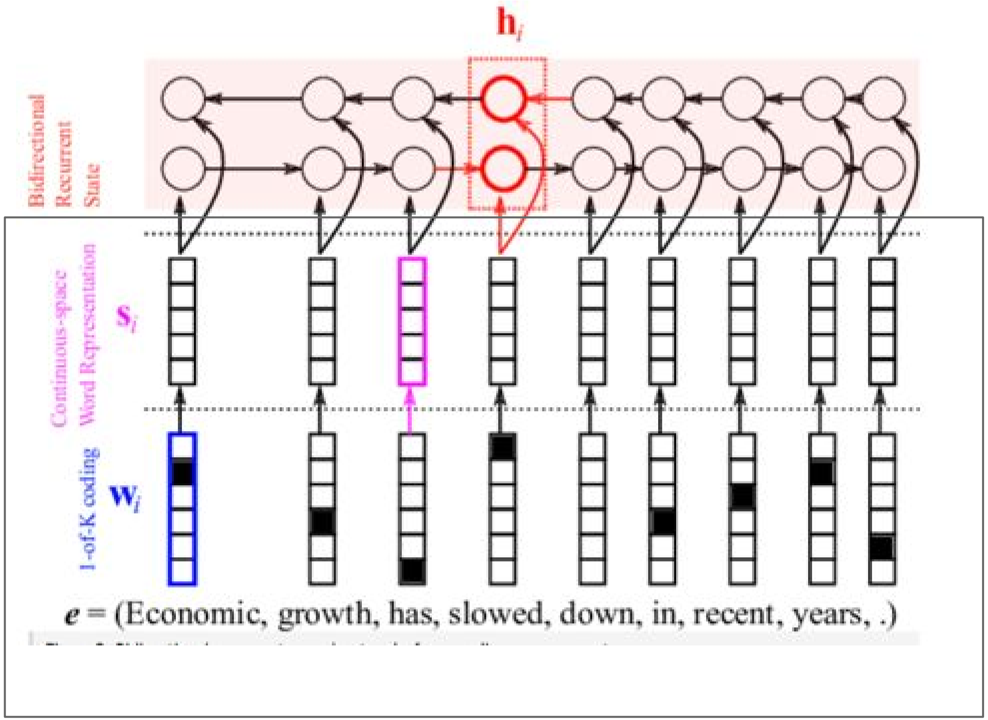
Kim et al, 2015

- The addition of two highway layers was shown to improve the quality of the language model in (Kim et al., 2016).
- The output of the second Highway layer will give us the final vector representation of each source word, replacing the standard source word embedding in the neural machine translation system.

Character-based Neural MT: Integration with NMT



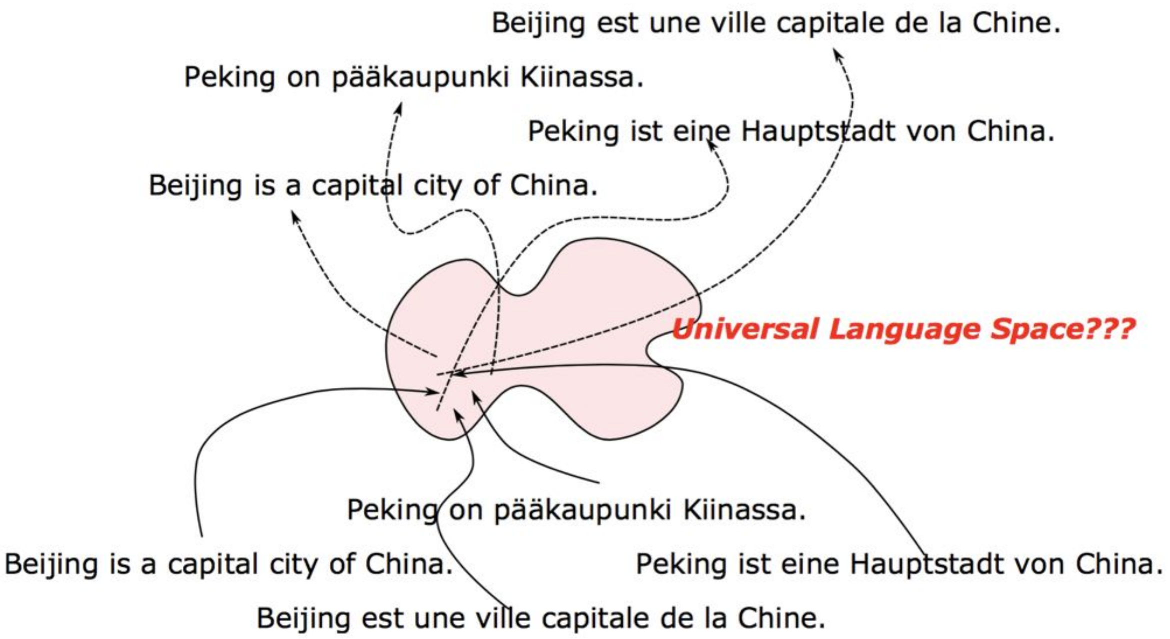
Kim et al, 2015



Examples

1	SRC	Berichten zufolge hofft Indien darber hinaus auf einen Vertrag zur **Verteidigungszusammenarbeit** zwischen den beiden Nationen .
	Phrase	reportedly hopes India , in addition to a contract for the defence cooperation between the two nations .
	NN	according to reports , India also hopes to establish a contract for the UNK between the two nations .
	CHAR	according to reports , India hopes to see a Treaty of **Defence Cooperation** between the two nations .
	REF	India is also reportedly hoping for a deal on **defence collaboration** between the two nations .
4	SRC	der durchtrainierte Mainzer sagt von sich , dass er ein " ambitionierter **Rennradler** " ist .
	Phrase	the will of Mainz says that he a more ambitious .
	NN	the UNK Mainz says that he is a " ambitious , . "
	CHAR	the UNK in Mainz says that he is a ' **ambitious racer** ' .
	REF	the well-conditioned man from Mainz said he was an " **ambitious racing cyclist** . "
3	SRC	die GDL habe jedoch nicht gesagt , wo sie **streiken** wolle , so dass es schwer sei , die Folgen konkret vorherzusehen .
	Phrase	the GDL have , however , not to say , where they strike , so that it is difficult to predict the consequences of concrete .
	NN	however , the UNK did not tell which they wanted to UNK , so it is difficult to predict the consequences .
	CHAR	however , the UNK did not say where they wanted to **strike** , so it is difficult to predict the consequences .
	REF	the GDL have not said , however , where they will **strike** , making it difficult to predict exactly what the consequences will be .
4	SRC	die Premierminister Indiens und Japans trafen sich in Tokio .
	Phrase	the Prime Minister of India and Japan in Tokyo .
	NN	the Prime Minister of India and Japan met in Tokyo
	CHAR	the Prime **Ministers** of India and Japan met in Tokyo
	REF	India and Japan prime **ministers** meet in Tokyo
5	SRC	wo die Beamten es aus den Augen verloren .
	Phrase	where the officials lost sight of
	NN	where the officials lost it out of the eyes
	CHAR	where officials **lose sight of it**
	REF	causing the officers to **lose sight of it**

Multilingual Translation



Kyunghyun Cho, "DL4MT slides" (2015)

Multilingual Translation Approaches

Sharing attention-based mechanism across language pairs

Orhan Firat et al, "Multi-way, Multilingual Neural Machine Translation with a Shared-based Mechanism" (2016)

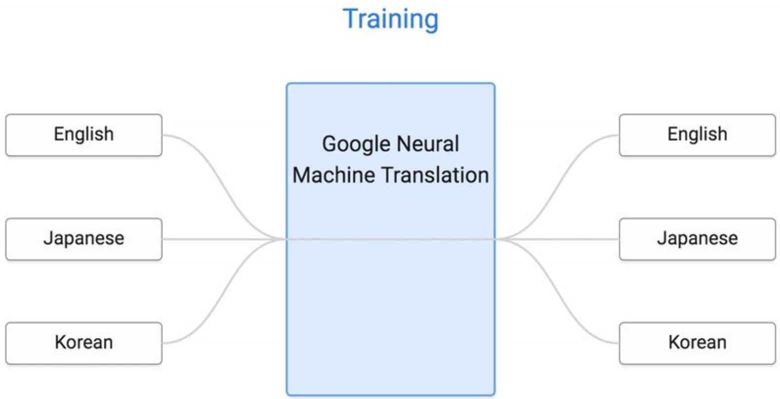
Multilingual Translation Approaches

Sharing attention-based mechanism across language pairs

Orhan Firat et al, "Multi-way, Multilingual Neural Machine Translation with a Shared-based Mechanism" (2016)

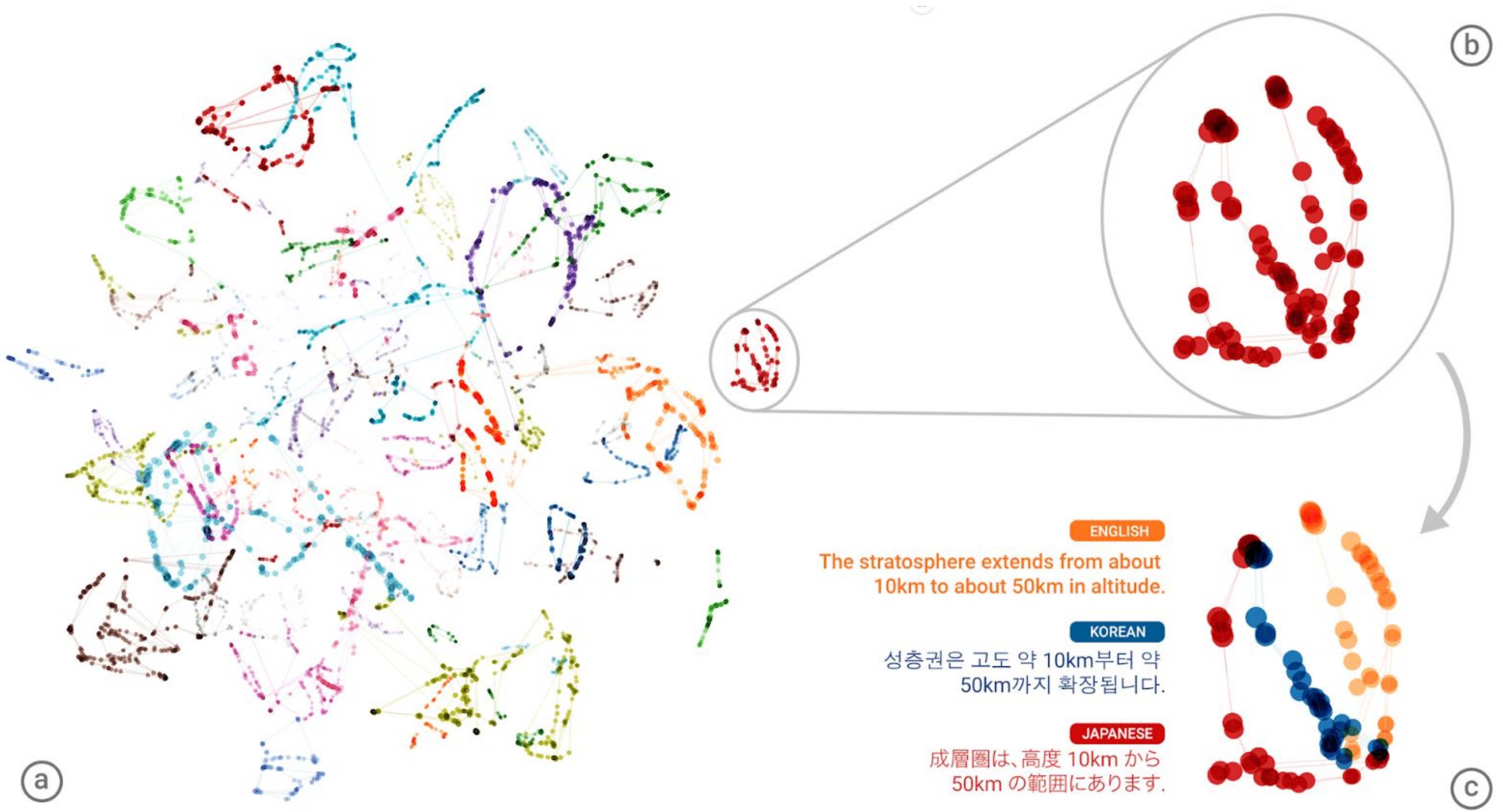
Share encoder, decoder, attention accross language pairs

Johnson et al, "Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation" (2016)



https://research.googleblog.com/2016/11/zero-shot-translation-with-googles.html

Is the system learning an Interlingua?



https://research.googleblog.com/2016/11/zero-shot-translation-with-googles.html

Available software on github

DL4MT

NEMATUS

Most publications have open-source code...

Summary

- Attention-based mechanism allows to achieve state-of-the-art results
- Progress in MT includes character-based, multilinguality...

Learn more

Natural Language Understanding with Distributed Representation, Kyunghyun Cho, Chapter 6, 2015 (available in github)

Thanks ! Q&A ?

https://www.costa-jussa.com

marta.ruiz@upc.edu