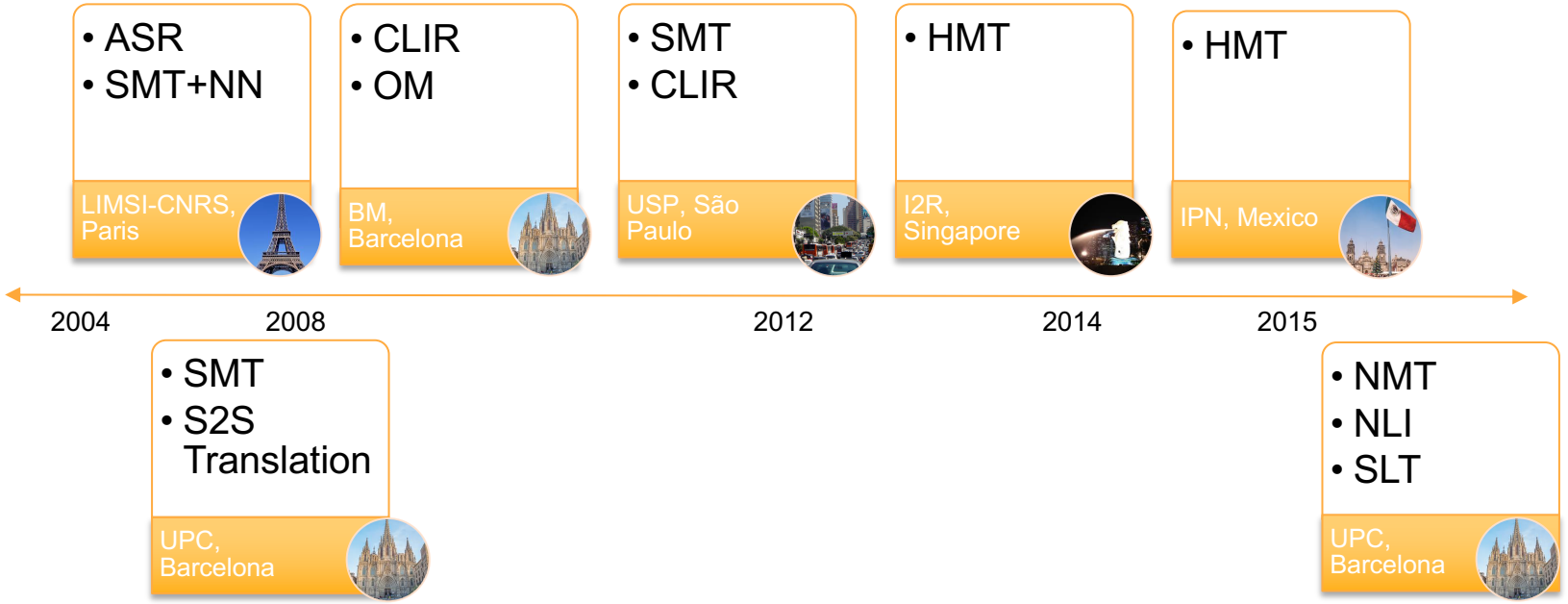
Neural Machine Translation: Basics



Marta R. Costa-jussà
marta.ruiz@upc.edu

Ramón y Cajal Researcher
Universitat Politecnica de Catalunya
Technical University of Catalonia

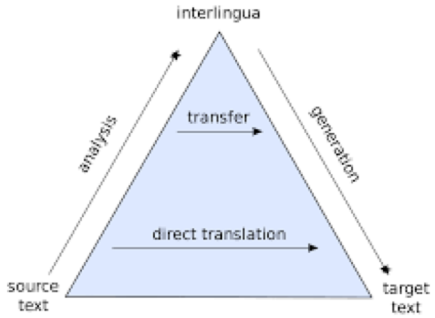
About me


• ASR
• SMT+NN
LIMSI-CNRS, Paris
• CLIR
• OM
BM, Barcelona
• SMT
• CLIR
USP, São Paulo
• HMT
I2R, Singapore
• HMT
IPN, Mexico
2004
2008
2012
2014
2015
• SMT
• S2S Translation
UPC, Barcelona
• NMT
• NLI
• SLT
UPC, Barcelona

Outline

- Machine translation background
- Neural Machine Translation (NMT): Basics
  - Motivation, Architecture
- Research advances in NMT
  - Characters, multimodality, multilinguality
- Conclusions

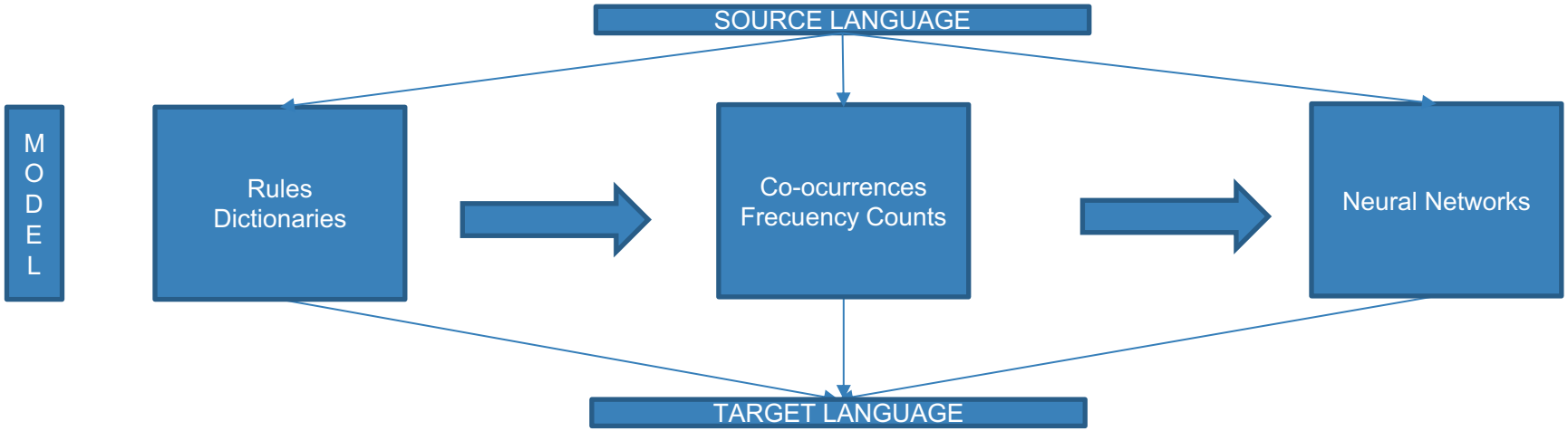
Machine Translation background



Machine Translation is the application that is able to automatically translate from source (S) to target (T).

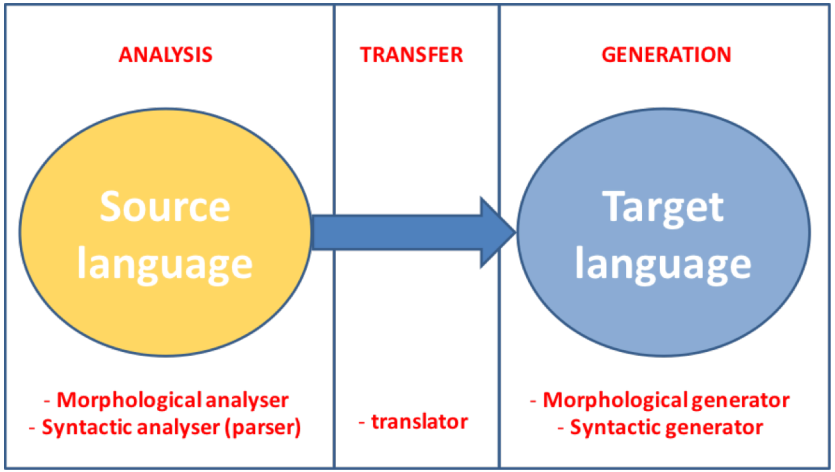
Main approaches have been either rule-based or statistical-based

MT Approaches

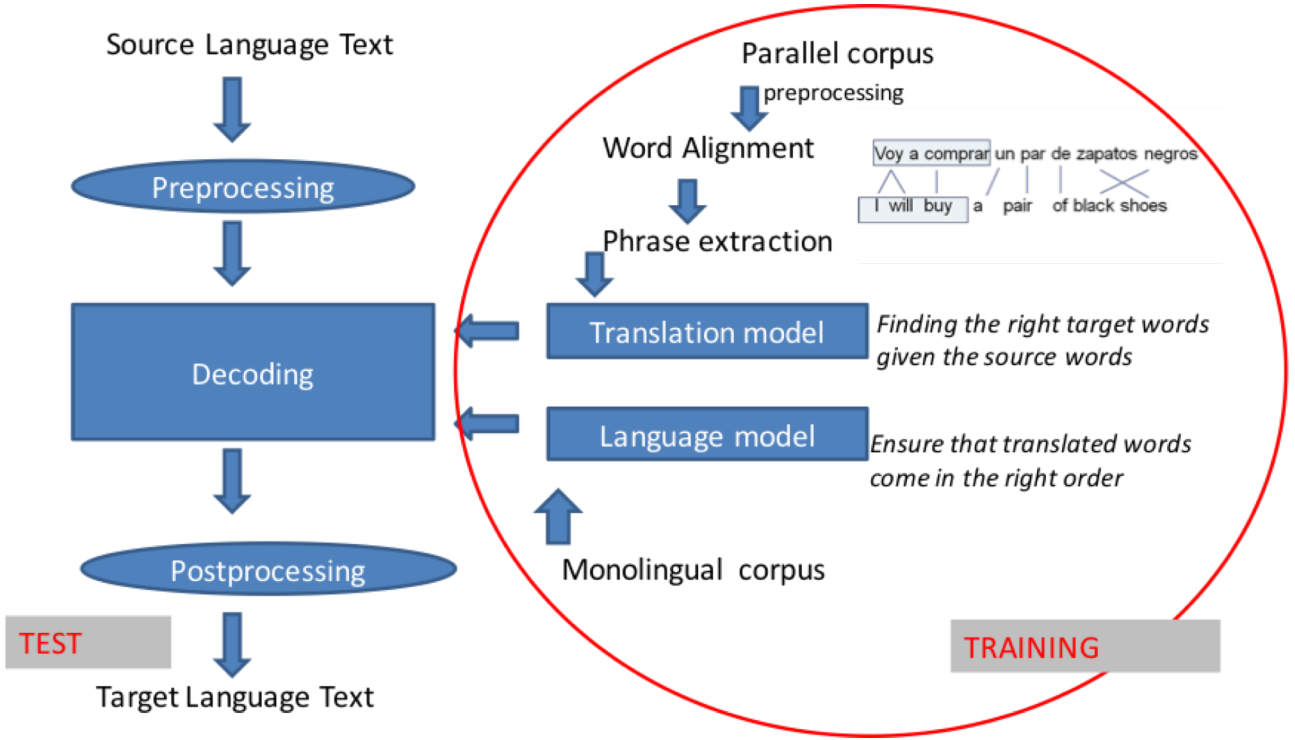


Dates Refs	From 1950s Eurotra, Apertium… (Forcada, 2005)	From 1990s TC-Star, Moses… (Koehn, 2010)	Starting in 2014… NEMATUS… (Cho, 2014)

Rule-based approach



Statistical-based approach



Why a new approach?

We need years to develop a nice rule-based approach

Regarding statistical systems:

(1) Word alignment and Translation are optimized separately

(2) Translation at the level of words, but difficulties with high variations in morphology (e.g. translation English-to-Finnish)

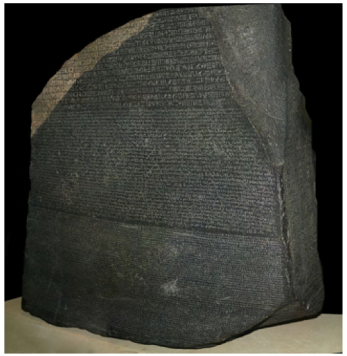
(3) Translation by language pairs
difficult to think of an automatic interlingua

Why Neural Machine Translation?

- Integrated MT paradigm
- Trainable at the subword/character level
- Multimodal/Multilingual advantages

What do we need?

- Parallel Corpus



English	Russian
This course is a thorough introduction to machine translation technology	Этот курс представляет собой интенсивное введение в технологию машинного перевода
We will describe all aspects of building a statistical machine translation system, from both formal and practical perspectives	Мы рассмотрим все аспекты построения системы статистического машинного перевода с теоретической и практической точки зрения

Same requirement than phrase-based systems

Sources of parallel corpus

- European Plenary Parlament Speeches (EPPS) transcriptions
- Canadian Handsards
- United Nations
- CommonCrawl
- ...



International evaluation campaigns:
Conference on Machine Translation (WMT)
International Workshop on Spoken Language Translation (IWSLT)

What else do we need?

Same requirements than phrase-based systems

Automatic measure

SYSTEM A: Israeli officials (2-GRAM MATCH) responsibility of airport (1-GRAM MATCH) safety

REFERENCE: Israeli officials are responsible for airport security

SYSTEM B: airport security (2-GRAM MATCH) Israeli officials are responsible (4-GRAM MATCH)

Metric	System A	System B
precision (1gram)	3/6	6/6
precision (2gram)	1/5	4/5
precision (3gram)	0/4	2/4
precision (4gram)	0/3	1/3
brevity penalty	6/7	6/7
BLEU	0%	52%

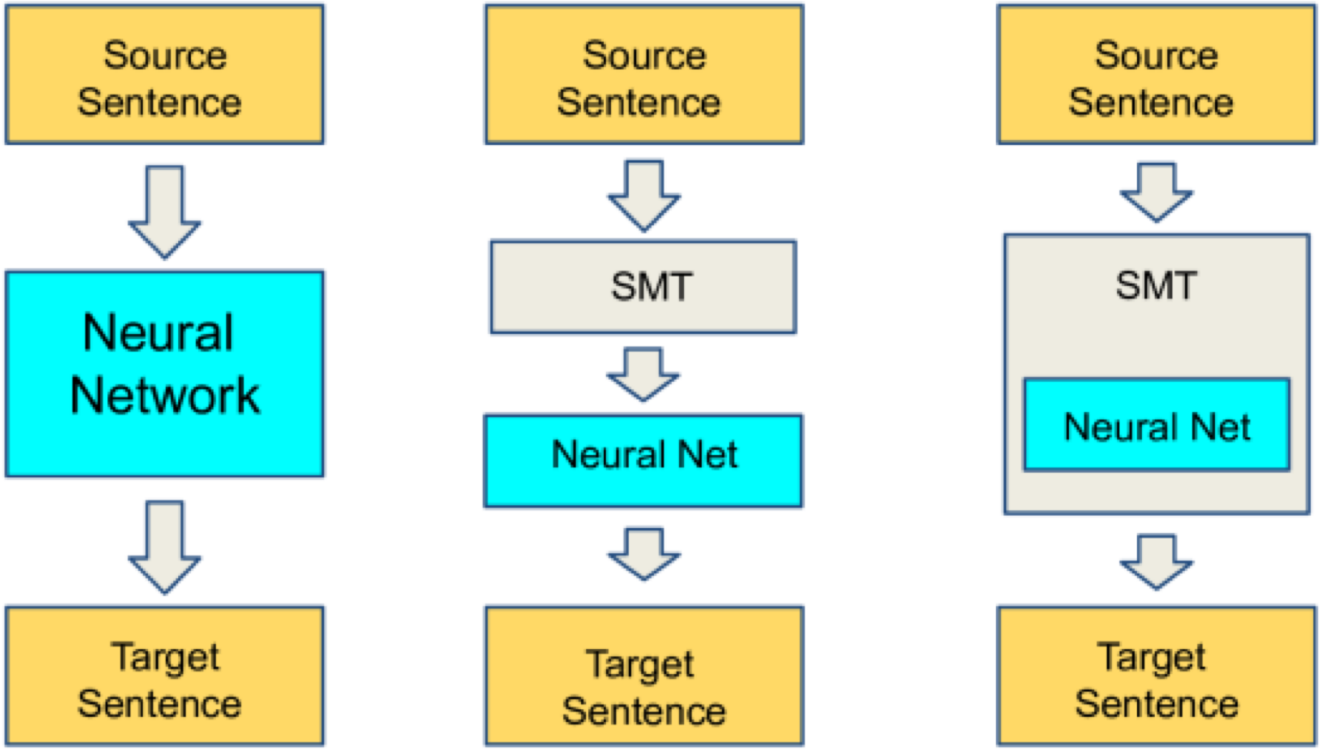
Neural nets are…

- Neural networks, a branch of machine learning, are a biologically-inspired programming paradigm which enables a computer to learn from observational data (http://neuralnetworksanddeeplearning.com/)

Deep learning is…

- A branch of machine learning based on a set of algorithms that attempt to model high-level abstractions in data by using model architectures, with complex structures or otherwise, composed of multiple non-linear transformations (wikipedia)
- A set of machine learning algorithms which attempt to learn multiple-layered models of inputs, commonly neural networks (Du et al, 2013)

Towards Neural Machine Translation



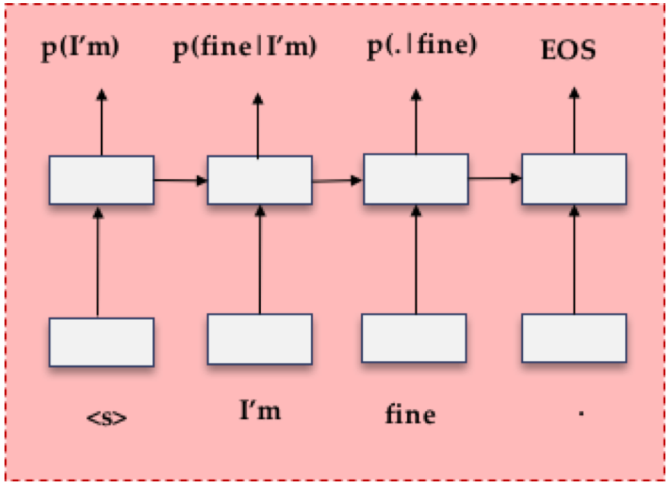
Kyunghyun Cho, "Introduction to Neural Machine Translation with GPUs" (2015)

Outline

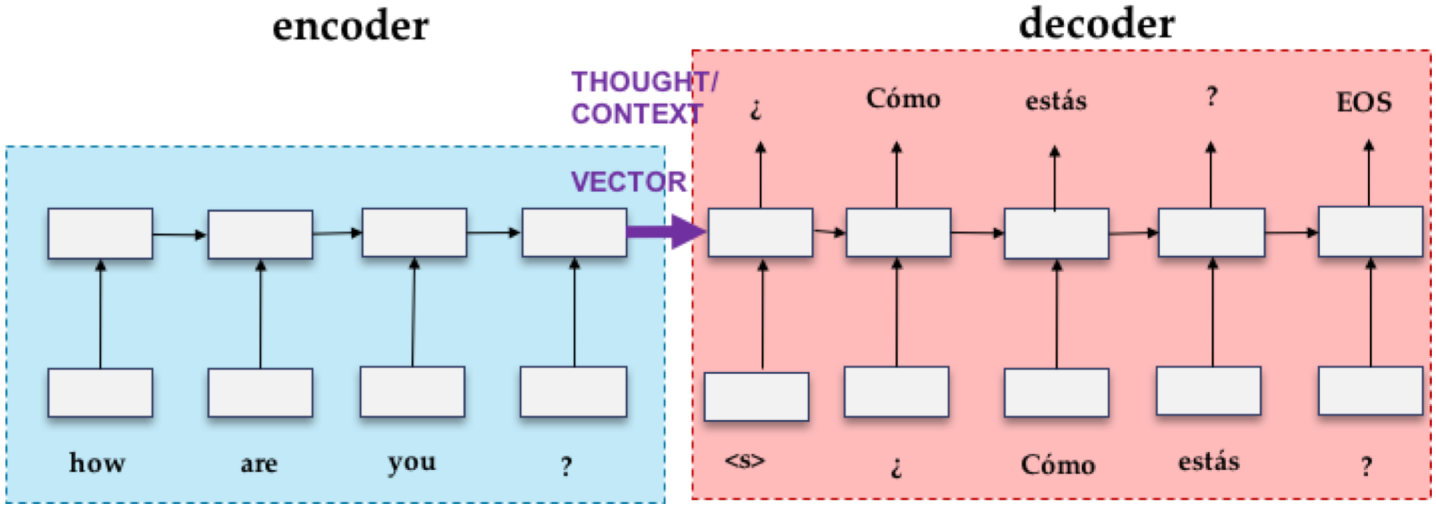
- Machine translation background
- Neural Machine Translation (NMT): Basics
  - Motivation, Architecture
- Research advances in NMT
  - Characters, multimodality, multilinguality
- Conclusions

Sequence modeling

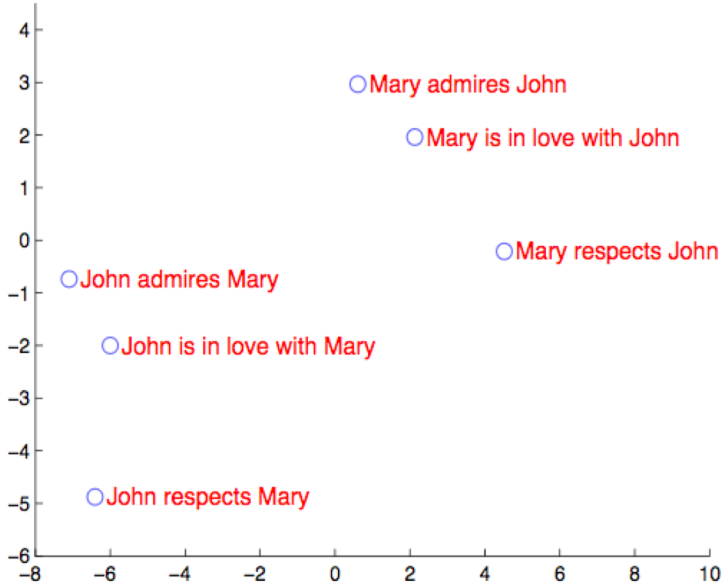
Model the probability of sequences of words

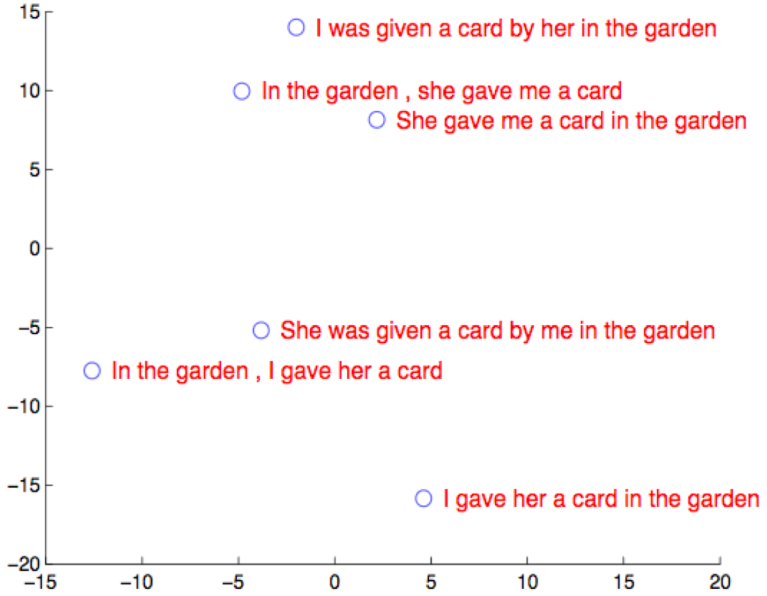


Sequence-to-sequence

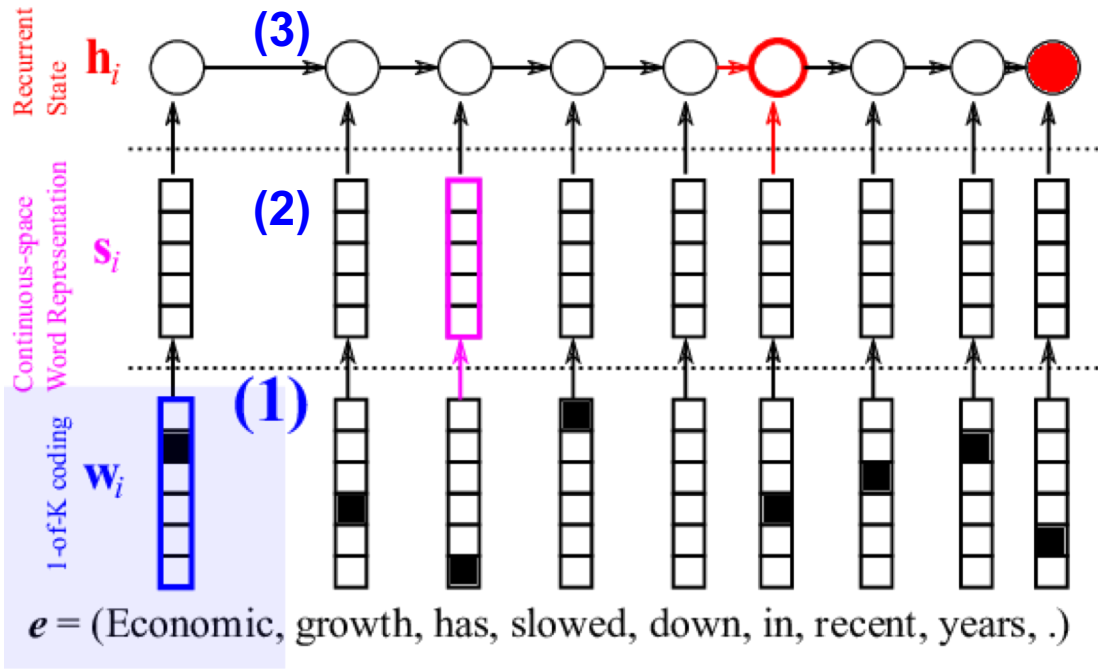


Motivation


Mary admires John
Mary is in love with John
Mary respects John
John admires Mary
John is in love with Mary
John respects Mary


I was given a card by her in the garden
In the garden , she gave me a card
She gave me a card in the garden
She was given a card by me in the garden
In the garden , I gave her a card
I gave her a card in the garden

Encoder



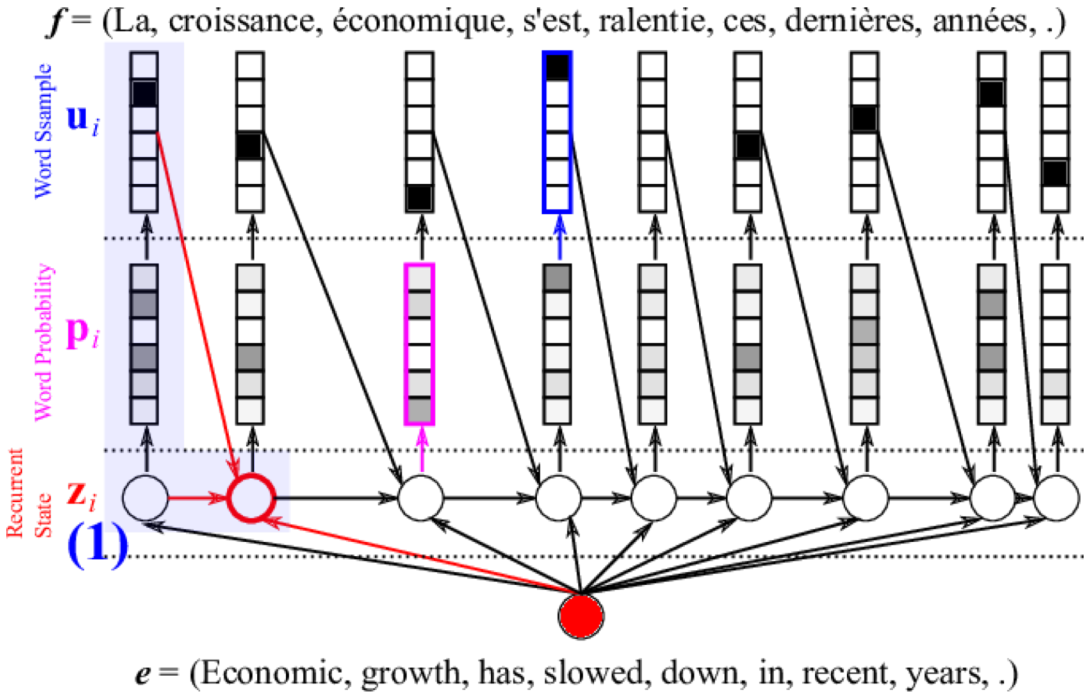
(1) One hot encoding
(2) Continuous space representation
(3) Sequence summarization

Decoder

$$e(k) = w_k^\top z_i + b_k,$$

Neuron weights for word k

RNN internal state



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

23

Why this may not work?

Why this may not work?

We are encoding the entire source sentence into a single context vector

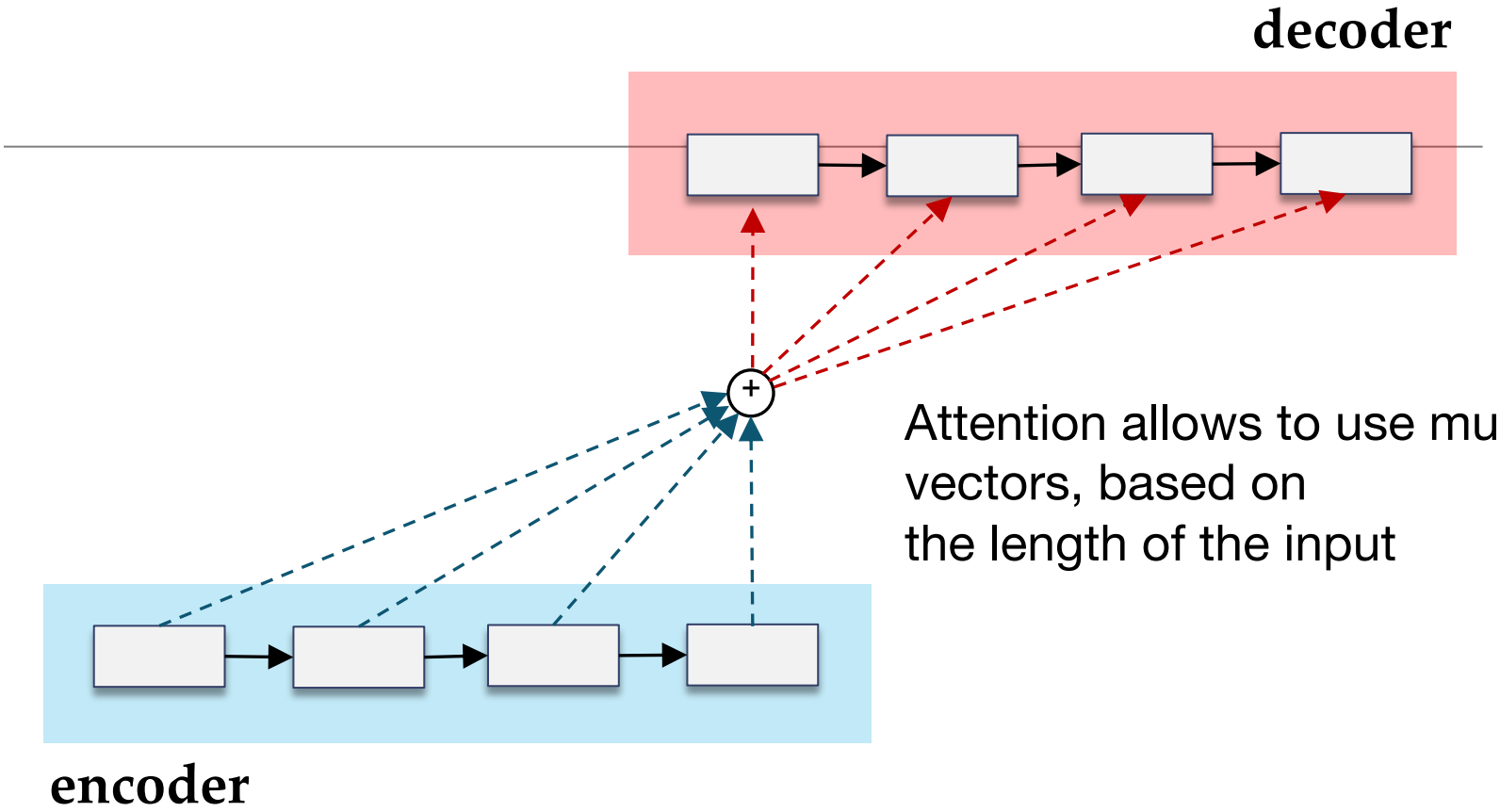
How to solve this?

With the attention-based mechanism…

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

Attention

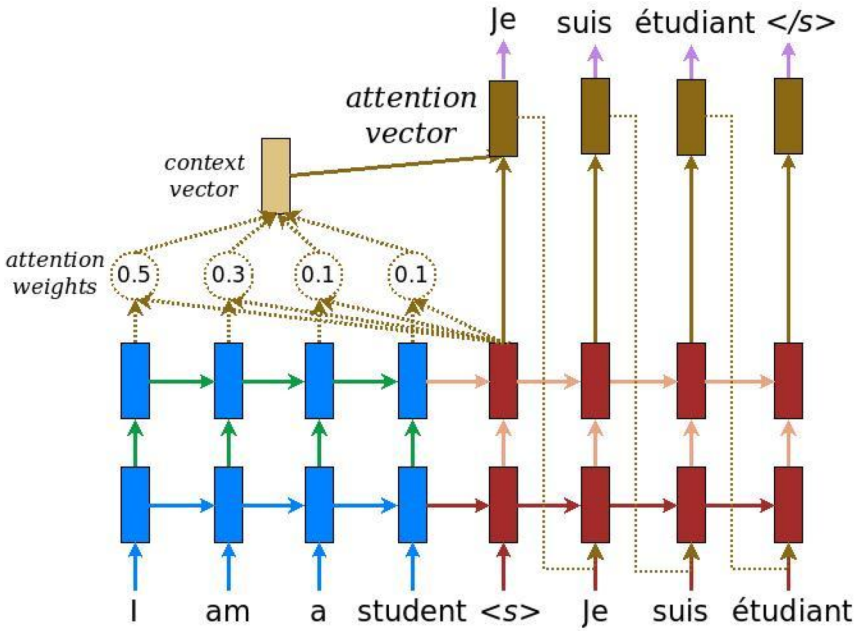


Attention allows to use multiple vectors, based on the length of the input

Attention Key Ideas

- Encode each word in the input and output sentence into a vector
- When decoding, perform a linear combination of these vectors, weighted by “attention weights”
- Use this combination in picking the next word

Attention Integration



Improvements in Attention

Coverage

Problem: Neural models tends to drop or repeat content

In MT,

1.Over-translation: some words are unnecessarily translated for multiple times;

2. Under-translation: some words are mistakenly untranslated.

SRC: Señor Presidente, abre la sesión.

TRG: Mr President Mr President Mr President.

Solution: Model how many times words have been covered e.g. maintaining a coverage vector to keep track of the attention history (Tu et al., 2016)

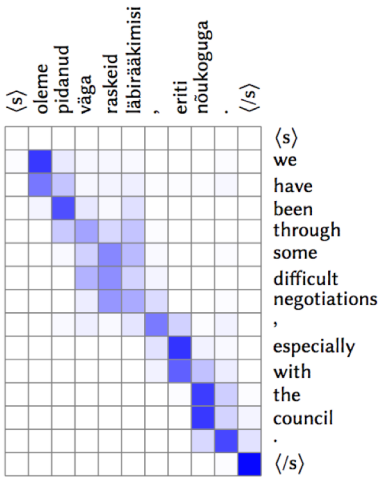
Modeling Coverage for Neural Machine Translation

Zhaopeng Tu[†] Zhengdong Lu[†] Yang Liu[‡] Xiaohua Liu[†] Hang Li[†]

Markov Properties

Intuition: Attention from last time tends to be correlated with attention this time

Approach: Add information about the last attention when making the next decision



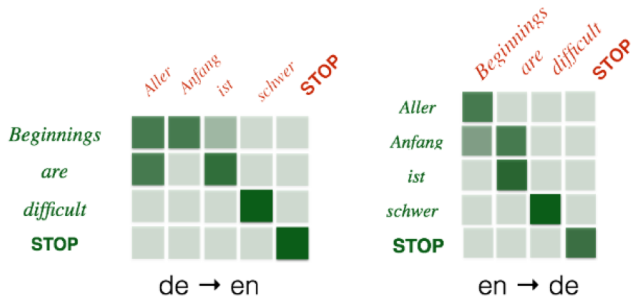
Incorporating Structural Alignment Biases into an Attentional Neural Translation Model

Trevor Cohn and **Cong Duy Vu Hoang** and **Ekaterina Vymolova**

Bidirectional training

-Background: Established that for latent variable translation models the alignments improve if both directional models are combined (koehn et al, 2005)

-Approach: joint training of two directional models



Incorporating Structural Alignment Biases into an Attentional Neural Translation Model

Trevor Cohn and **Cong Duy Vu Hoang** and **Ekaterina Vymolova**

Supervised Training

Sometimes we can get “gold standard” alignments a –priori

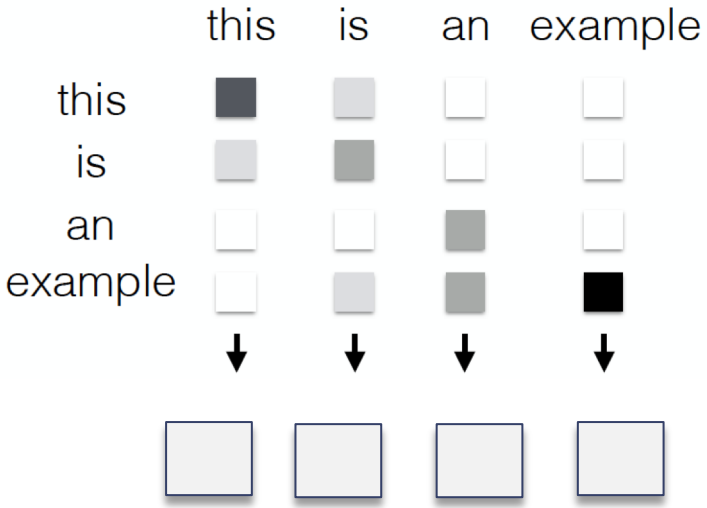
- Manual alignments
- Pre-trained with strong alignment model

Train the model to match these strong alignments

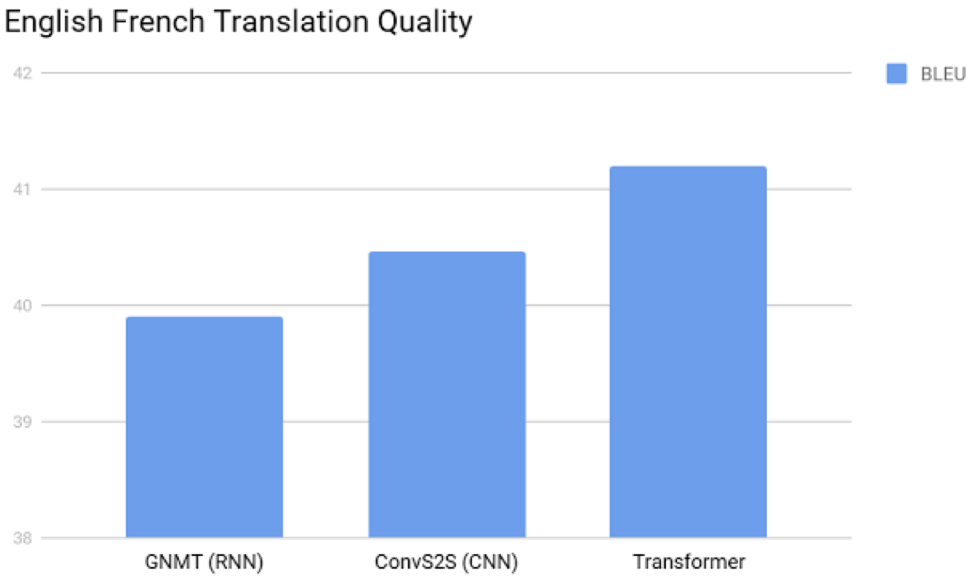
Attention is all you need

Intra-Attention

Each element in the sentence attends to other elements from the SAME sentence → context sensitive encodings!



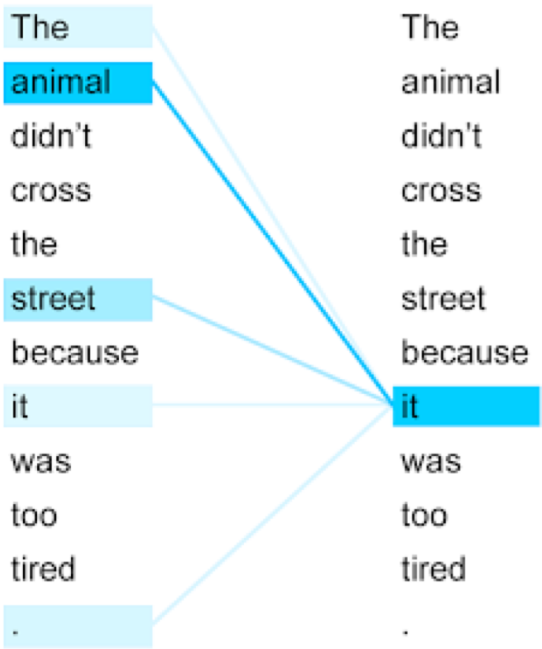
Results

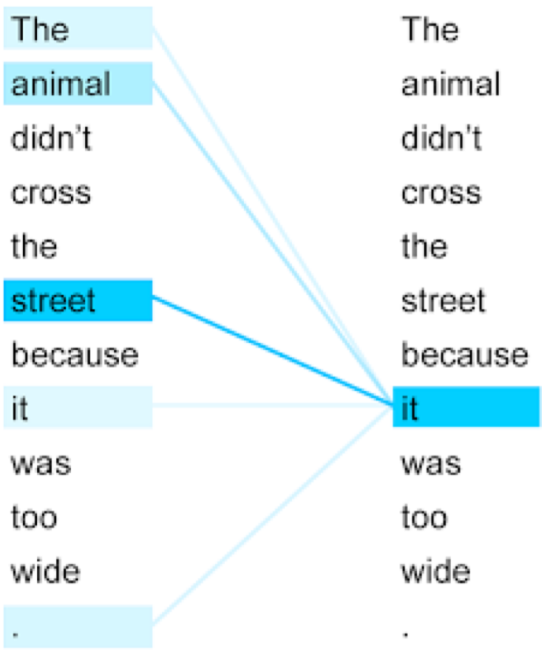

English French Translation Quality
BLEU
42
41
40
39
38
GNMT (RNN)
ConvS2S (CNN)
Transformer

Attention Weights

The animal didn't cross the street because it was too tired.
L'animal n'a pas traversé la rue parce qu'il était trop fatigué.

The animal didn't cross the street because it was too wide.
L'animal n'a pas traversé la rue parce qu'elle était trop large.

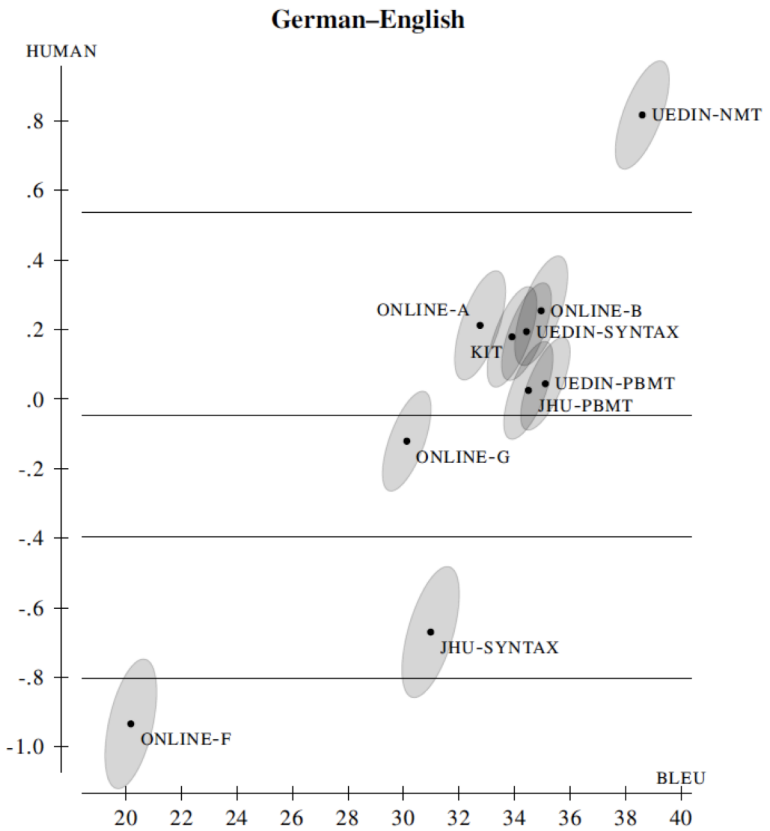

The
animal
didn't
cross
the
street
because
it
was
too
tired
.
The
animal
didn't
cross
the
street
because
it
was
too
tired
.

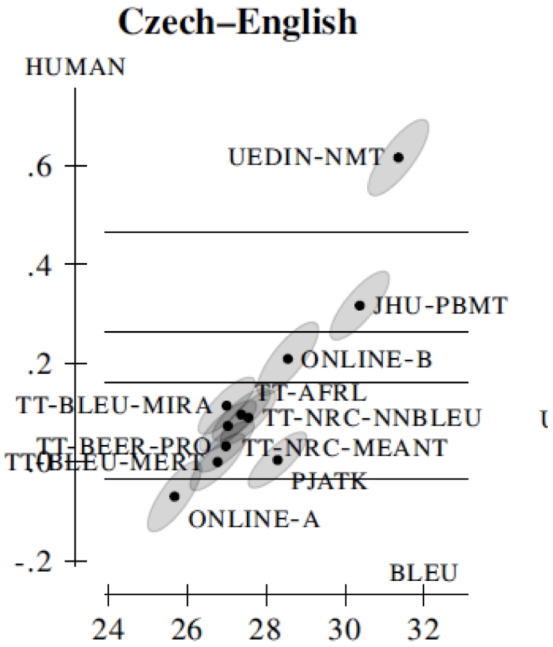

The
animal
didn't
cross
the
street
because
it
was
too
wide
.
The
animal
didn't
cross
the
street
because
it
was
too
wide
.

Neural MT is better than phrase-based

Neural Network for Machine Translation at Production Scale

Results in WMT 2016 international evaluation





Case Study: Catalan-Spanish

System	CAES						ESCA					
	In-domain			Out-of-domain			In-domain			Out-of-domain		
	METEOR	BLEU	TT	METEOR	BLEU	TT	METEOR	BLEU	TT	METEOR	BLEU	TT
Rule	87.11	75.20	0,35'	67.22	50.53	0,19'	83.92	70.15	0,35'	63.77	50.41	0,19'
Phrase	**90.59**	81.80	50,59'	**72.31**	**57.20**	3,01'	90.88	84.24	51,56'	60.73	47.63	3,10'
Neural	90.26	**83.72**	25,98'	67.21	53.69	3,38'	**92.82**	**87.44**	28,10'	**67.89**	**53.68**	2,76'
MBR	92.76	83.73	-	74.04	59.61	-	92.99	57.55	-	74.07	54.46	-

Case Study: Catalan-Spanish

1	SRC	una cosa lúdica y divertida
	Rule	una cosa **lúdica** i **divertida**
	Phrase	una cosa lúdic i divertit
	Neural	una cosa **lúdica** i **divertida**
	REF	una cosa lúdica i divertida
2	SRC	noche de impacto Antena 3 TV , 21.45 Antena 3 TV apuesta de nuevo por las imgenes
	Rule	nit d ' impacte Antena 3 TV , 21.45 Antena 3 TV aposta de nou per les imatges
	Phrase	nit d ' impacte ' Antena 3 TV , 21.45 Antena 3 TV aposta per les imatges
	Neural	**noche de impacto** Antena 3 TV , 21.45 Antena 3 TV torna a apostar per les imatges
	REF	noche de impacto Antena 3 TV , 21.45 Antena 3 TV aposta un altre cop per les imatges
3	SRC	este rgano de gestin decidira qu unidades del aeropuerto podran privatizarse .
	Rule	aquest rgan de gesti decidiria qu unitats de l ' aeroport podrien **privatitzar-se** .
	Phrase	aquest rgan de gesti decidiria quin unitats de l ' aeroport podrien privatitzar .
	Neural	aquest rgan de gesti decidiria **quines unitats** de l aeroport podrien privatitzar .
	REF	aquest rgan de gesti decidiria quines unitats de l aeroport es podrien privatitzar .
4	SRC	quan el pilot ja era a l ' entrada del parc (...)
	Rule	cuando el piloto ya era a la entrada del parque (...)
	Phrase	cuando el piloto era ya a la entrada del parque (...)
	Neural	cuando el piloto **estaba** ya a la entrada del parque (...)
	REF	cuando el pelotón ya se hallaba en la entrada del parque (...)
5	SRC	i s hi han d afegir dues querelles de particulars .
	Rule	y **han tenido que** añadir dos querellas de particulares .
	Phrase	y se han de añadir dos querellas de particulares .
	Neural	y **hay que** añadir dos querallas de particulares
	REF	a estos hay que añadir dos querellas de particulares .

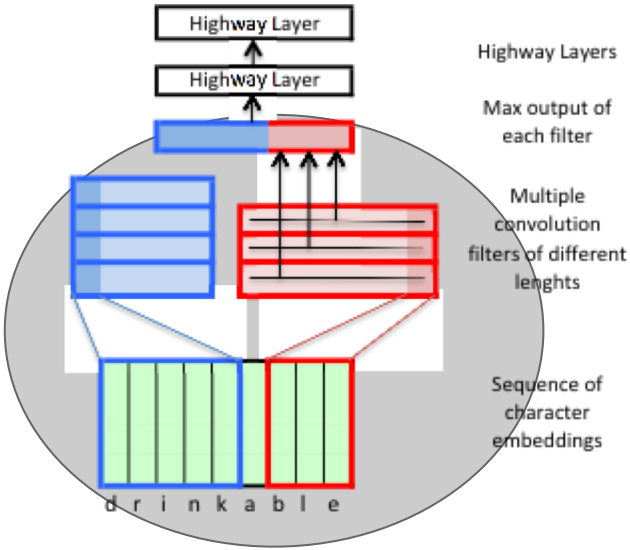
Outline

- Machine translation background
- Neural Machine Translation (NMT): Basics
  - Motivation, Architecture
- Research advances in NMT
  - Characters, multimodality, multilinguality
- Conclusions

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.

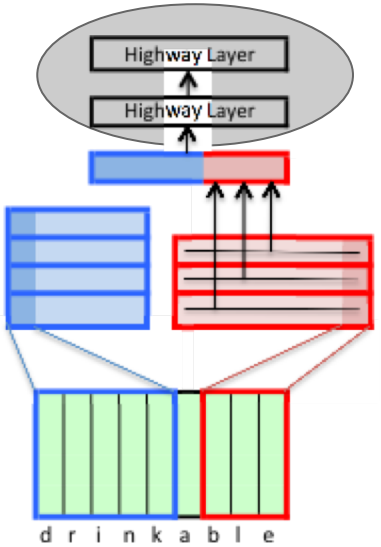
Character-based Neural MT: Proposal



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.

Character-based Neural MT: Integration with NMT



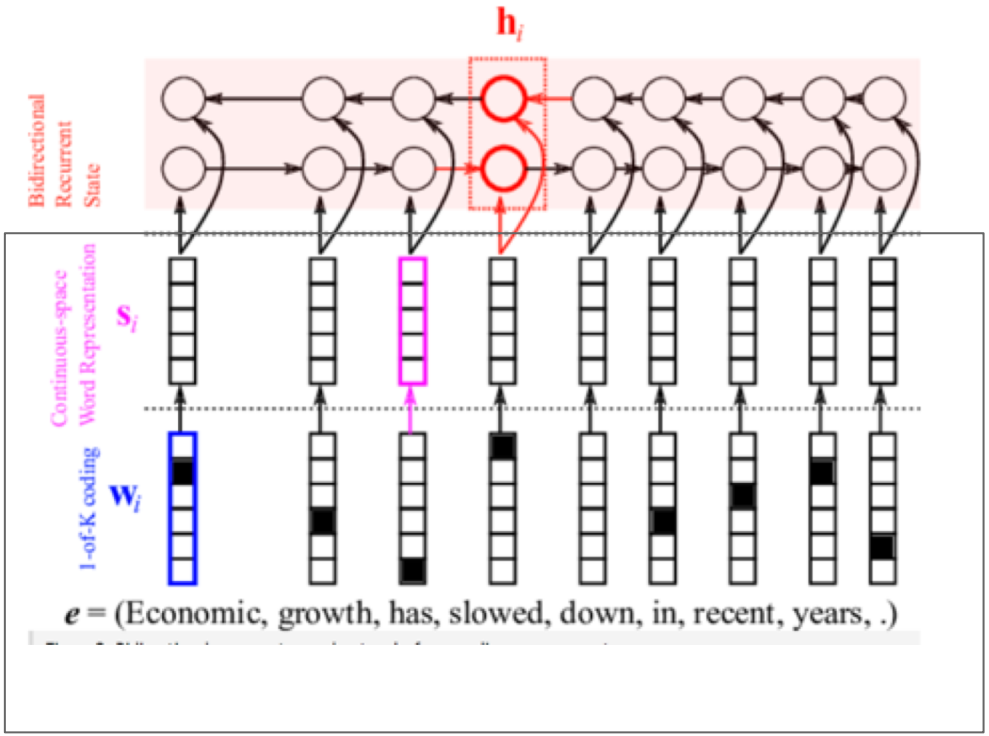
Highway Layers

Max output of each filter

Multiple convolution filters of different lenghts

Sequence of character embeddings

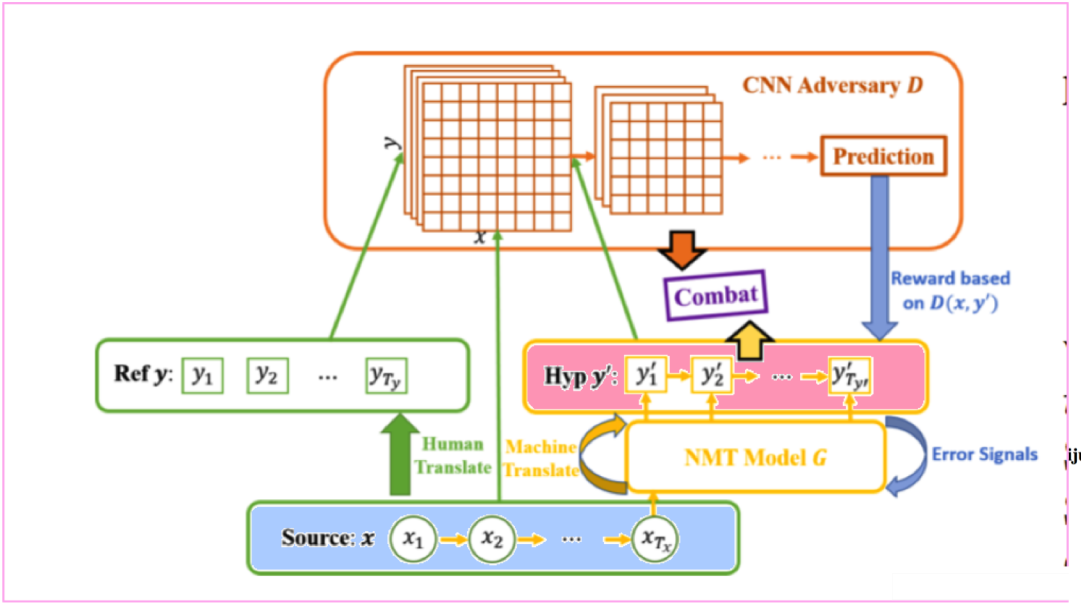
Kim et al, 2015



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 .
2	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 .

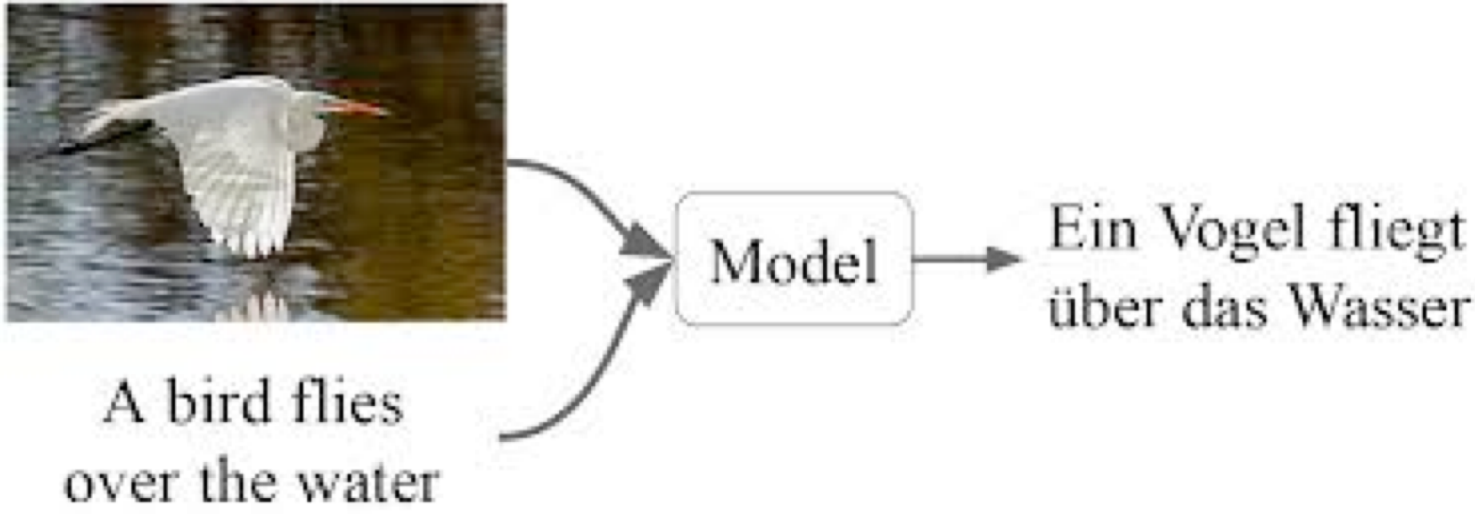
Generative Adversarial Networks



Adversarial Neural Machine Translation

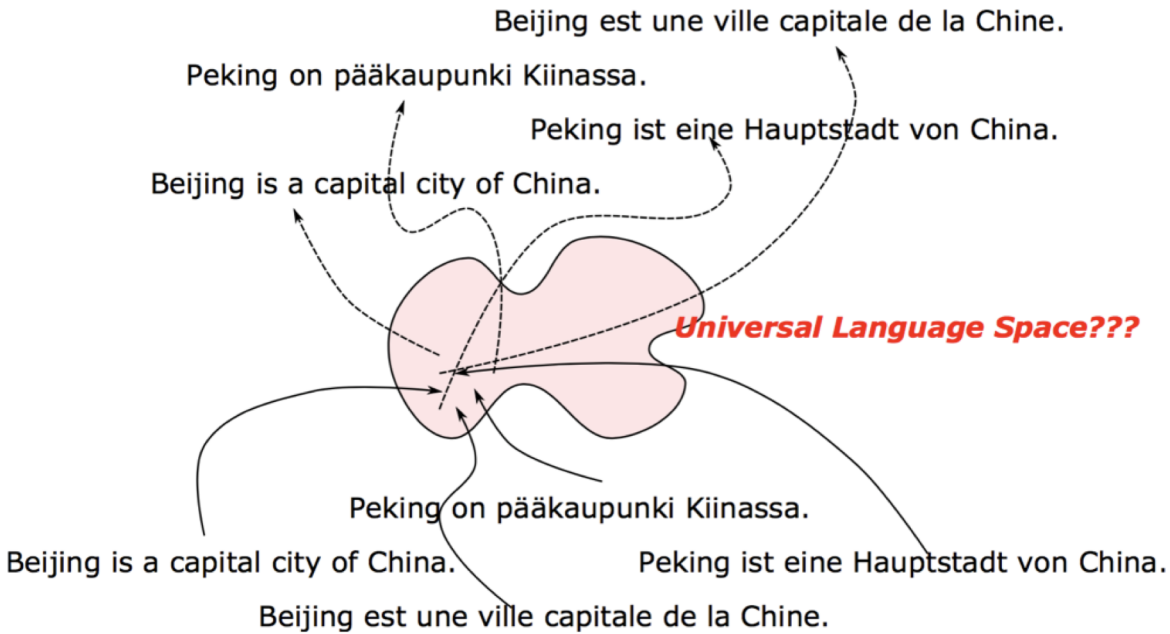
ijun Wu[1], Yingce Xia[2], Li Zhao[3], Fei Tian[3], Tao Qin[3], Jianhuang Lai[1,4] and Tie-Yan Liu[3]
[1]School of Data and Computer Science, Sun Yat-sen University
[2]University of Science and Technology of China
[3]Microsoft Research Asia
[4]Guangdong Key Laboratory of Information Security Technology
wulijun3@mail2.sysu.edu.cn; yingce.xia@gmail.com;
{lizo,fetia,taoqin,tie-yan.liu}@microsoft.com; stsljh@mail.sysu.edu.cn

Multimodal Translation



http://www.statmt.org/wmt18/multimodal-task.html

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)

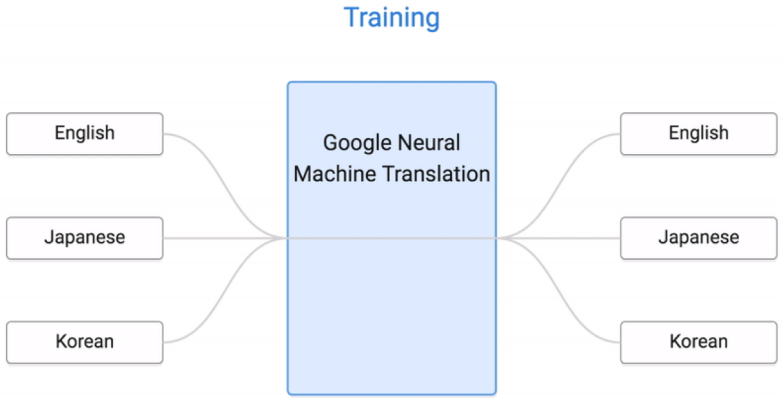
Multilingual Translation Approaches

Sharing attention-based mechanism across language pairs

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

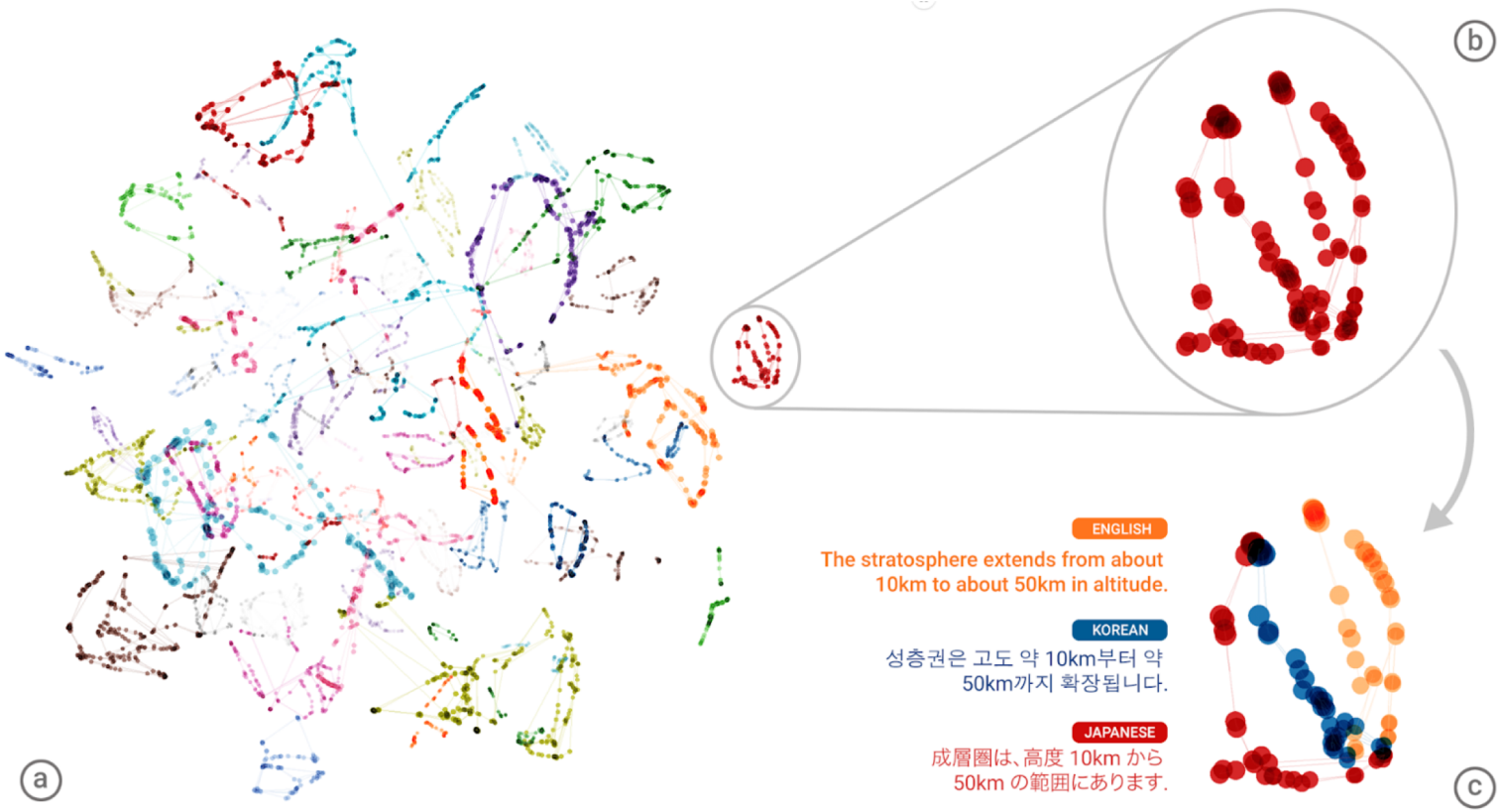
Share encoder, decoder, attention accross language pairs

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

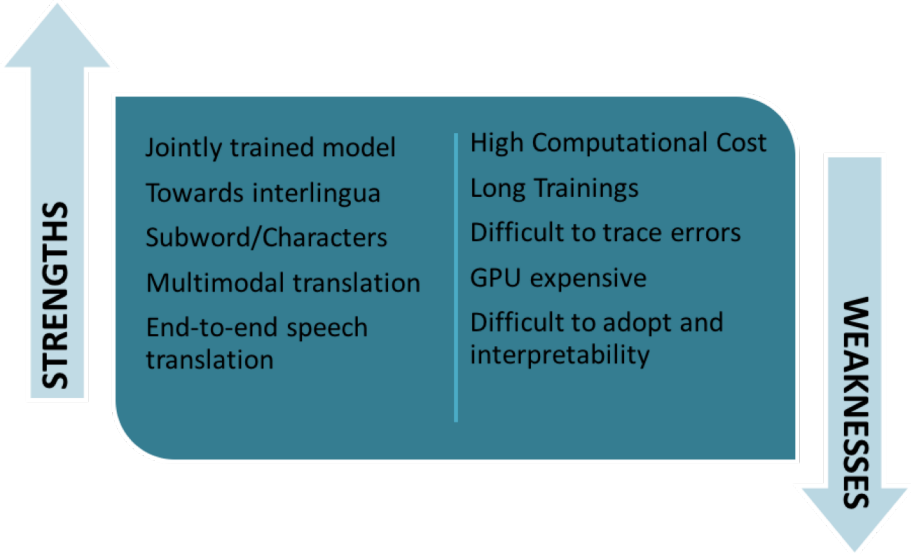
Outline

- Machine translation background
- Neural Machine Translation (NMT): Basics
  - Motivation, Architecture
- Research advances in NMT
  - Characters, multimodality, multilinguality
- Conclusions

Conclusions

- Machine Translation is faced as a **sequence-to-sequence** problem
- The source sentence is **encoded** into a fixed length vector and this fixed length vector is **decoded** into the final most probable target sentence
- Only **parallel corpus** and **automatic evaluation** measures are required to train a neural machine translation system
- Attention-based mechanism allows to achieve state-of-the-art results
- Progress in MT includes character-based, multilinguality...

Strengths/Weaknesses



Thanks ! Q&A ?

https://www.costa-jussa.com
marta.ruiz@upc.edu

Agència de Gestió d'Ajuts Universitaris i de Recerca
AGAUR
chist-era
GOBIERNO DE ESPAÑA
MINISTERIO DE ECONOMÍA, INDUSTRIA Y COMPETITIVIDAD