

Neural Machine Translation

Kyunghyun Cho

New York University

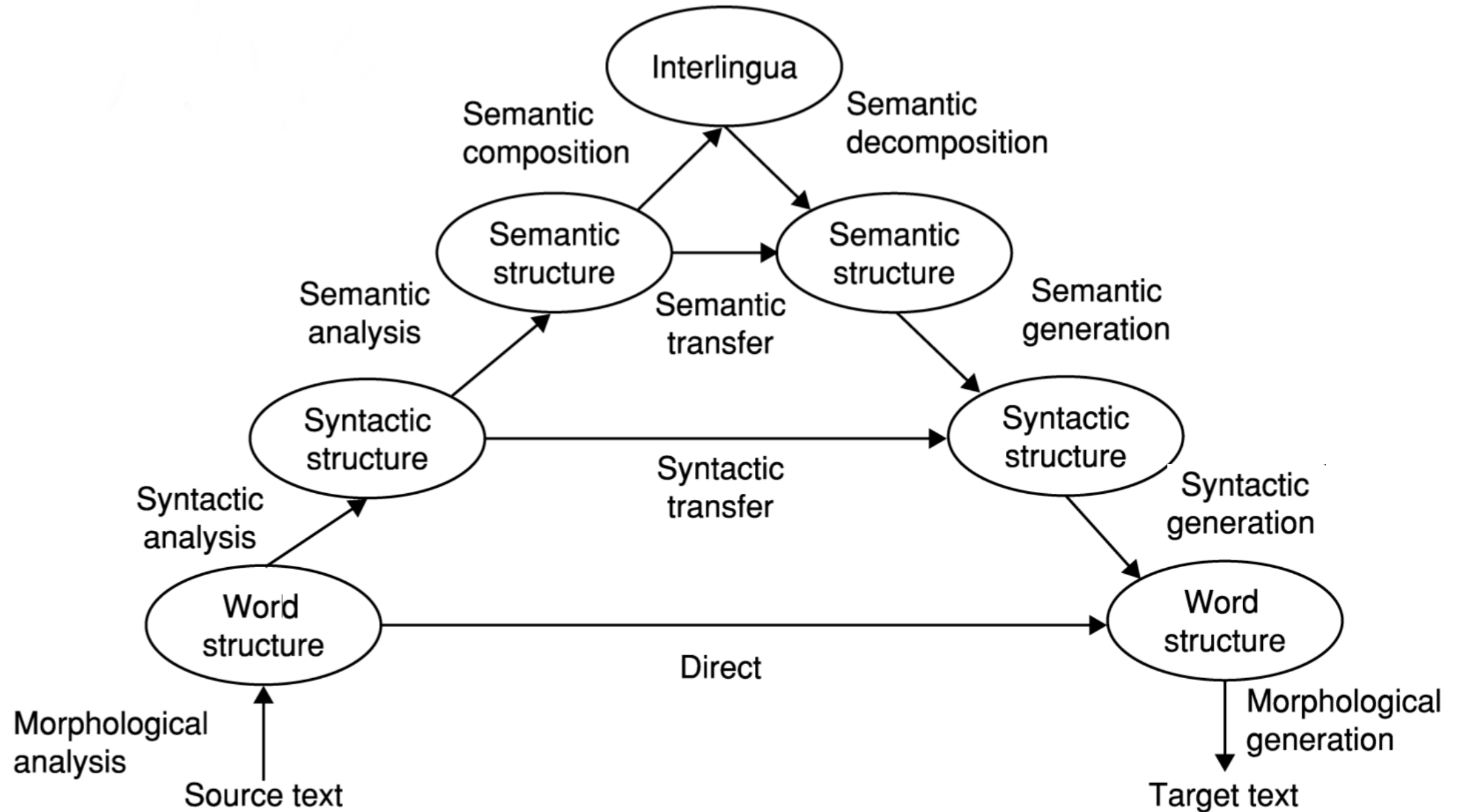
Courant Institute (Computer Science) and Center for Data Science

Facebook AI Research

A bit of historical remark

When did neural machine translation start, and where does it fit?

Only

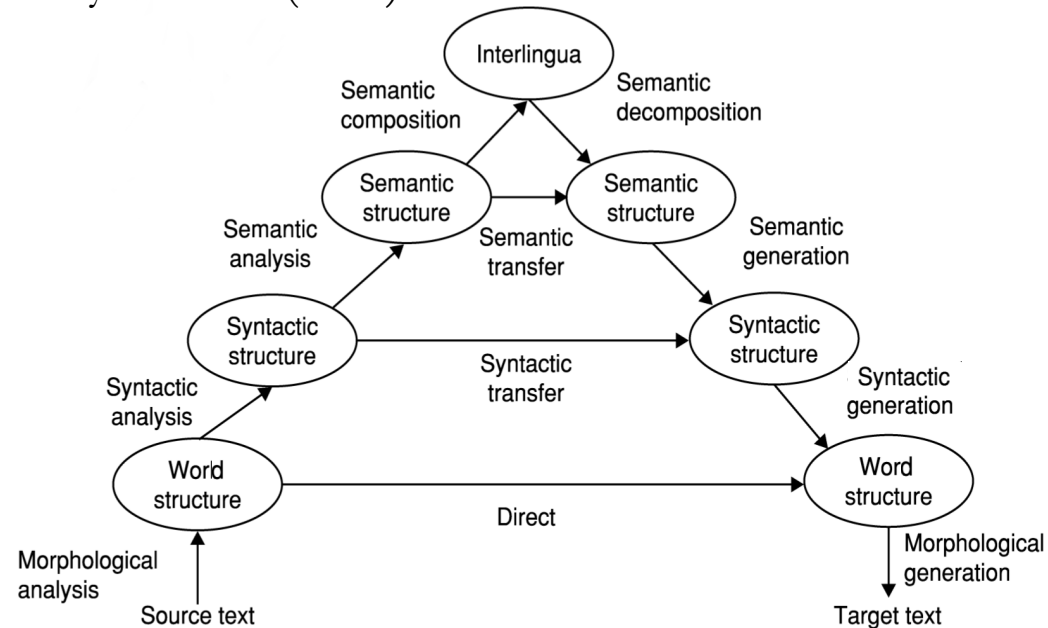


Unfortunately not...

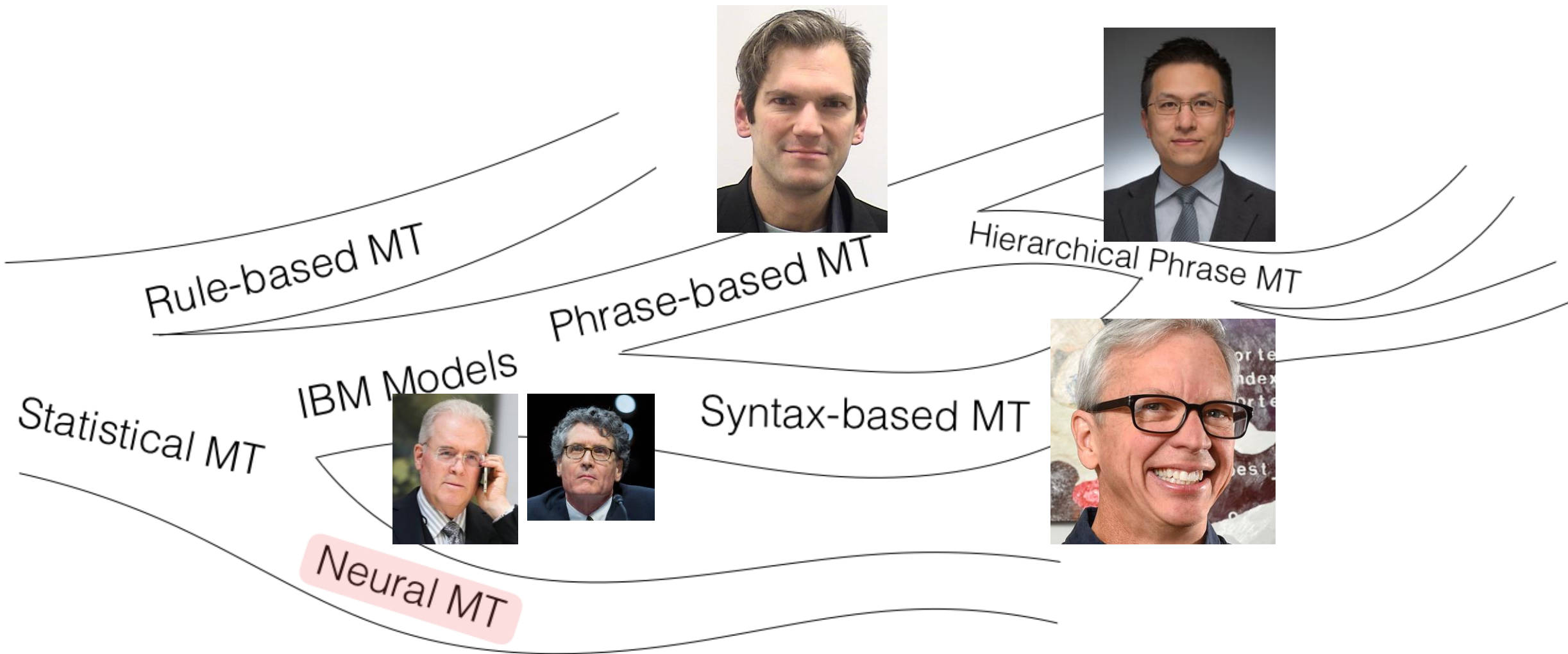
*“Every time I fire a linguist,
the performance of the
recognizer goes up.”*

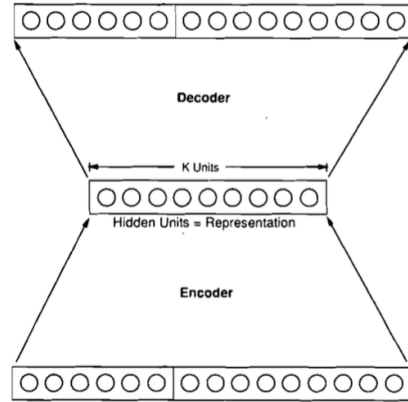
- Fred Jelinek (IBM), 1988

Borr, Hovy & Levin (2006)

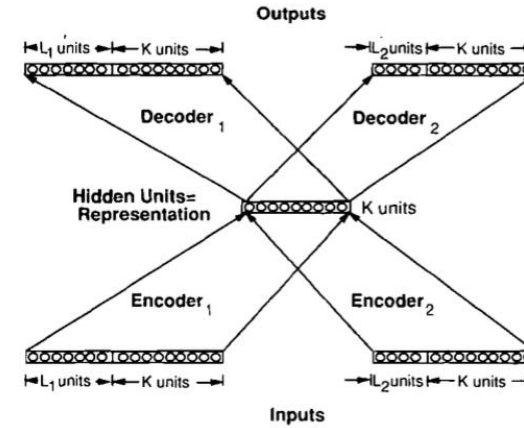


Machine Translation

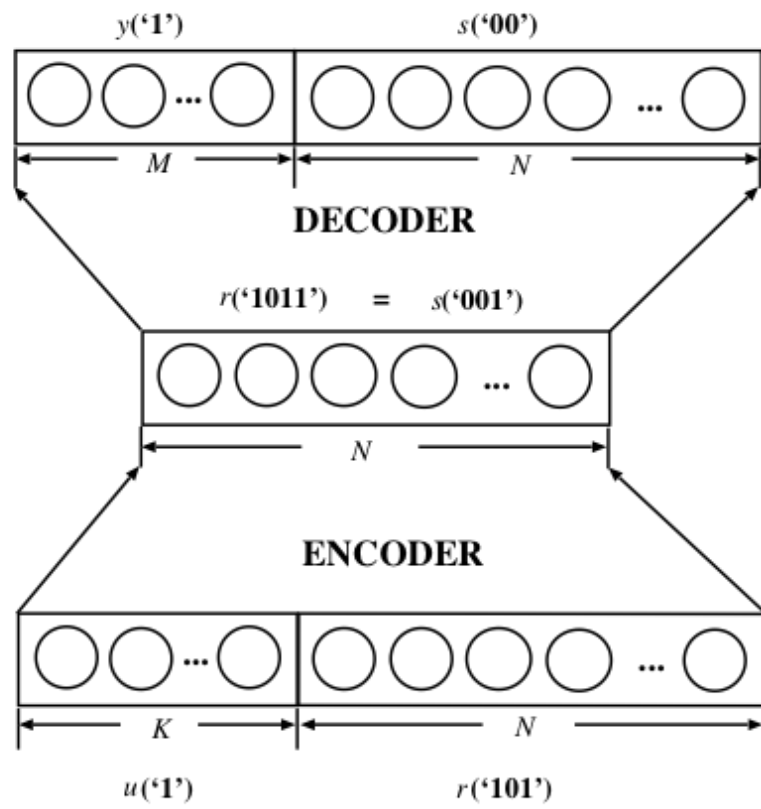




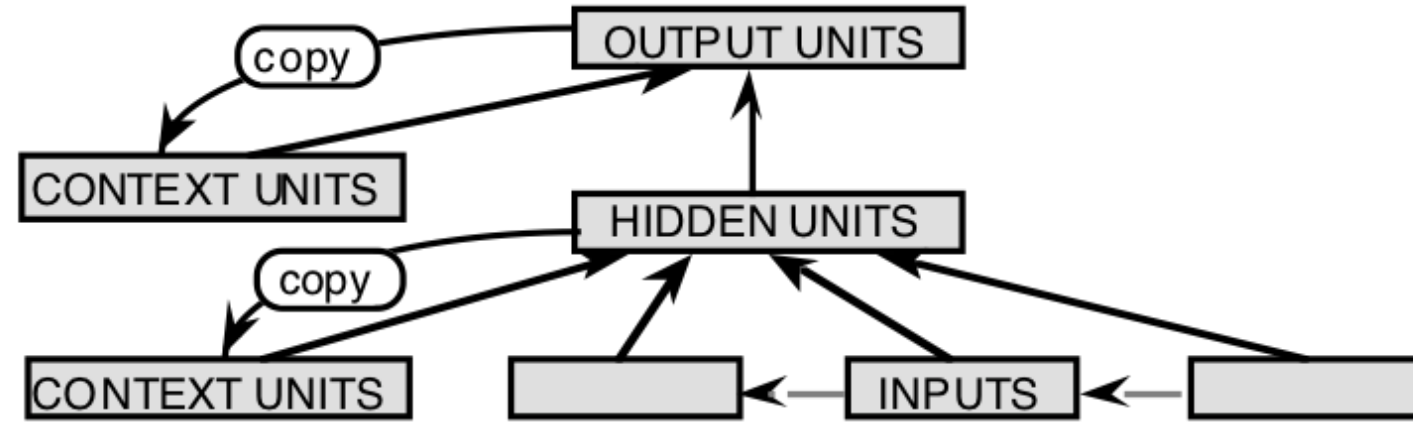
- [Allen 1987 IEEE 1st ICNN]
- 3310 En-Es pairs constructed on 31 En, 40 Es words, max 10/11 word sentence; 33 used as test set
- Binary encoding of words – 50 inputs, 66 outputs; 1 or 3 hidden 150-unit layers. Ave WER: 1.3 words



- [Chrisman 1992 *Connection Science*]
- Dual-ported RAAM architecture [Pollack 1990 *Artificial Intelligence*] applied to corpus of 216 parallel pairs of simple En-Es sentences:
- Split 50/50 as train/test, 75% of sentences correctly translated!



Brief resurrection in 1997



"We propose .. **Recursive Hetero-Associative Memory** which .. may be applied **to learn general translations from examples** in which different sentences have the same translation."

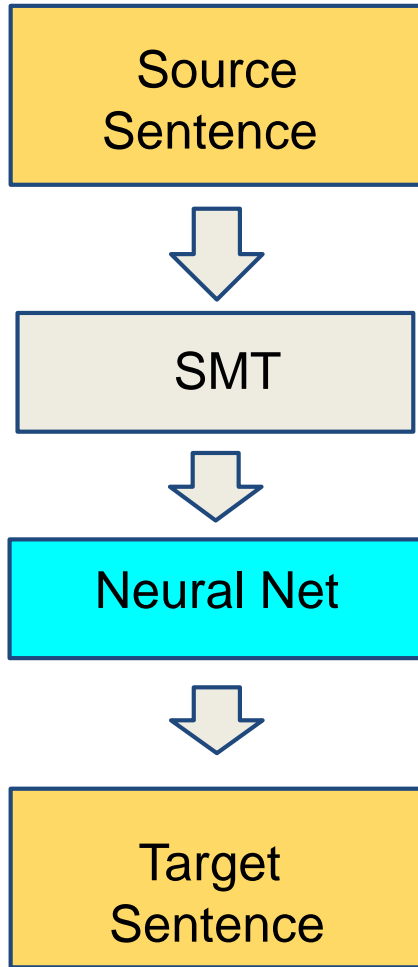


— Forcada & Neco, 1997

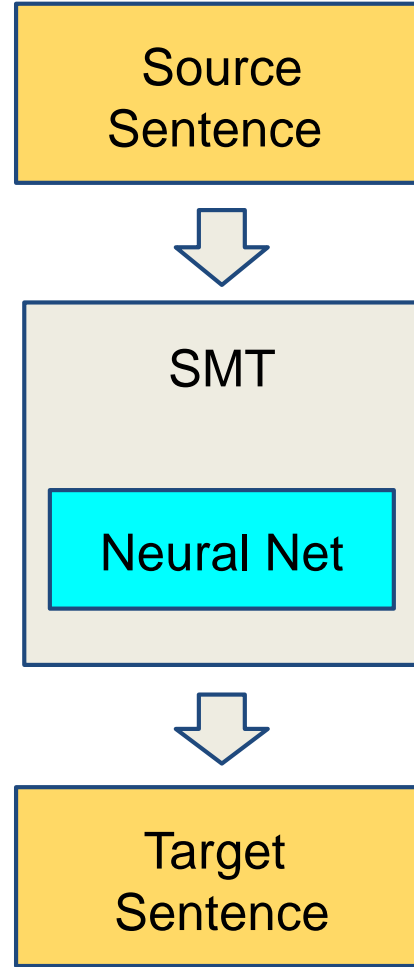
"Based on these encouraging performances, future work dealing with more complex limited-domain translations seems to be feasible. **However, the size of the neural nets required for such applications (and consequently, the learning time) can be prohibitive**"

- Castano & Casacuberta, 1997

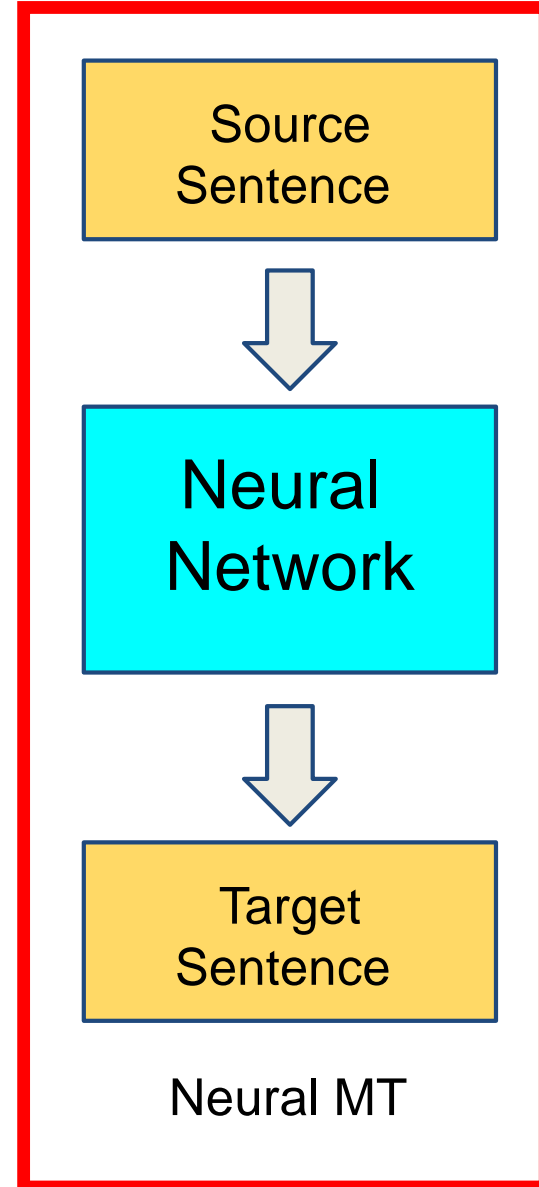
Modern neural machine translation



(Schwenk et al. 2006)



(Devlin et al. 2014)



Neural MT

Machine Translation

- Input: a sentence written in a source language L_S
- Output: a corresponding sentence in a target language L_T
- Problem statement:
 - Supervised learning: given the input sentence, output its translation
 - Compute the conditional distribution over all possible translation given the input
$$p(Y = (y_1, \dots, y_T) | X = (x_1, \dots, x_{T'}))$$
- *We have already learned every necessary ingredient for building a full neural machine translation system.*

Token Representation – One-hot Vectors

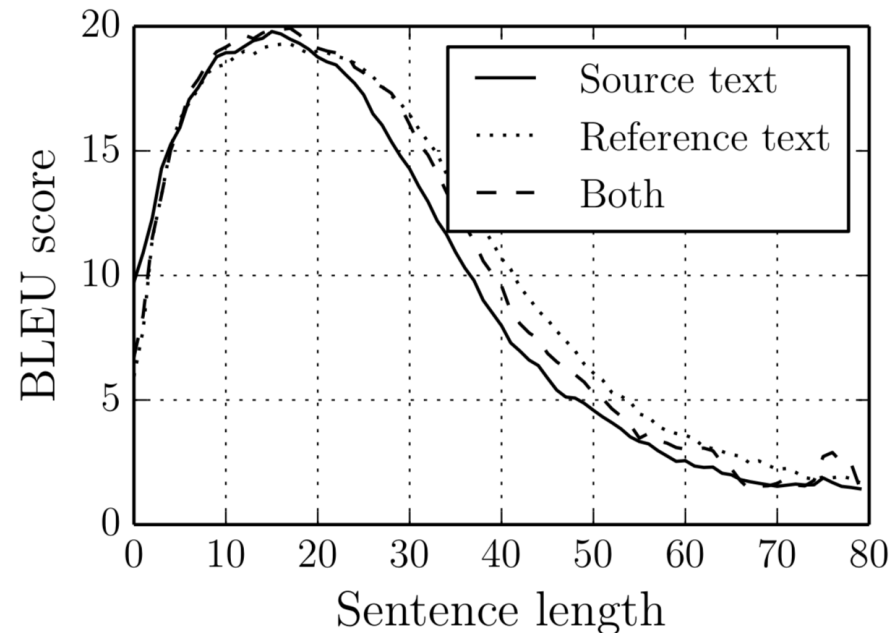
1. Build source and target vocabularies of unique tokens
 - For each of source and target languages,
 1. Tokenize: separate punctuations, normalize punctuations, ...
e.g., “I’m going” => (“I”, “m”, “going”), replace ‘, ’, ` , into “`”, ...
use Spacy.io, NLTK or Moses’ tokenizer.
 2. Subword segmentation: segment each token into a sequence of subwords
e.g., “going” => (“go”, “ing”), use BPE [Sennrich et al., 2015]
 3. Collect all unique subwords, sort them by their frequencies (descending) and assign indices.
2. Transform each subword token into a corresponding one-hot vector.*
 - See Lecture 2.

Encoder – Source Sentence Representation

- Encode the source sentence into a set of sentence representation vectors
 - # of encoded vectors is proportional to the source sentence length: often same.
 $H = (h_1, \dots, h_{T'})$
 - Recurrent networks have been widely used [Cho et al., 2014; Sutskever et al., 2014], but CNN [Gehring et al., 2017; Kalchbrenner&Blunsom, 2013] and self-attention [Vaswani et al., 2017] are used increasingly more often. See Lecture 2 for details.
- We do not want to collapse them into a single vector.
 - Collapsing often corresponds to information loss.
 - Increasingly more difficult to encode the entire source sentence into a single vector, as the sentence length increases [Cho et al., 2014b].
 - We didn't know initially until [Bahdanau et al., 2015].

Encoder – Source Sentence Representation

- Encode the source sentence into a set of sentence representation vectors
- We do not want to collapse them into a single vector.
 - Increasingly more difficult to encode the entire source sentence into a single vector, as the sentence length increases [Cho et al., 2014b].



Encoder – Source Sentence Representation

- Encode the source sentence into a set of sentence representation vectors
- We do not want to collapse them into a single vector.
 - Increasingly more difficult to encode the entire source sentence into a single vector, as the sentence length increases [Cho et al., 2014b].
 - When collapsed, the system fails to translate a long sentence correctly.
 - **Source:** *An admitting privilege is the right of a doctor to admit a patient to a hospital or a medical centre to carry out a diagnosis or a procedure, based on his status as a health care worker at a hospital.*
 - **When collapsed:** *Un privilège d'admission est le droit d'un médecin de reconnaître un patient à l'hôpital ou un centre médical d'un diagnostic ou de prendre un diagnostic en fonction de son état de santé.*
- The system translates reasonable up to a certain point, but starts drifting away.

Decoder – Language Modelling

- Autoregressive Language modelling with an infinite context $n \rightarrow \infty$
 - Larger context is necessary to generate a coherent sentence.
 - Semantics could be largely provided by the source sentence, but syntactic properties need to be handled by the language model directly.
 - Recurrent networks, self-attention and (dilated) convolutional networks
 - Causal structure must be followed.
 - See Lecture 3.
- **Conditional** Language modelling
 - The context based on which the next token is predicted is **two-fold**

$$p(Y|X) = \prod_{t=1}^T p(y_t | y_{<t}, X)$$

Decoder – Conditional Language Modelling

- Conditional Language modelling

- The context based on which the next token is predicted is **two-fold**.

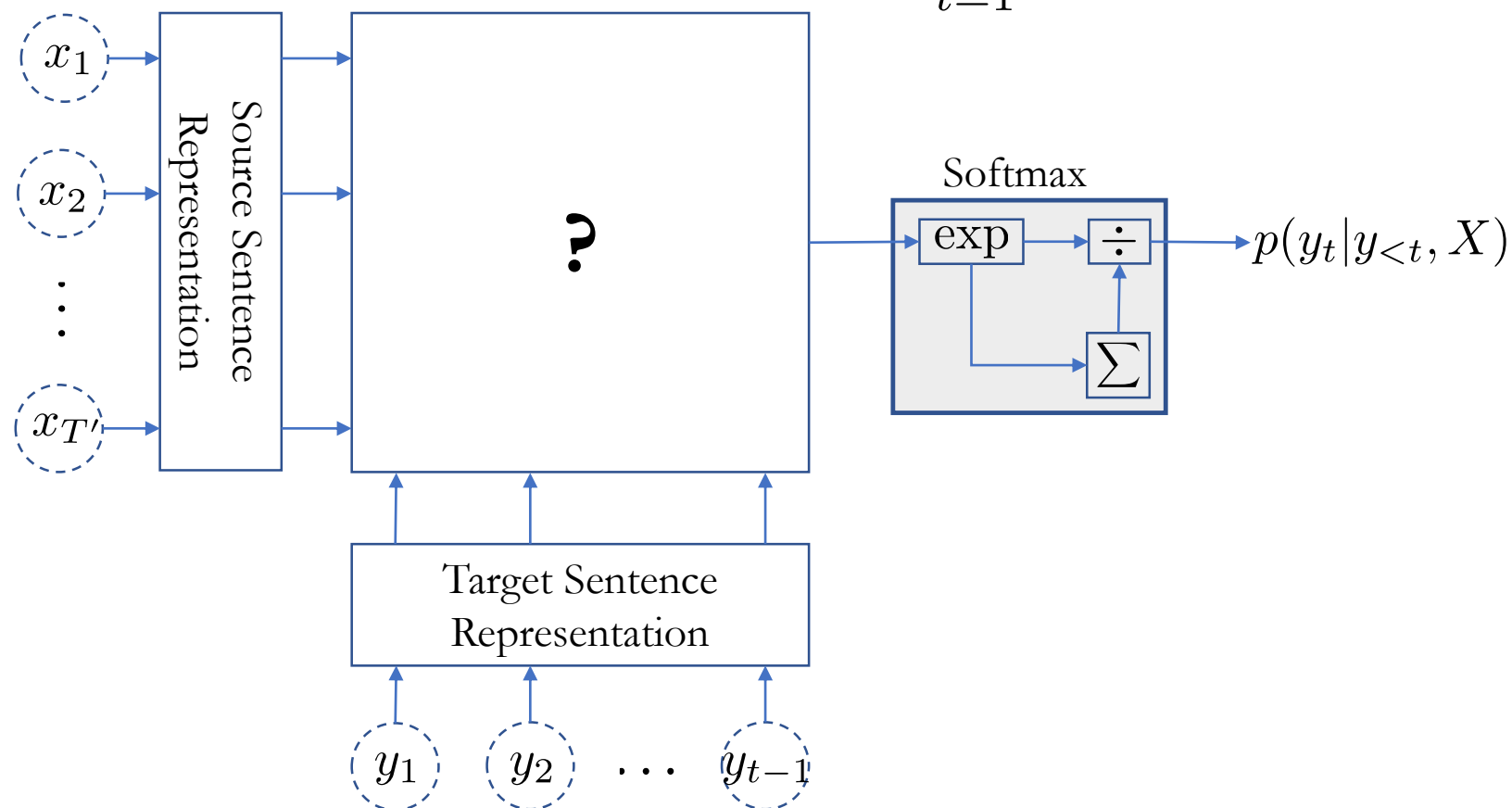
$$p(Y|X) = \prod_{t=1}^T p(y_t | y_{<t}, X)$$

- Supervised learning: T input-output training pairs per sentence

- Input: the entire source sentence X and the preceding target tokens $y_{<t}$
 - Output: the next token y_t

Decoder – Conditional Language Modelling

- Conditional Language modelling $p(Y|X) = \prod_{t=1}^T p(y_t|y_{<t}, X)$

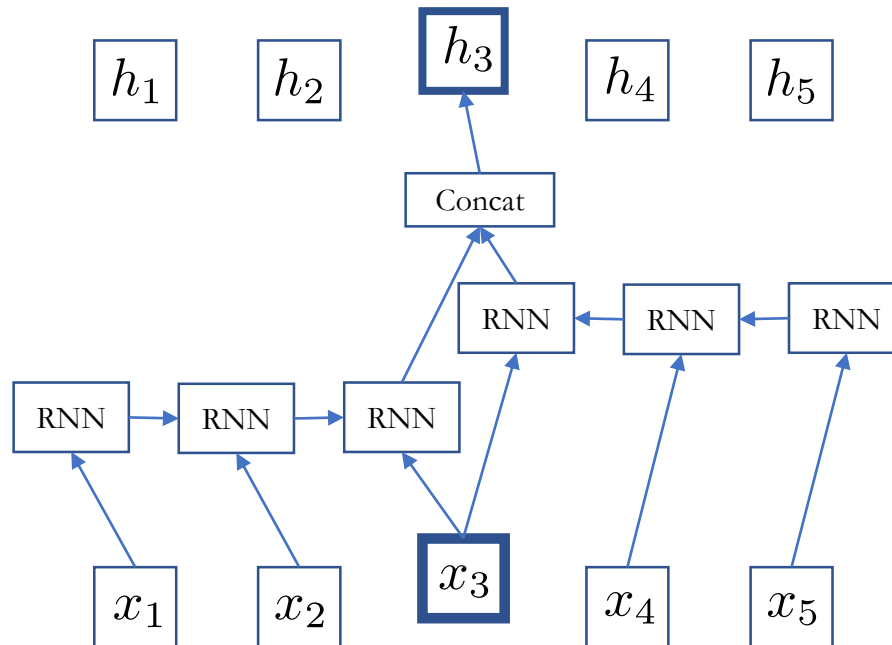


RNN Neural Machine Translation

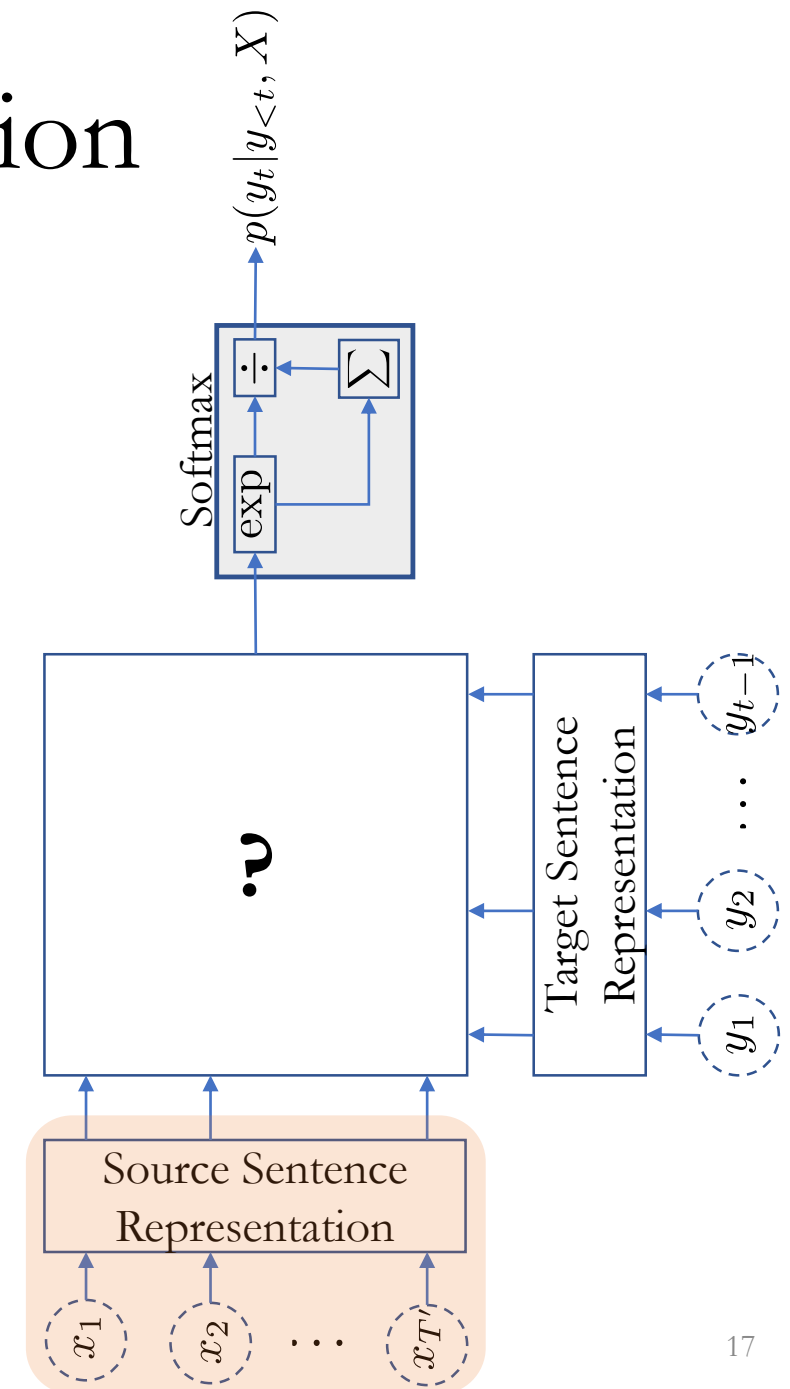
[Bahdanau et al., 2015]

1. Source sentence representation

- A stack of bidirectional RNN's



- The extracted vector at each location is a **context-dependent vector representation**.

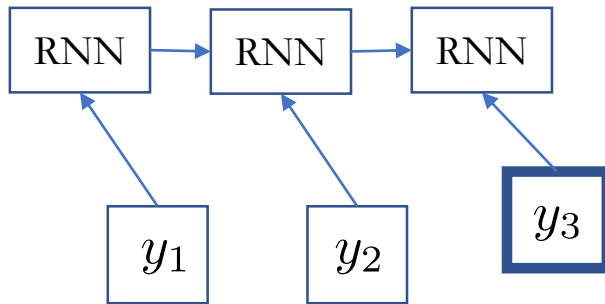


RNN Neural Machine Translation

[Bahdanau et al., 2015]

2. Target prefix representation

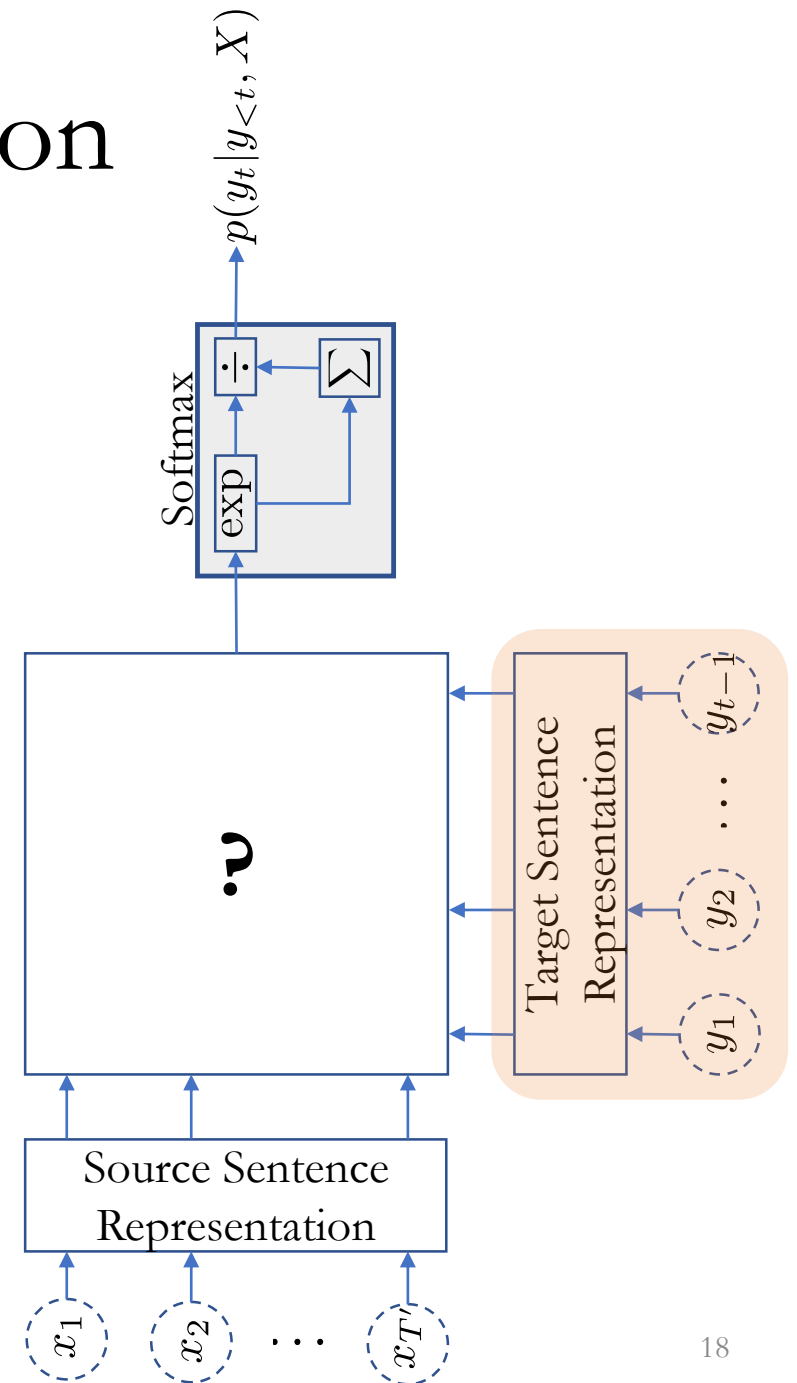
- A unidirectional recurrent network



- Compression of the target prefix

$$z_t = \text{RNN}_{\text{decoder}}(z_{t-1}, y_{t-1})$$

- Summarizes what has been translated so far

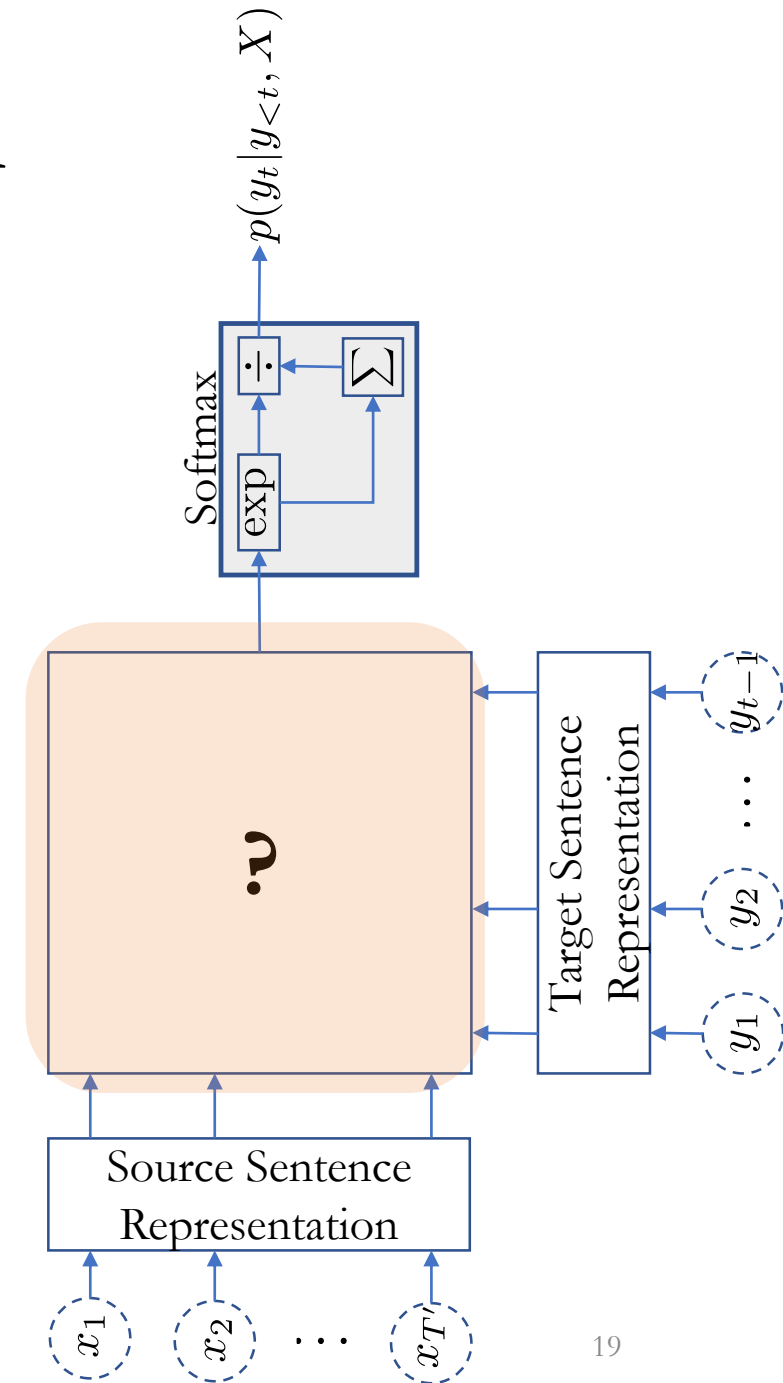
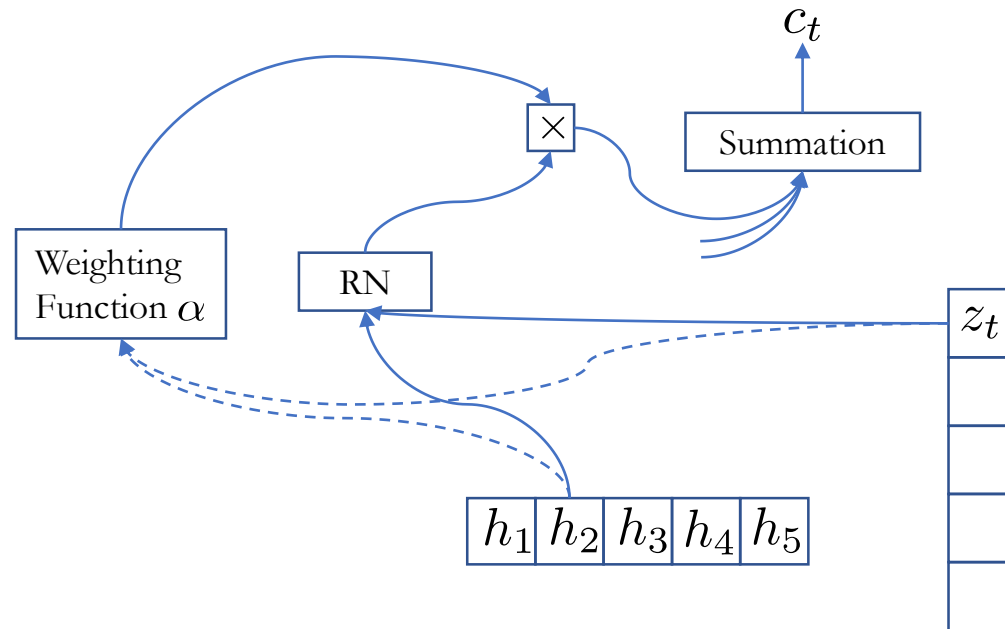


RNN Neural Machine Translation

[Bahdanau et al., 2015]

3. Attention mechanism

- Which part of the source sentence is relevant for predicting the next target token?
- Recall self-attention from Lecture 2

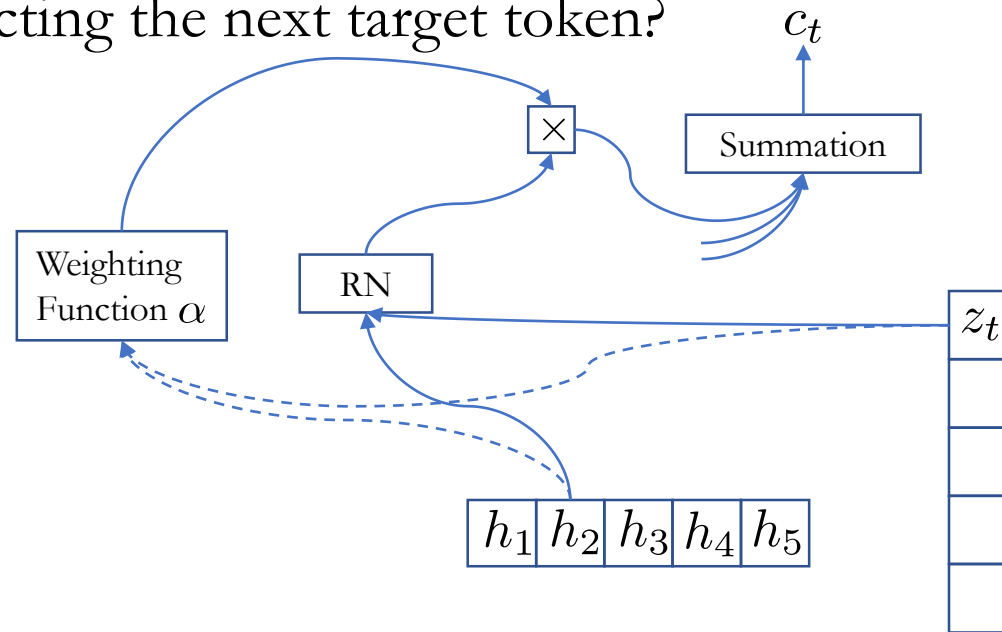


RNN Neural Machine Translation

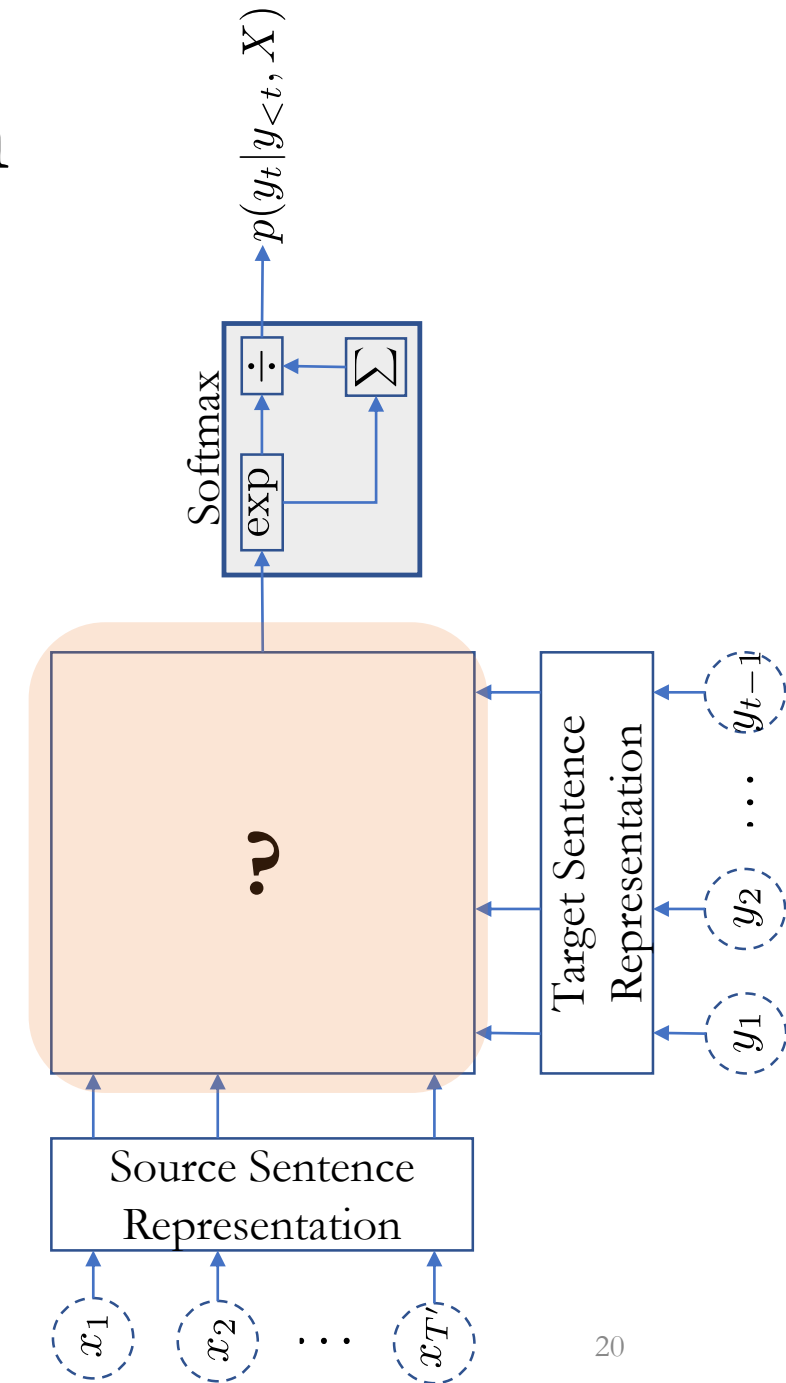
[Bahdanau et al., 2015]

3. Attention mechanism

- Which part of the source sentence is relevant for predicting the next target token?



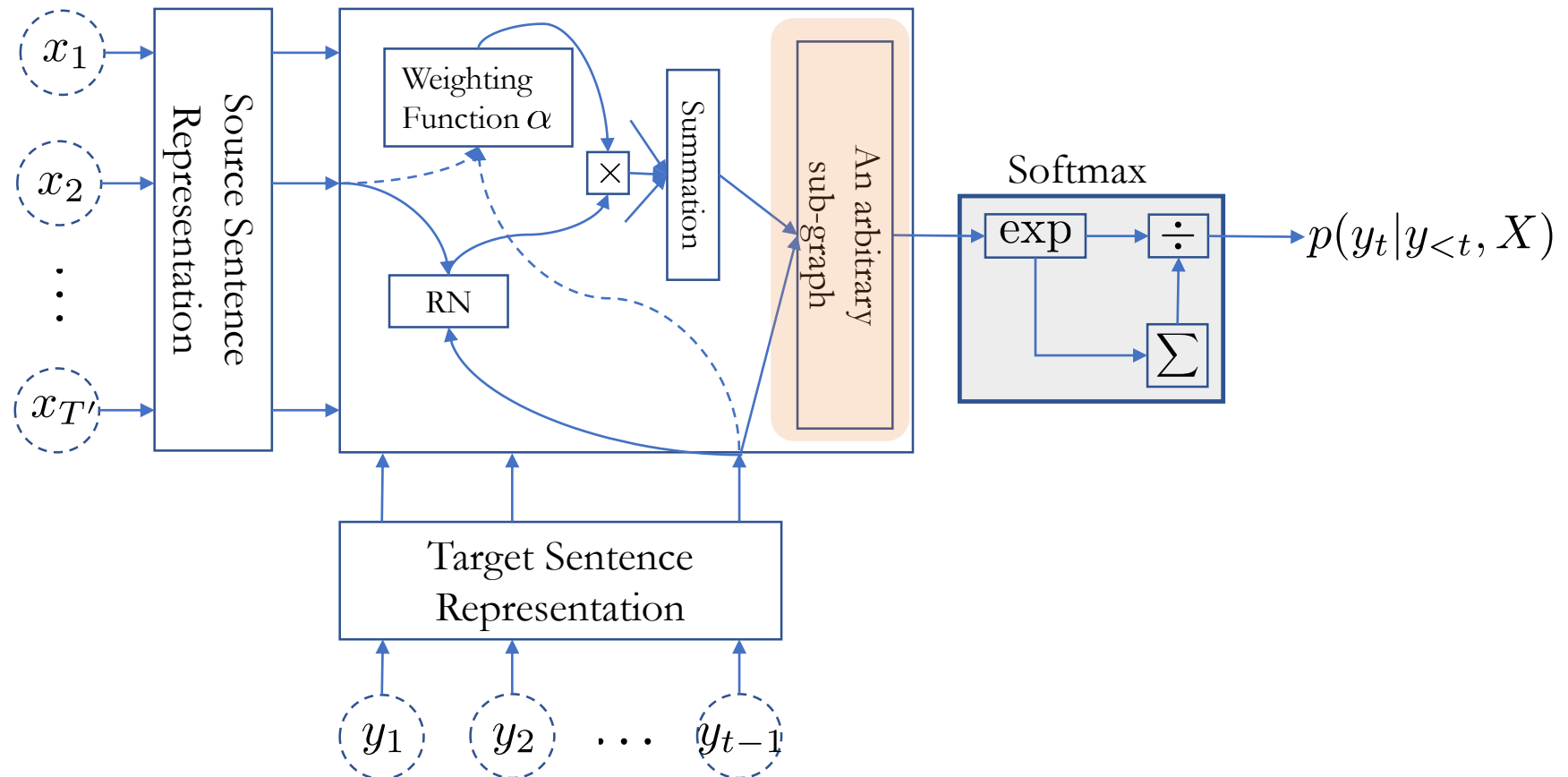
- Time-dependent source context vector c_t



RNN Neural Machine Translation

[Bahdanau et al., 2015]

4. Fuse the source context vector and target prefix vector
- Combines z_t and c_t into a single vector

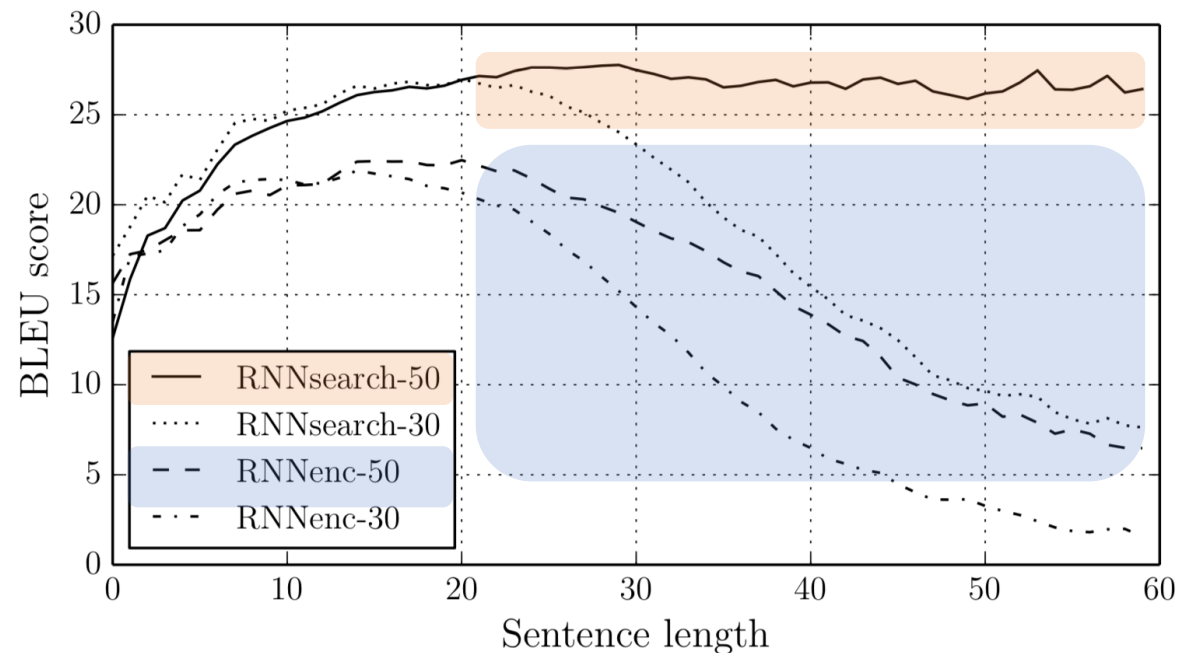


RNN Neural Machine Translation

- Conceptual process
 1. Encode: read the entire source sentence to know what to translate
 2. Attention: at each step, decide which source token(s) to translate next
 3. Decode: based on what has been translated and what need to be translated, predict the next target token.
 4. Repeat 2-3 until the <end-of-sentence> special token is generated.

RNN Neural Machine Translation

- The model is not pressured to compress the entire source sentence into a single, fixed-size vector:
 - Greatly improves the translation quality, especially of long sentences.
 - Much more efficient: less parameters are necessary.
- Bahdanau et al. [2015] showed for the first time the machine translation purely based on neural networks could be as good as then-state-of-the-art alternatives (e.g., PBMT).



RNN Neural Machine Translation

- **Source:** *An admitting privilege is the right of a doctor to admit a patient to a hospital or a medical centre to carry out a diagnosis or a procedure, based on his status as a health care worker at a hospital.*
- **When collapsed:** *Un privilège d'admission est le droit d'un médecin de reconnaître un patient à l'hôpital ou un centre médical d'un diagnostic ou de prendre un diagnostic en fonction de son état de santé.*
- **RNNSearch:** *Un privilège d'admission est le droit d'un médecin d'admettre un patient à un hôpital ou un centre médical pour effectuer un diagnostic ou une procédure, selon son statut de travailleur des soins de santé à l'hôpital.*

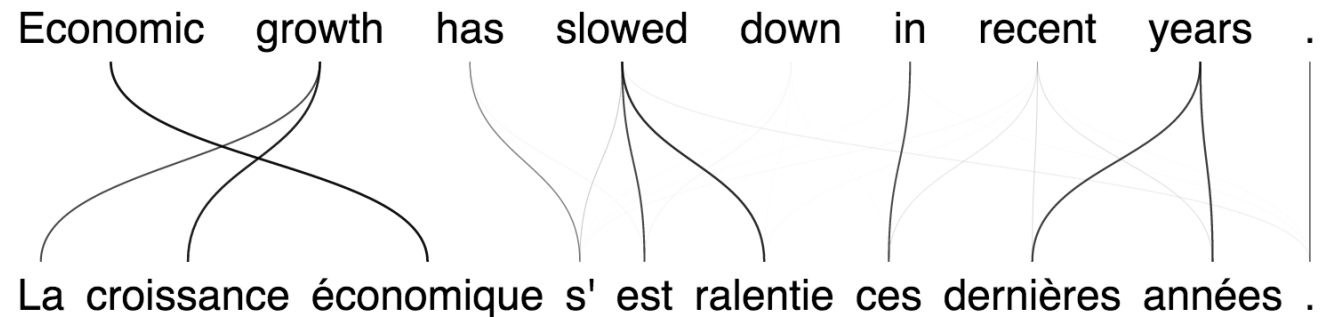
RNN Neural Machine Translation

- **Source:** *An admitting privilege is the right of a doctor to admit a patient to a hospital or a medical centre to carry out a diagnosis or a procedure, based on his status as a health care worker at a hospital.*
- **When collapsed:** *Un privilège d'admission est le droit d'un médecin de reconnaître un patient à l'hôpital ou un centre médical d'un diagnostic ou de prendre un diagnostic en fonction de son état de santé.*
- **RNNSearch:** *Un privilège d'admission est le droit d'un médecin d'admettre un patient à un hôpital ou un centre médical pour effectuer un diagnostic ou une procédure, selon son statut de travailleur des soins de santé à l'hôpital.*

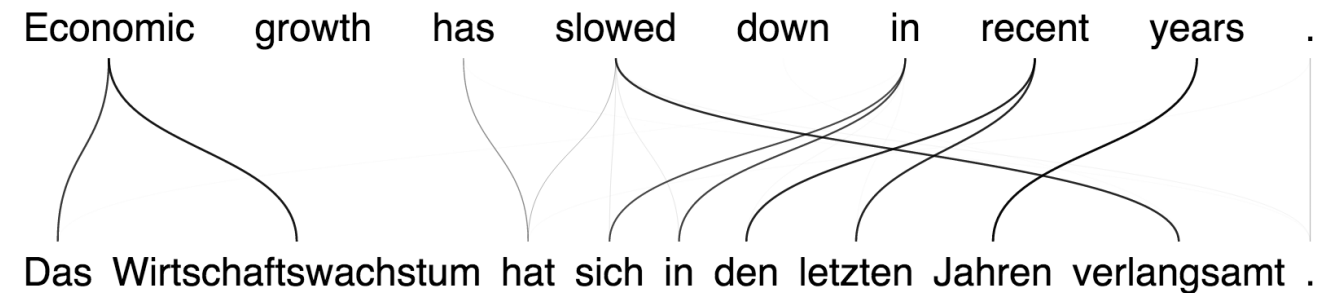
RNN Neural Machine Translation

- Sensible alignment between source and target tokens
- Capture long-range reordering/dependencies
- Without strong supervision on the alignment
 - Weakly supervised learning

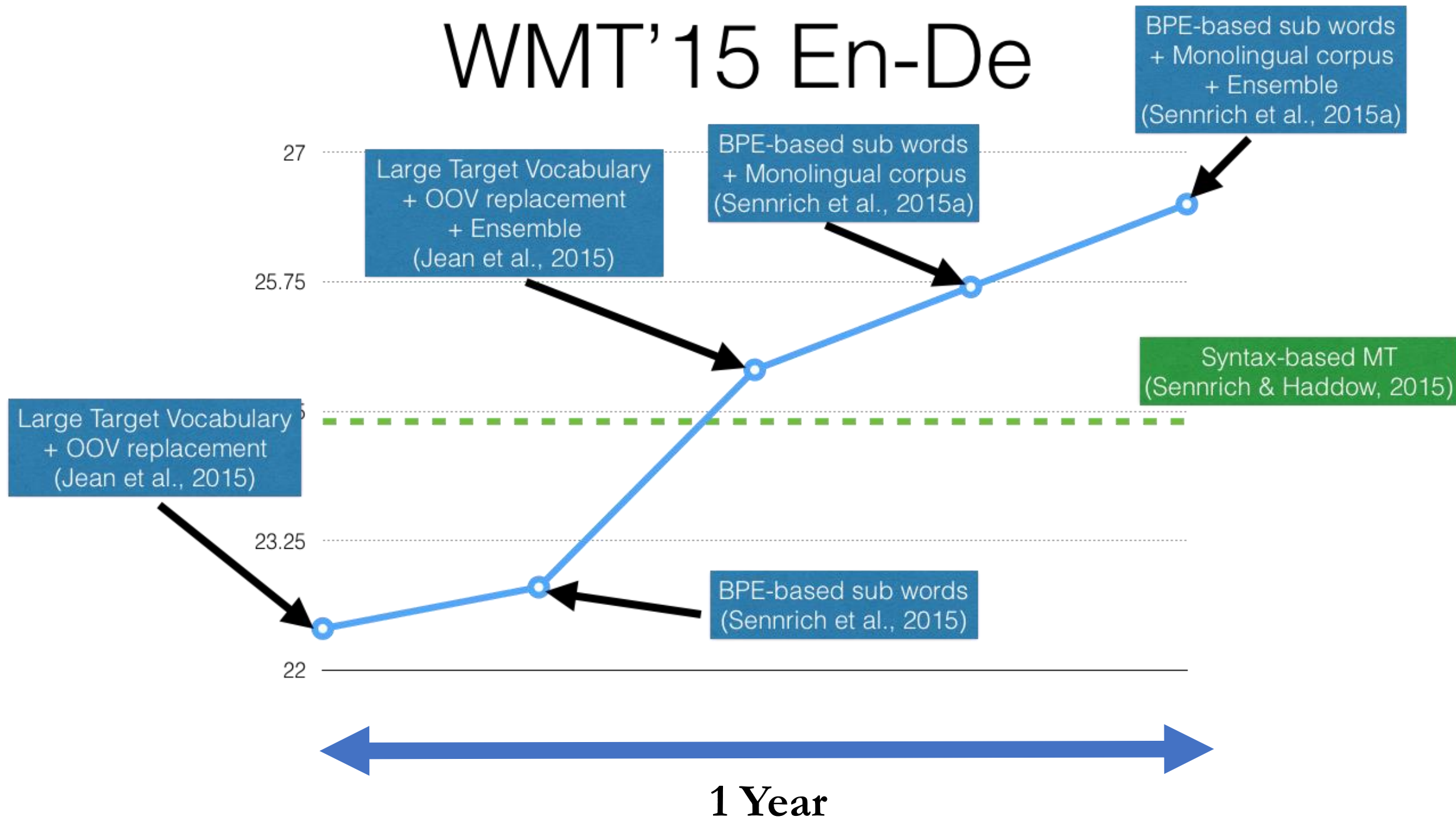
English-French











English-German



WMT'15 En-De



		output language						
input language	Czech 		barry uedin-nmt					
		German 	rsennrich uedin-nmt-					
	barry uedin-nmt	rsennrich uedin-nmt-	English 	jorgtied HY-HNMT	tilde tilde-nc-s	lexi EDIN-ens4	annacurrey uedin-nmt	barry uedin-nmt
			jhu-smt Moses	Finnish 				
			tilde tilde-nc-n		Latvian 			
			jeremy.gwinnup afri-mitll			Russian 		
			annacurrey uedin-nmt				Turkish 	
			Zhixing Tan xmunmt_ens					Chinese 

WMT 2017: news translation task

A Neural Network for Machine Translation, at Production Scale

Tuesday, September 27, 2016

Posted by Quoc V

Ten years ago, w
Based Machine
machine intelligence
improving mach

Today
of-th
quali

Amazon Translate

Natural and fluent language translation

Try the Preview

Sysran launch translation engine

Language barriers represent one of the most significant challenges in the world. Now, thanks to advances in artificial intelligence, we are beginning to see solutions.

By Eileen Brown for Social Business | November 2016



Inside the EPO's Machine-Powered Mission to Unlock Europe's Multilingual Patents

by Eden Estopace on June 6, 2017

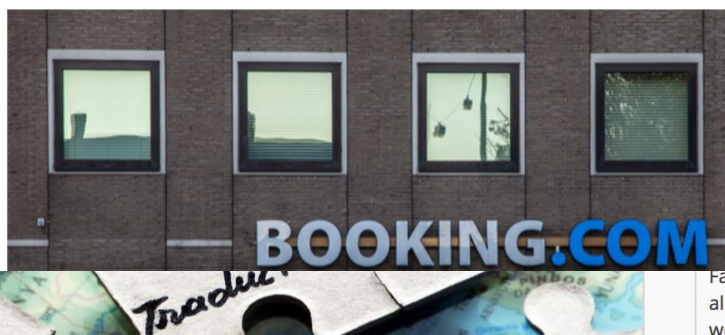


Adoption of Neural Machine Translation (NMT) in production environments is gathering pace. In a blog post on May 15,

PO)
e, the
n 2013
ite is a
rch

Booking.com Builds on Harvard Framework to Run Neural MT at Scale

by Eden Estopace on July 31, 2017



Three major trends shaping the language technology space converged at Booking.com in what is likely a harbinger of things to come in the language industry.

work based

Facebook is an online platform that allows its users to connect with friends and family, as well as make new connections.

Share 461

Rat

Microsoft Translator is now powering all speech translation through state-of-the-art neural networks.

In practice,

- Many excellent open-source packages exist:
 - Nematus <https://github.com/EdinburghNLP/nematus>
 - Compute backend: TensorFlow (originally Theano)
 - Used to build state-of-the-art translation systems for WMT'16 and WMT'17.
 - Supported by U. Edinburgh (Rico Sennrich's group)
 - OpenNMT-py <https://github.com/OpenNMT/OpenNMT-py>
 - Compute backend: PyTorch (originally Lua-Torch)
 - Implements latest architectures and algorithms
 - Supported by Harvard NLP (Sasha Rush's group)
 - FairSeq <https://github.com/facebookresearch/fairseq>
 - Compute backend: PyTorch
 - Focuses on the convolutional seq2seq
 - Supported by Facebook AI Research
 - Sockeye <https://github.com/awslabs/sockeye>
 - Compute backend: MXNet
 - Supported by Amazon

In practice,

- Many new architectures are being proposed constantly
- Convolutional sequence-to-sequence models [Gehring et al., 2017]
 - Encoder: CNN-based sentence representation
 - Decoder: CNN-based conditional language model
- Transforms [Vaswani et al., 2017]
 - Encoder: Self-attention based sentence representation
 - Decoder: Self-attention based conditional language model
- It has been three years only, and a long road lies ahead...

Further application – machine reading

- Input: a paragraph + a question
- Output: the position of the answer in the paragraph

Paragraph

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, graupel and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud. Short, in- tense periods of rain in scattered locations are called “showers”.

Question

What causes precipitation to fall?

Further application – machine reading

1. Encode the paragraph into a set of continuous vectors $\{p_1, p_2, \dots, p_{T'}\}$
 - e.g., use a bidirectional recurrent network.
2. Encode the question into a single, fixed-size vector q
 - e.g., use a self-attention network followed by averaging.
3. Compare each token in the paragraph against the question using the RN module: $s_i = \sigma(\text{RN}(p_i, q))$
4. Turn the scores into a Categorical distribution using softmax.
5. Train it by maximizing the log-probability of the correct answer.

Nothing really changes other than the network architecture (DAG)

In this lecture, we've learned

- What machine translation is:
 - It maps a sentence in a source language into its translation in a target language.
- What neural machine translation is:
 - A single neural network is used to approximate the entire translation process.
- How to build an RNN neural machine translation system:
 - Encoder: a bidirectional RNN
 - Decoder: a unidirectional RNN coupled with the attention mechanism

From here on...

- I will cover a few recent research projects:
 1. Multimedia description
 2. Character-level neural machine translation
 3. Non-parametric neural machine translation
 4. Real-time translation