UCU ML Summer Workshops 2017

Natural Language Processing: Part 2

DataRobot

Yuriy Guts
ML Engineer

Remaining Agenda

Day 4: Short Text Similarity & Paraphrase Detection

- Paraphrase identification problem
- Traditional and DL approaches to semantic similarity
- Architectural principles for state-of-the-art neural NLP
- DL feature extraction for NLP

Day 5: Sequence-to-Sequence Modeling

- Attention mechanisms in DL for NLP
- Sequence-to-sequence neural models

Paraphrase Identification

Where can I get very professional and reliable envelope printing service in Sydney?

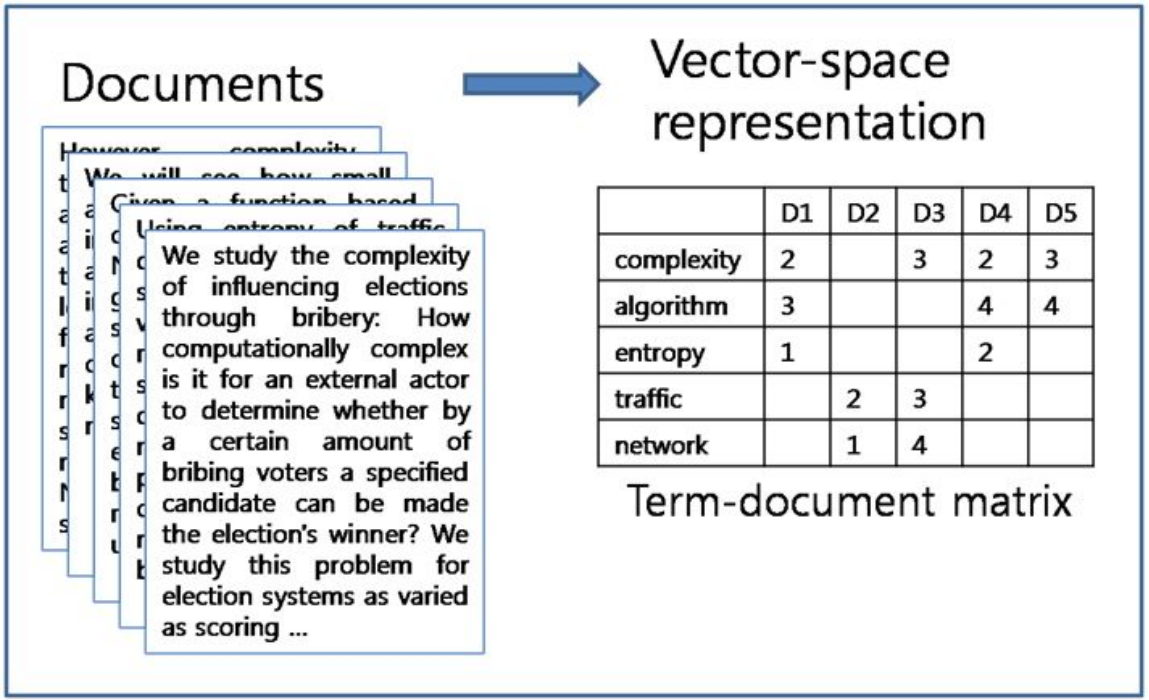
Where can I get very affordable branded envelope printing service in Sydney?

Why are doctors always late?

Why doctors always make you wait for 15-20 minutes before they see you?

Source: Quora Question Pairs Dataset
https://data.quora.com/First-Quora-Dataset-Release-Question-Pairs

Recap: BoW Representations



	D1	D2	D3	D4	D5
complexity	2		3	2	3
algorithm	3			4	4
entropy	1			2	
traffic		2	3		
network		1	4		

Image copyright © S. Mukherjee

TFIDF

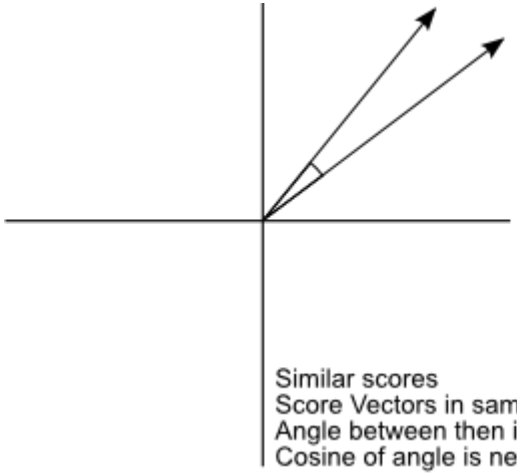
For a term i in document j:

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

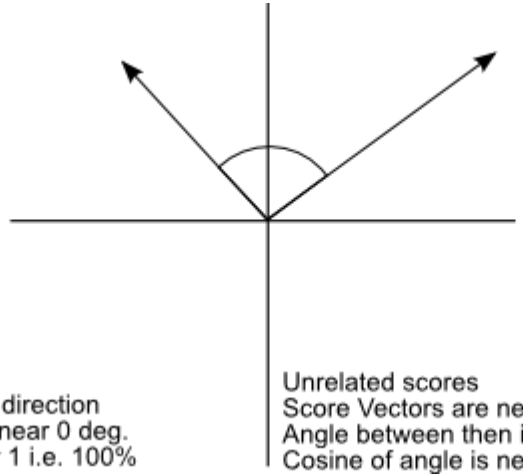
$tf_{i,j}$ = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents

Recap: Cosine Similarity

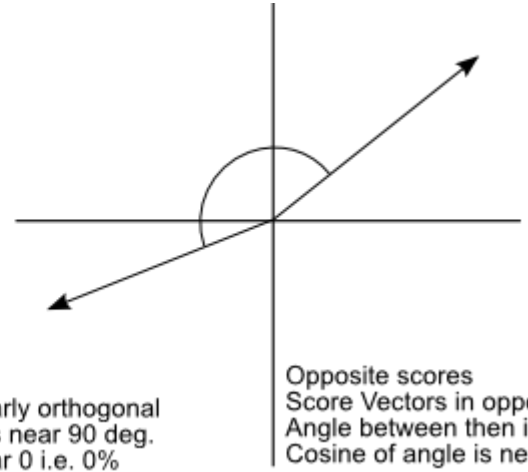
$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|_2 \|\mathbf{B}\|_2} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$



Similar scores
Score Vectors in same direction
Angle between then is near 0 deg.
Cosine of angle is near 1 i.e. 100%



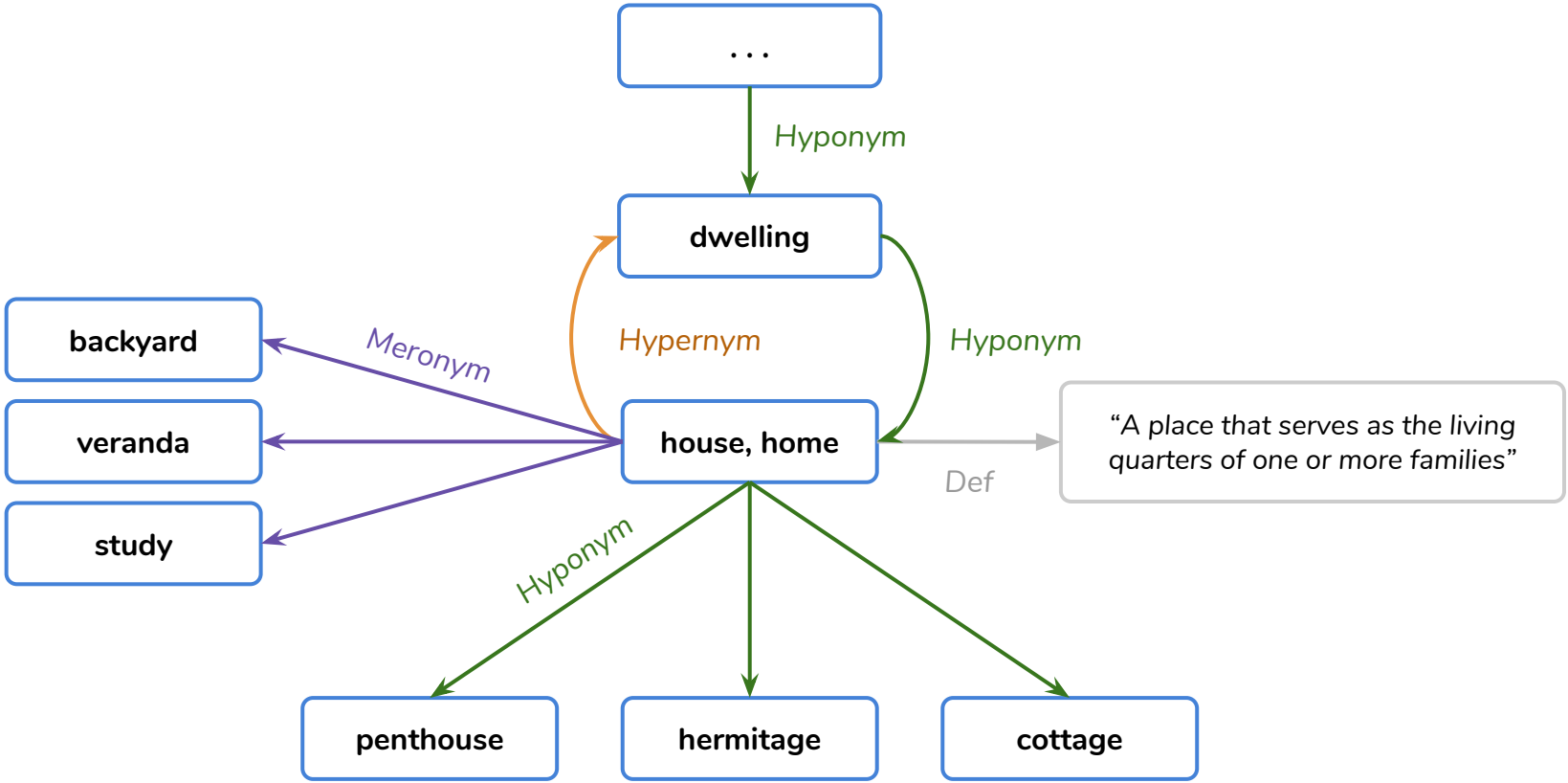
Unrelated scores
Score Vectors are nearly orthogonal
Angle between then is near 90 deg.
Cosine of angle is near 0 i.e. 0%



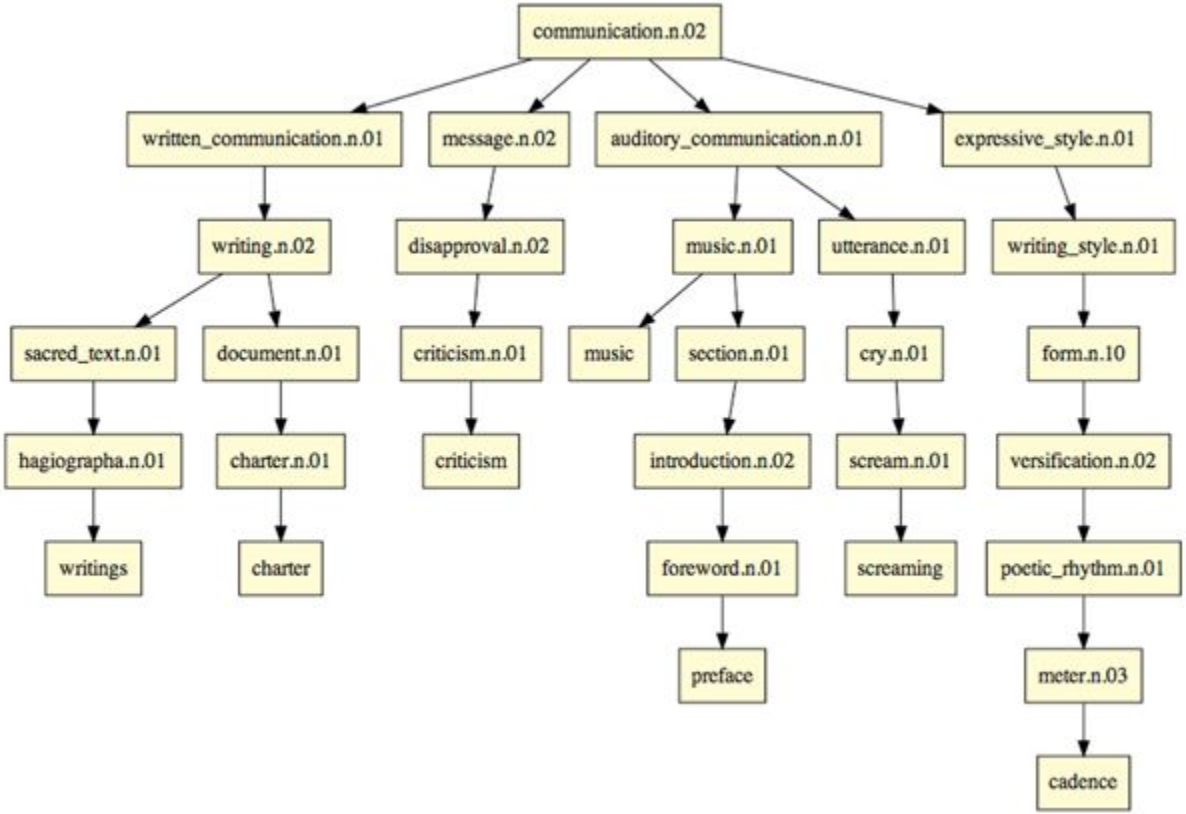
Opposite scores
Score Vectors in opposite direction
Angle between then is near 180 deg.
Cosine of angle is near -1 i.e. -100%

Image copyright © C. Perone

Lexical Databases and Ontologies



WordNet


communication.n.02
written_communication.n.01
message.n.02
auditory_communication.n.01
expressive_style.n.01
writing.n.02
disapproval.n.02
music.n.01
utterance.n.01
writing_style.n.01
sacred_text.n.01
document.n.01
criticism.n.01
music
section.n.01
cry.n.01
form.n.10
hagiographa.n.01
charter.n.01
criticism
introduction.n.02
scream.n.01
versification.n.02
writings
charter
foreword.n.01
screaming
poetic_rhythm.n.01
preface
meter.n.03
cadence

Distributed Hypothesis of Language

“The complete meaning of a word is always contextual, and no study of meaning apart from context can be taken seriously”

John R. Firth. The technique of semantics.

Transactions of the Philological Society, 1935.

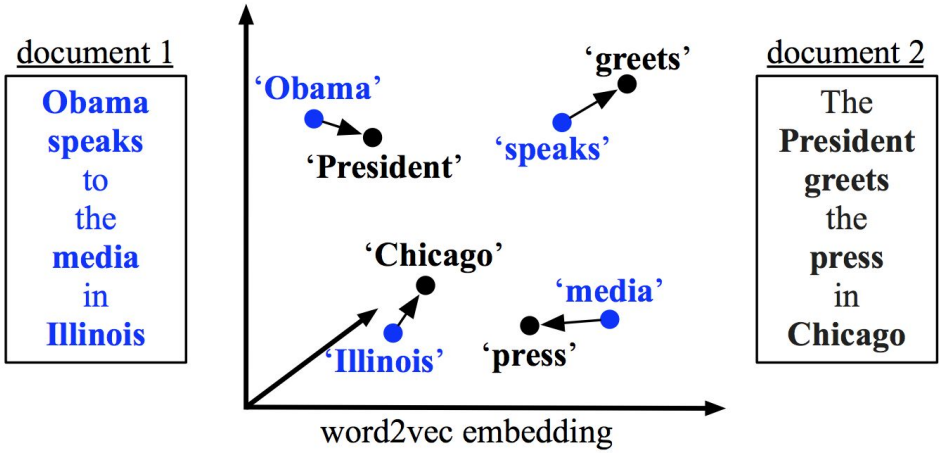
Distributional Hypothesis of Language [Harris, 1954]

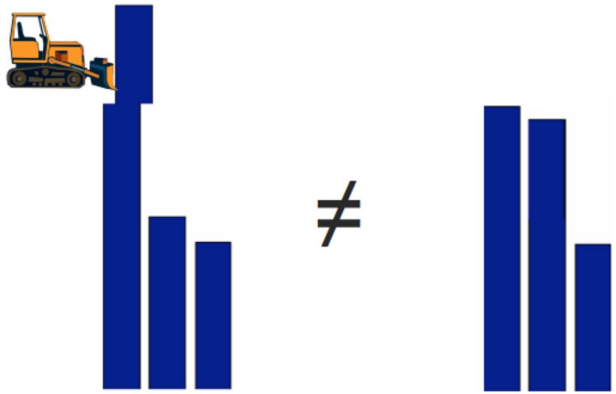
"Words that occur in the same contexts tend to have similar meanings"

The functional interplay of philosophy and	science	should, as a minimum, guarantee...
...and among works of dystopian	science	fiction...
The rapid advance in	science	today suggests...
...calculus, which are more popular in	science	-oriented schools.
But because	science	is based on mathematics...
...the value of opinions formed in	science	as well as in the religions...
...if	science	can discover the laws of human nature....
...is an art, not an exact	science	.
...factors shaping the future of our civilization:	science	and religion.
...certainty which every new discovery in	science	either replaces or reshapes.
...if the new technology of computer	science	is to grow significantly
He got a	science	scholarship to Yale.
...frightened by the powers of destruction	science	has given...
...but there is also specialization in	science	and technology...

Image copyright © D. Kartsaklis

Word Mover's Distance (WMD)

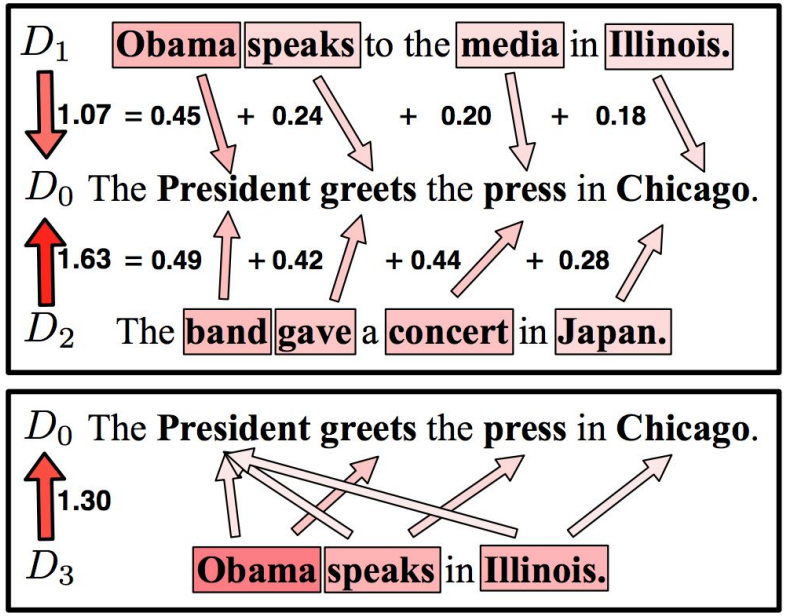




The minimum amount of "work" needed to transform document **1** to document **2**.
Inspired by Earth Mover's Distance, a well-studied transportation problem.

M. Kusner et al. "From Word Embeddings to Document Distances", 2015.
http://proceedings.mlr.press/v37/kusnerb15.pdf

WMD: Linear Optimization Problem



$$d_i = \frac{c_i}{\sum_{j=1}^{n} c_j}$$

nBOW frequency of the i-th word in the document

$$\mathbf{T}_{ij} \geq 0$$

"How much" of word ***i*** travels to word ***j***

$$\min_{\mathbf{T} \geq 0} \sum_{i,j=1}^{n} \mathbf{T}_{ij} c(i,j)$$

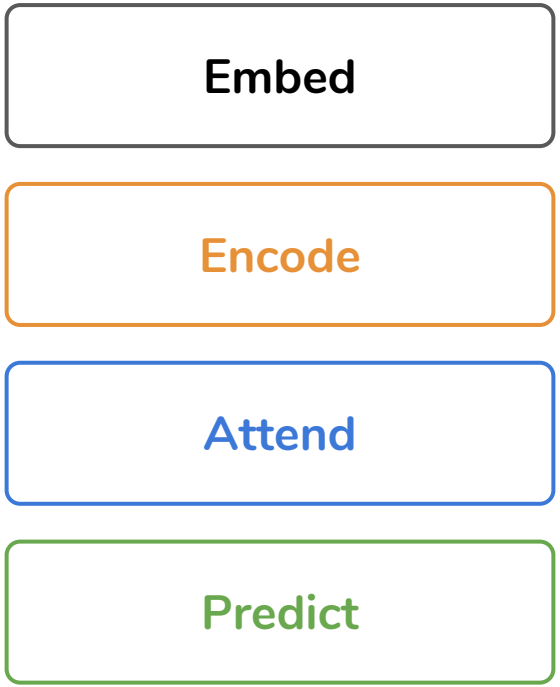
$$\text{subject to: } \sum_{j=1}^{n} \mathbf{T}_{ij} = d_i \quad \forall i \in \{1, \ldots, n\}$$

$$\sum_{i=1}^{n} \mathbf{T}_{ij} = d'_j \quad \forall j \in \{1, \ldots, n\}.$$

M. Kusner et al. "From Word Embeddings to Document Distances", 2015.
http://proceedings.mlr.press/v37/kusnerb15.pdf

Architectural Principles

for State-of-the-Art Neural NLP


Embed
Encode
Attend
Predict

State-of-the-Art NLP Pipeline

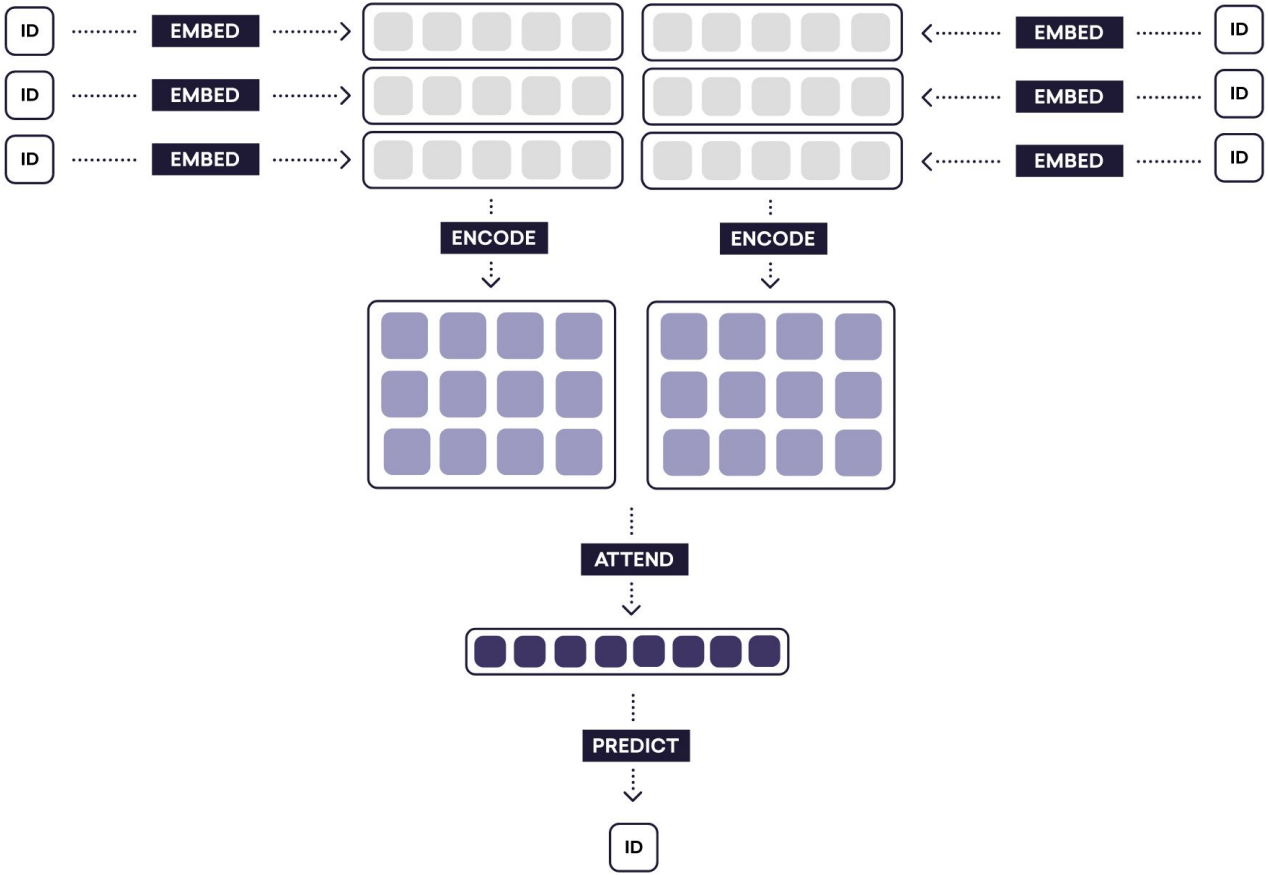


Image copyright © Explosion AI

Embed

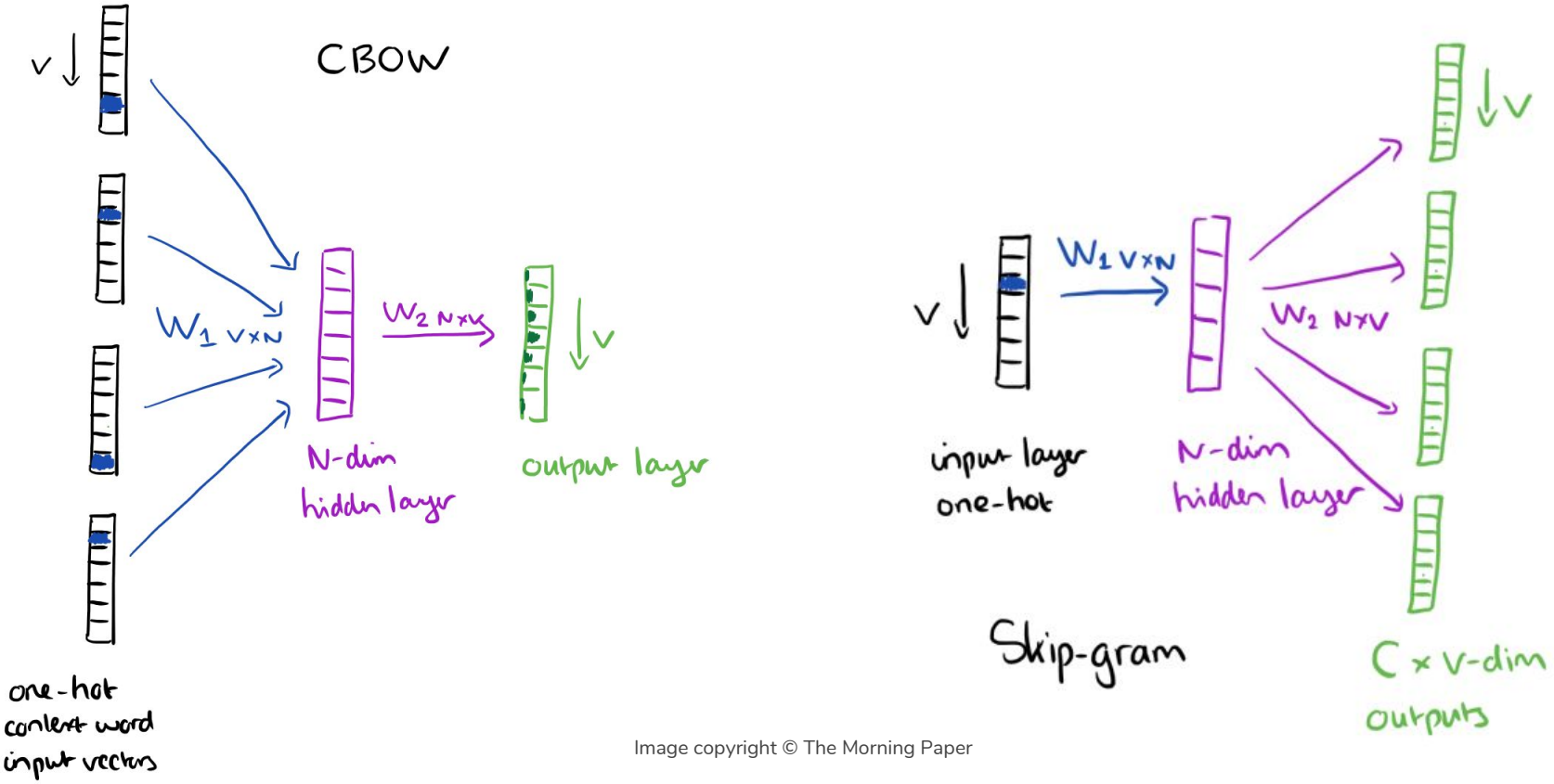


Image copyright © The Morning Paper

Encode: Convolutions over Word Embeddings

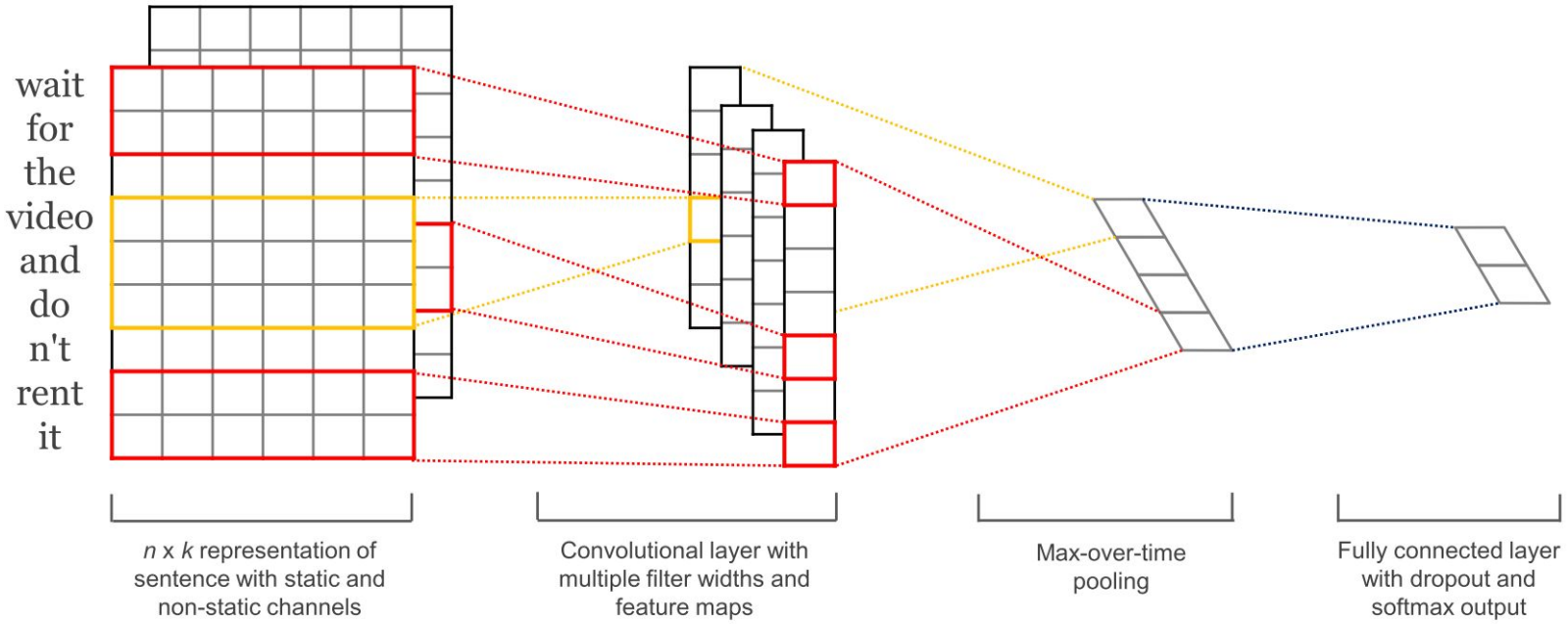
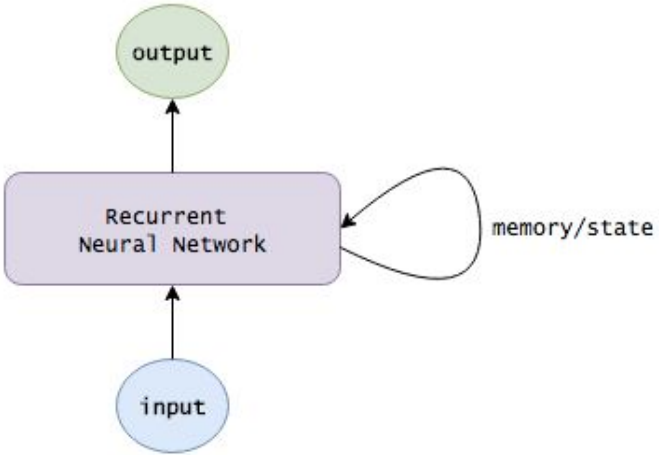


Image copyright © WildML

Y. Kim. “Convolutional Neural Networks for Sentence Classification” [2014]

Encode: RNNs over Word Embeddings



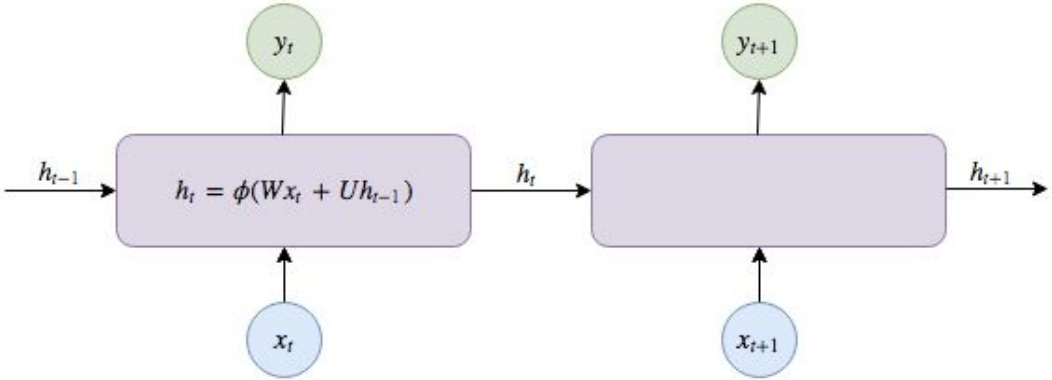


Image copyright © Edwin Chen

Encode: RNNs over Word Embeddings

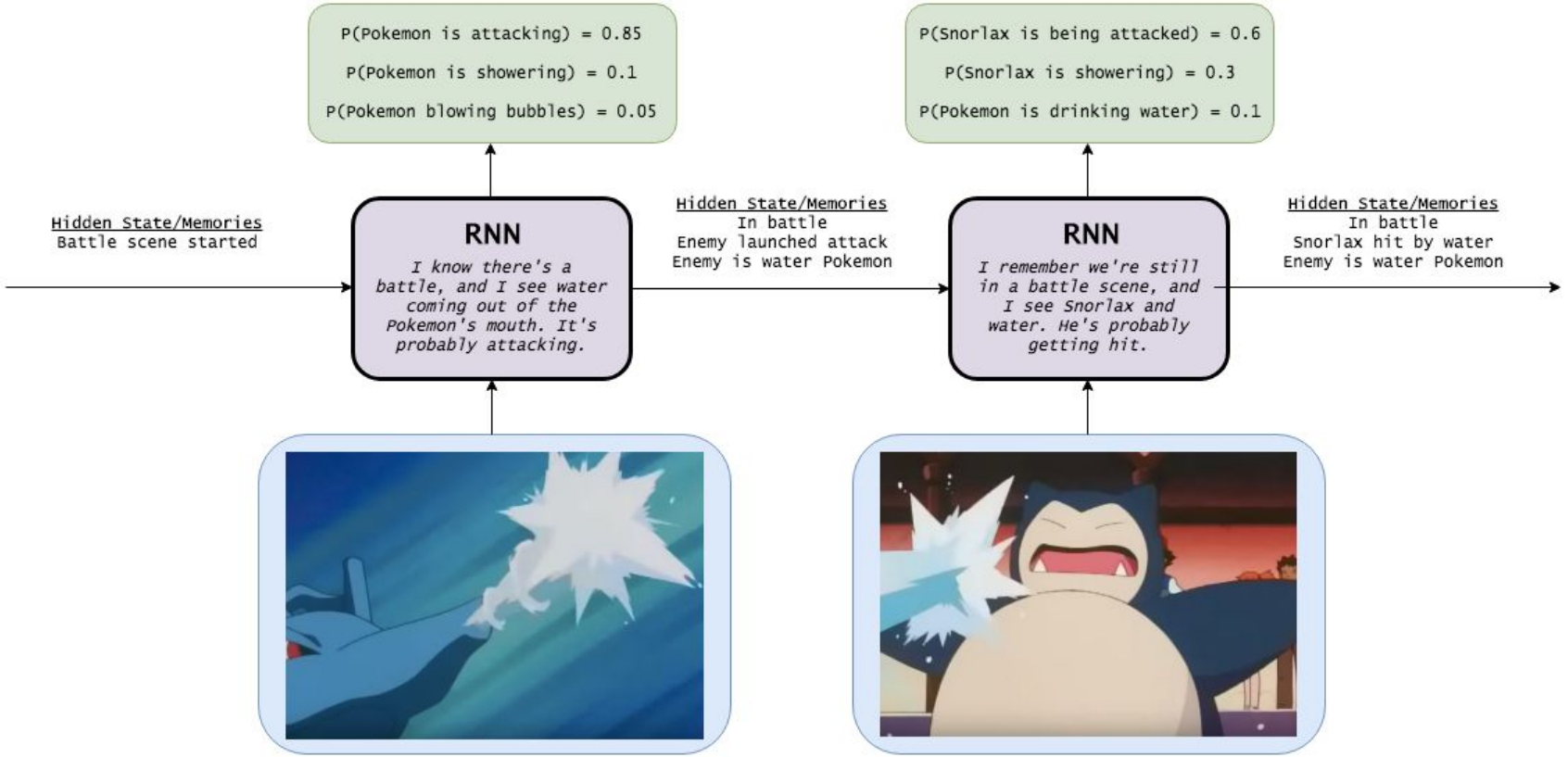
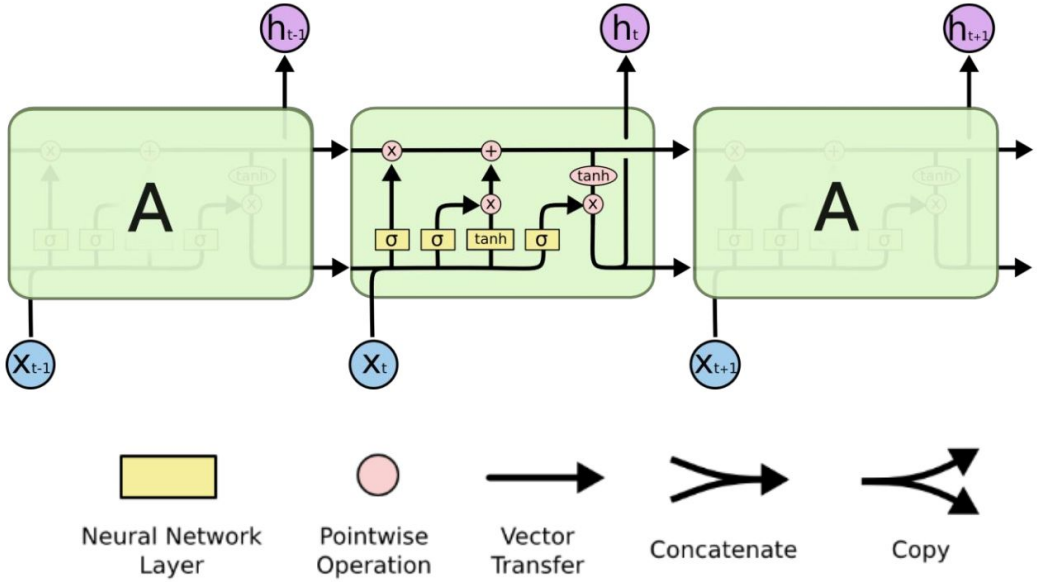


Image copyright © Edwin Chen

Encode: Gated Recurrent Architectures



$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \; + \; b_f\right)$$

$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \; + \; b_i\right)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \; + \; b_C)$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

$$o_t = \sigma\left(W_o \; [h_{t-1}, x_t] \; + \; b_o\right)$$

$$h_t = o_t * \tanh\left(C_t\right)$$

Image copyright © Chris Olah

Encode: Gated Recurrent Architectures

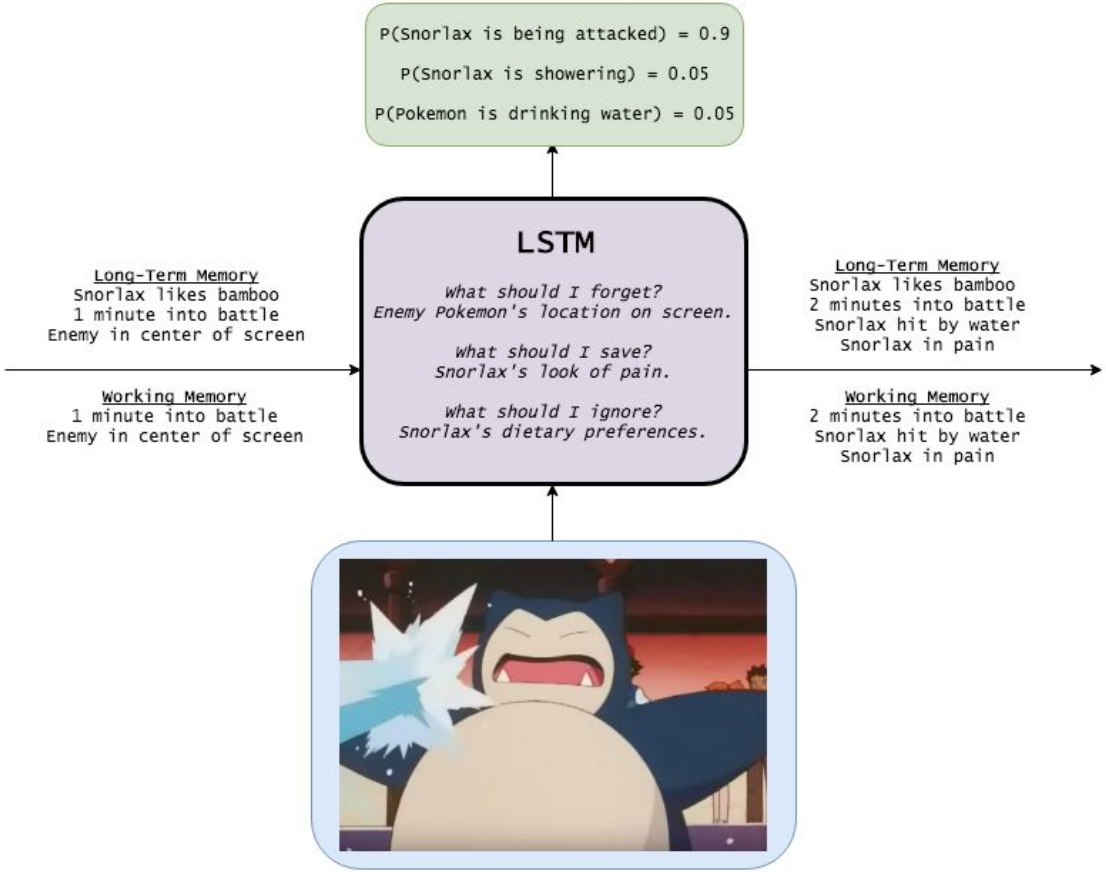
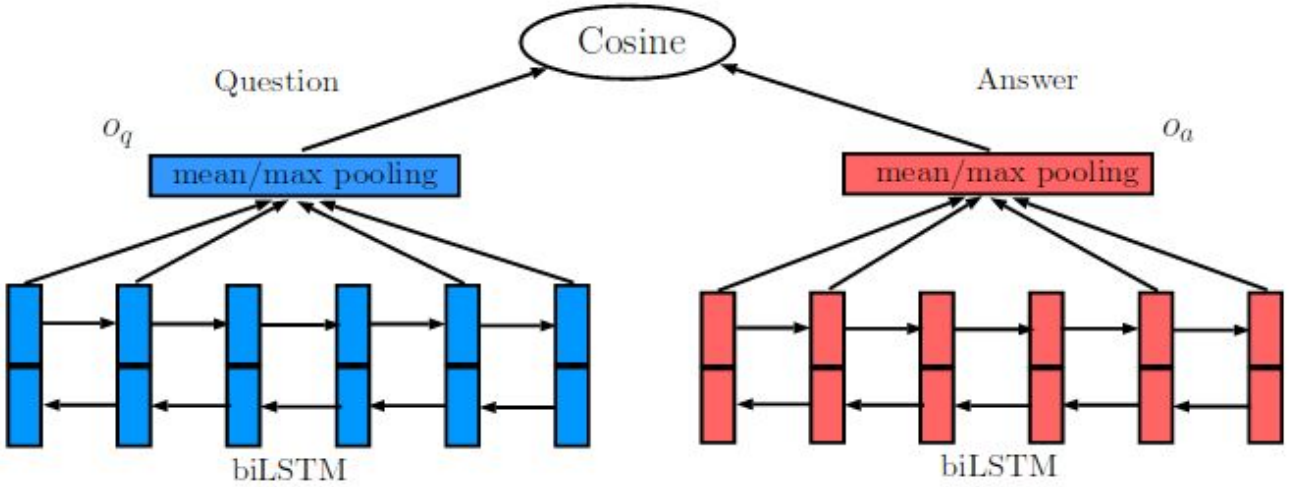


Image copyright © Edwin Chen

Two Input Documents? Siamese Network



Run the same encoding step for every input.

Share encoder weights, don't learn W_{E1} and W_{E2} separately.

Distance Learning, Contrastive Loss

$$L(W, Y, \vec{X}_1, \vec{X}_2) = (1 - Y)\frac{1}{2}(D_W)^2 + (Y)\frac{1}{2}\{max(0, m - D_W)\}^2$$

Penalize similar pairs by a monotonically **increasing** function of their learned distance.

Penalize different pairs by a monotonically **decreasing** function of their learned distance.

Hadsell, Chopra, LeCun “Dimensionality Reduction by Learning an Invariant Mapping”, 2006.
http://yann.lecun.com/exdb/publis/pdf/hadsell-chopra-lecun-06.pdf

Sequence-to-Sequence Modeling

and Attention Mechanisms in Neural NLP

State-of-the-Art NLP Pipeline

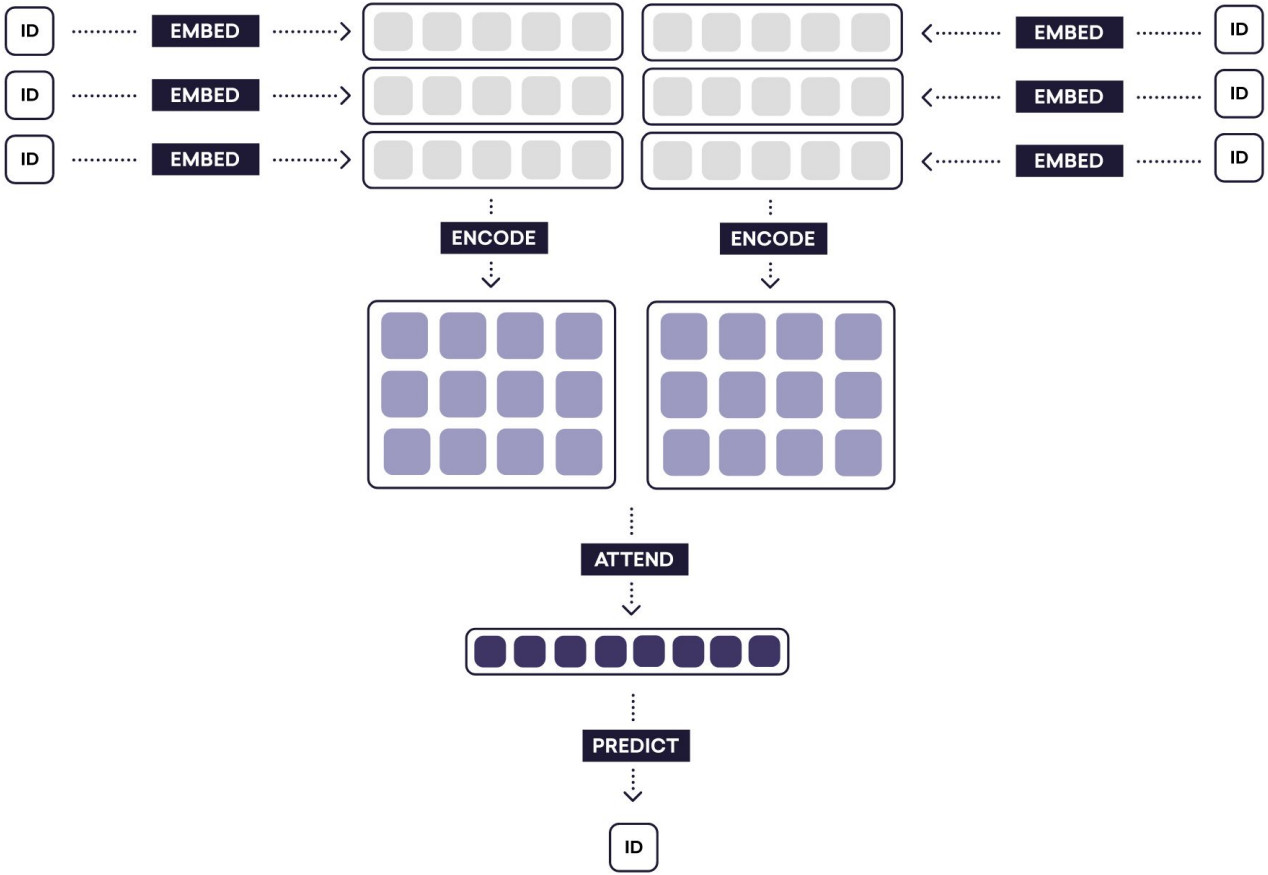


Image copyright © Explosion AI

Relative Complexity of NLP Problems

Easy

- Chunking
- Part-of-Speech Tagging
- Named Entity Recognition
- Spam Detection
- Spellchecking

Medium

- Syntactic Parsing
- Paraphrase Identification
- Sentiment Analysis
- Topic Modeling
- Information Retrieval

Hard

- Machine Translation
- Text Generation
- Automatic Summarization
- Question Answering
- Conversational Interfaces

Relative Complexity of NLP Problems

Easy

- Chunking
- Part-of-Speech Tagging
- Named Entity Recognition
- Spam Detection
- Spellchecking

Medium

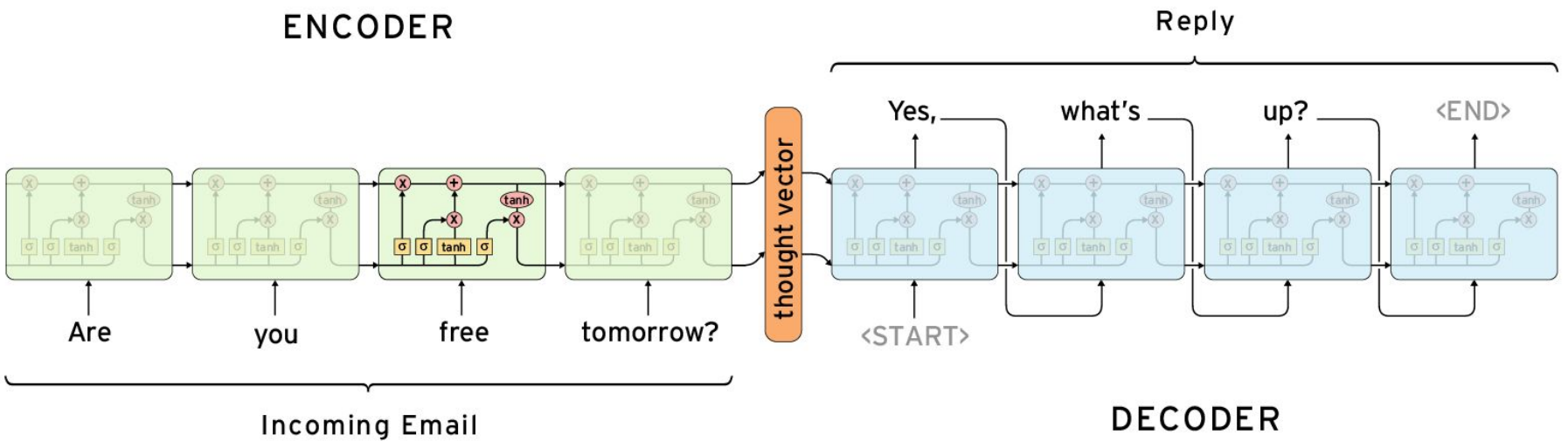
- Syntactic Parsing
- Paraphrase Identification
- Sentiment Analysis
- Topic Modeling
- Information Retrieval

Hard

- Machine Translation
- Text Generation
- Automatic Summarization
- Question Answering
- Conversational Interfaces

Seq2Seq can be applied here

Seq2Seq Modeling at a Glance

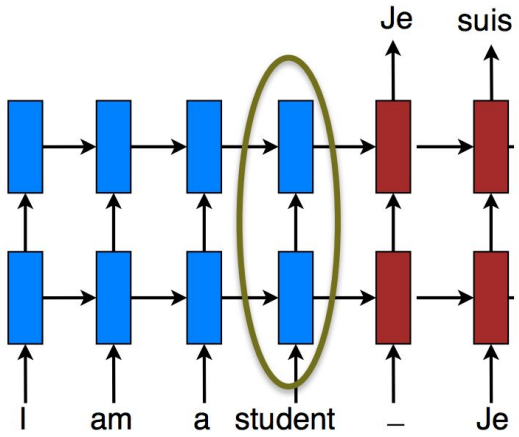


- Two different models: an **encoder** and a **decoder** (typically RNNs).
- Encoder represents an input document as a fixed size “thought vector”.
- Decoder unwinds the thought vector into an output sequence.

Image copyright © Suriyadeepan Ram

Vanilla Seq2Seq: Challenges

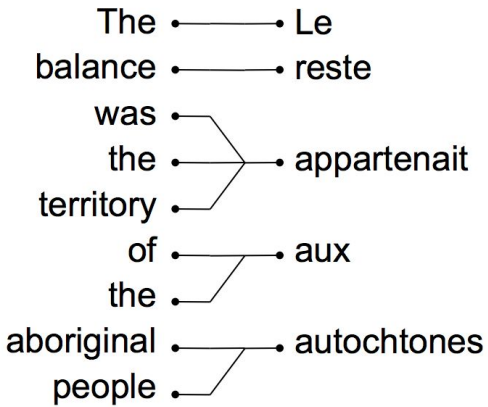
- Hard for encoder to compress the source documents into a fixed size vector.
- Performance deteriorates rapidly as the size of the document increases (e.g. good local grammar, but lots of word repetition).

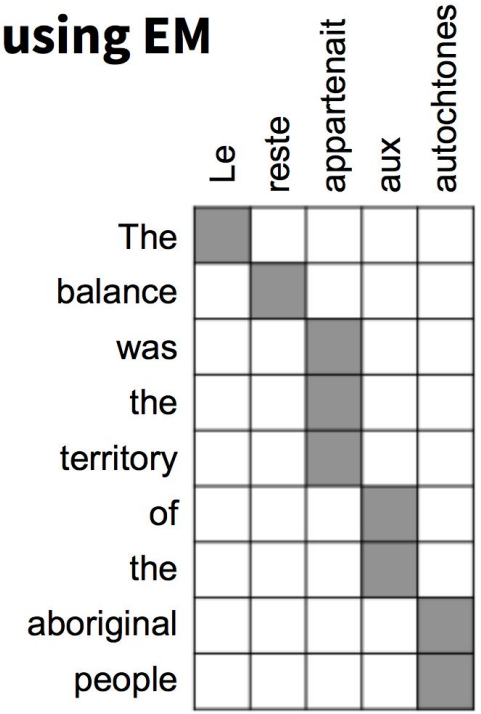


Content credit: Stanford CS224n

Traditional MT: “Word Alignment”

Phrase-based SMT aligned words in a preprocessing-step, usually using EM





Content credit: Stanford CS224n

Jointly Learning to Align and Translate

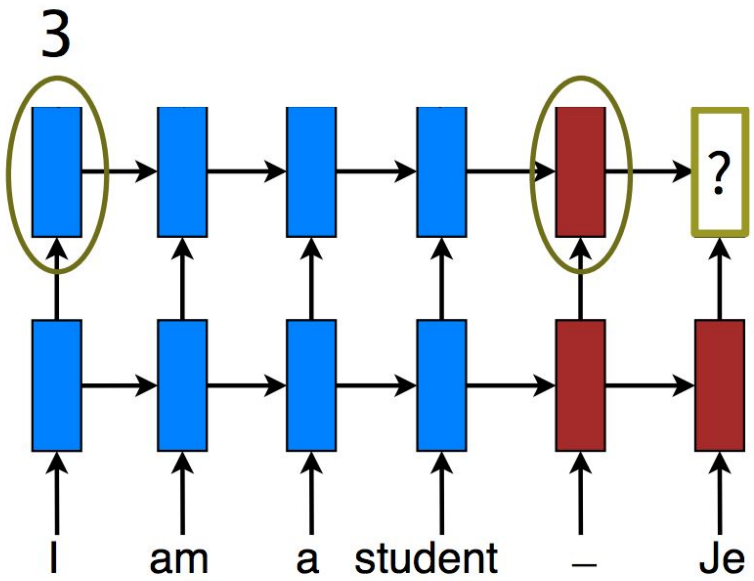
It would be a lot more beneficial to let the decoder **peek at the original input** at **each step** of the decoding phase.

D. Bahdanau, K. Cho et al. "Neural Machine Translation by Jointly Learning to Align and Translate", 2014.

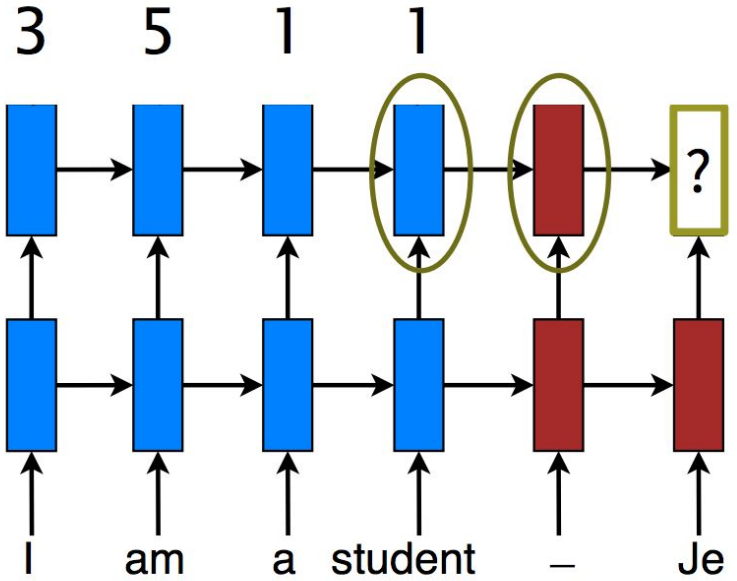
https://arxiv.org/pdf/1409.0473.pdf

Neural Attention: Scoring

$$\text{score}(\boldsymbol{h}_{t-1}, \bar{\boldsymbol{h}}_s)$$



$$\text{score}(\boldsymbol{h}_{t-1}, \bar{\boldsymbol{h}}_s)$$

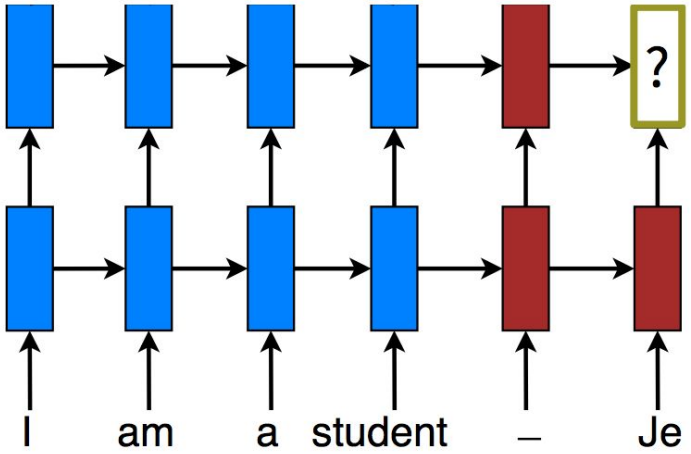


Content credit: Stanford CS224n

Neural Attention: Scoring

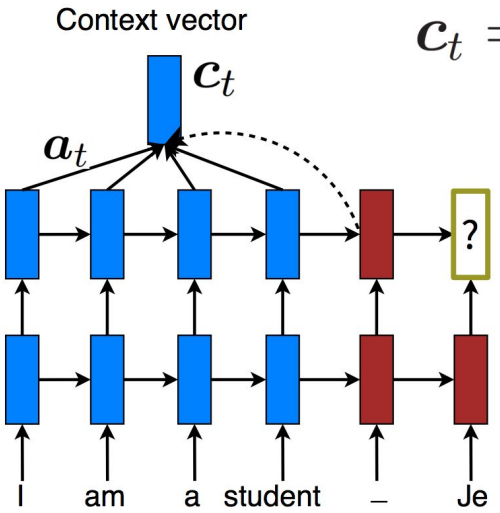
$$\boldsymbol{a}_t(s) = \frac{\mathrm{e}^{\mathrm{score}(s)}}{\sum_{s'} \mathrm{e}^{\mathrm{score}(s')}}$$

$\boldsymbol{a}_t$ 0.3 0.5 0.1 0.1

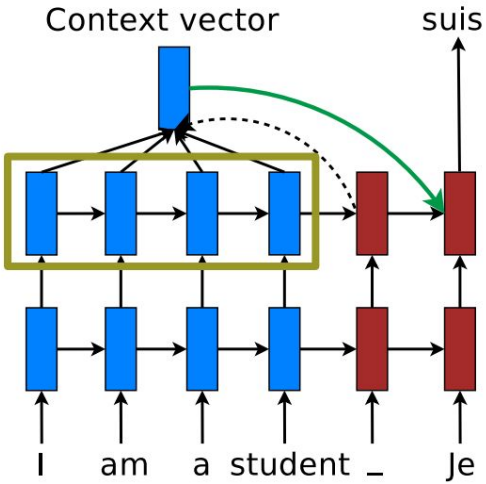


Content credit: Stanford CS224n

Attention Mechanism: Context

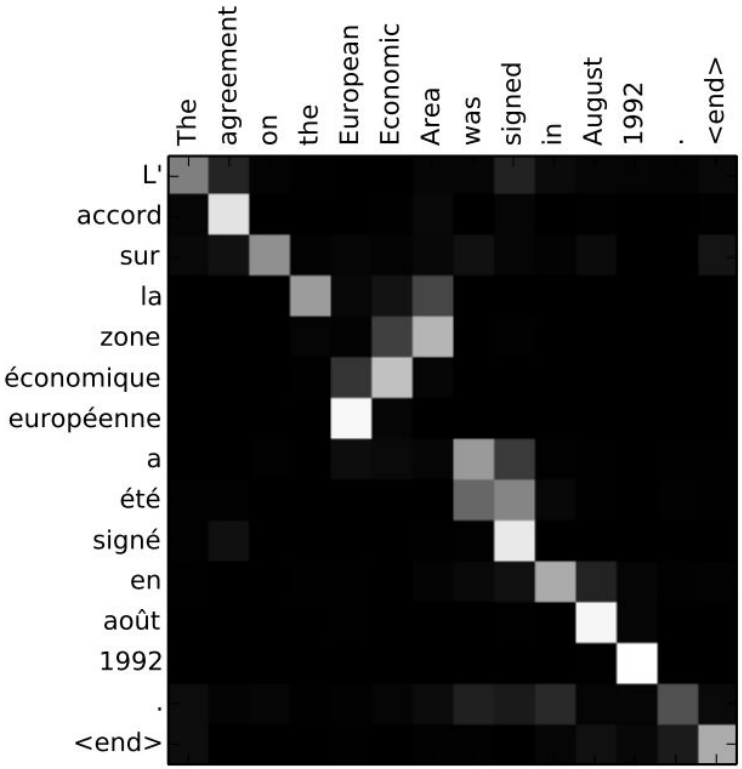


$$c_t = \sum_s a_t(s)\bar{h}_s$$



Content credit: Stanford CS224n

Interpretability of Attention Scores



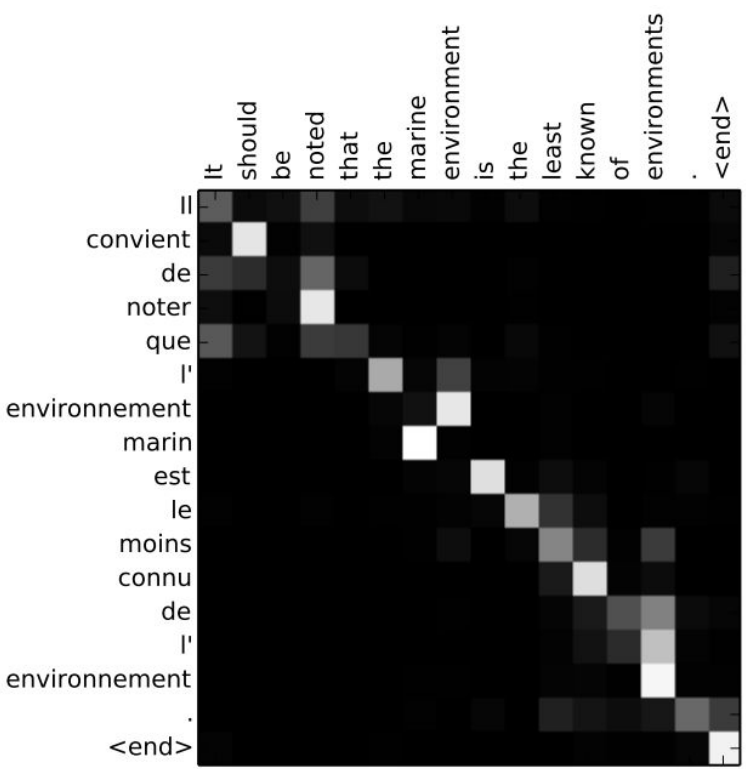


Image copyright © Bahdanau et al.

Interpretability of Attention Scores

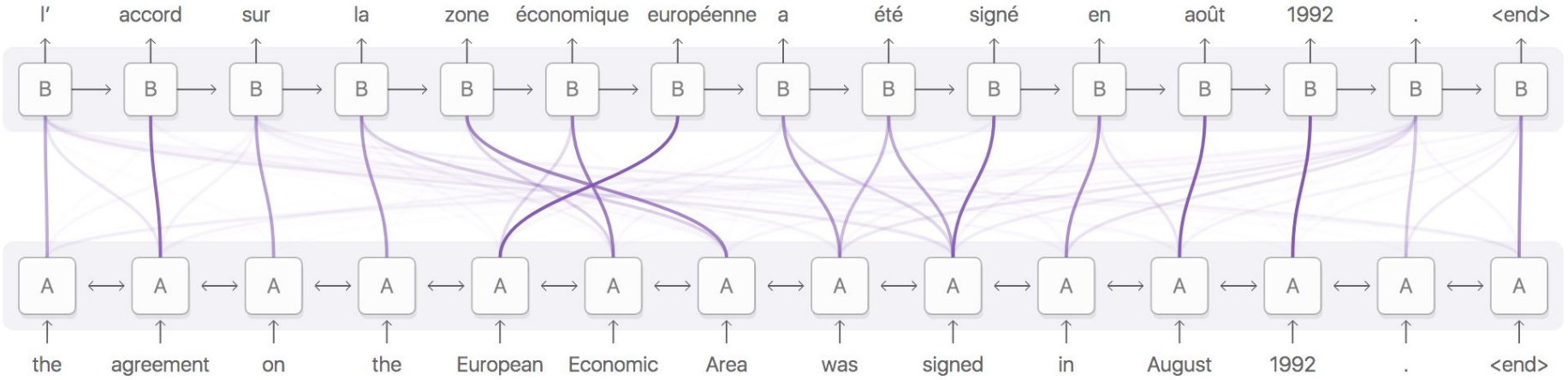
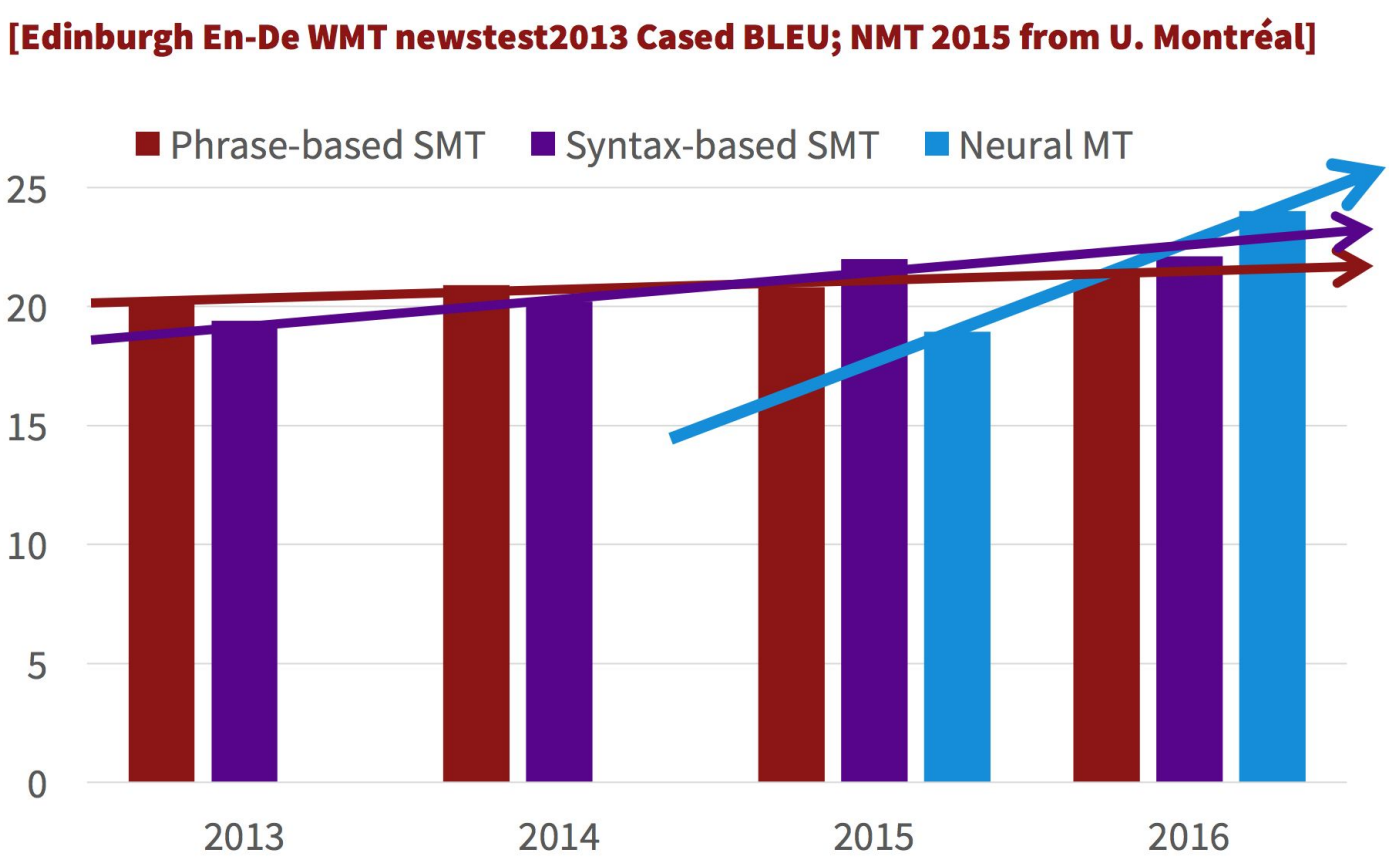


Image copyright © Distill.Pub

MT Performance



Content credit: Stanford CS224n

Attention-Only NMT

Recent paper: SotA NMT using only stacked attention units and positional embeddings (no CNN, RNN etc.).
Very computationally efficient.

Vaswani, Shazeer et al. (Google Brain)
"Attention is All You Need", Jun 2017.
https://arxiv.org/pdf/1706.03762.pdf