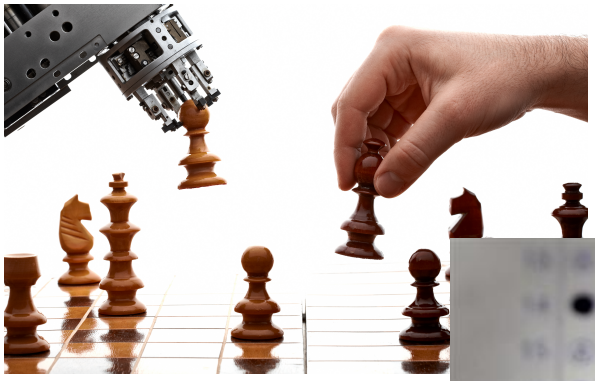
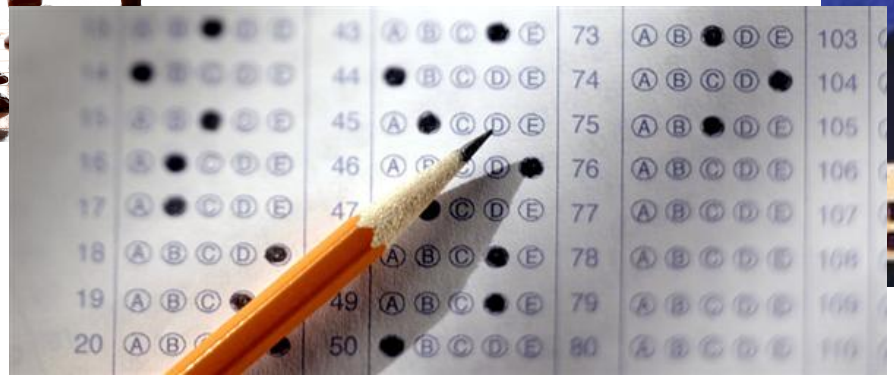
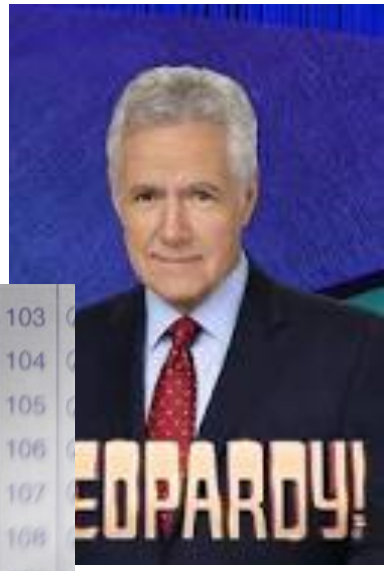
Deep Learning & the Future of Natural Language Processing: Google's New Translation Engine







Bill Vander Lugt J.D. Ph.D.

Michael Jordan (2014):

“I’d use the billion dollars to build a NASA-size program focusing on natural language processing, in all of its glory (semantics, pragmatics, etc.).”

Overview:

1) History of NLP
2) My Experiment:
   LSAT Questions
3) Deep NLP's Architecture

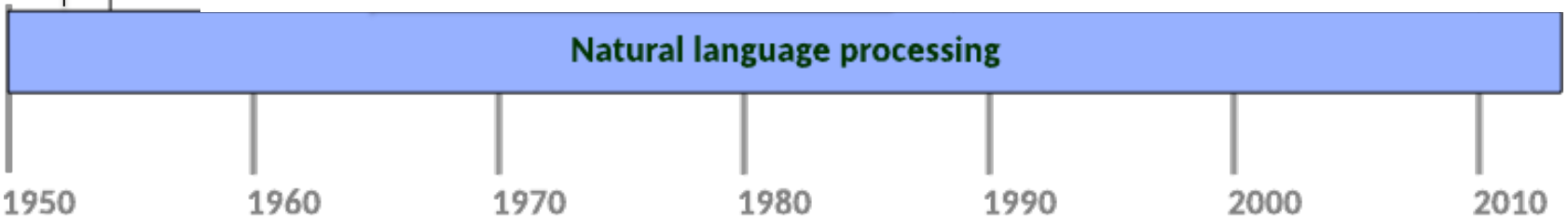
Part 1: History of NLP/ Machine Translation

Does NLP need linguists? Or are texts just data, and words just like numbers?

Are texts just data?

Warren Weaver (to Norbert Wiener) & "Translation" memo

Yes, because words are just another kind of information.



Are texts just data?

Warren Weaver (to Norbert Wiener) & "Translation" memo

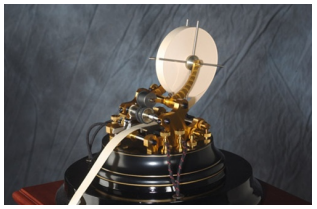
Yes, because words are just another kind of information.

"Recognizing fully, even though necessarily vaguely, the semantic difficulties because of multiple meanings, etc., I have wondered if it were unthinkable to design a computer which would translate…

One … 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**." --1947

1950 1960 010

English > *Morse code* < English
Stock prices > *ticker* < Stock prices
German > *noise* < German
(English > *Russian*) < English



Warren Weaver (to Norbert Wiener) & "Translation" memo

Are texts just data?

Yes, because words are just another kind of information.

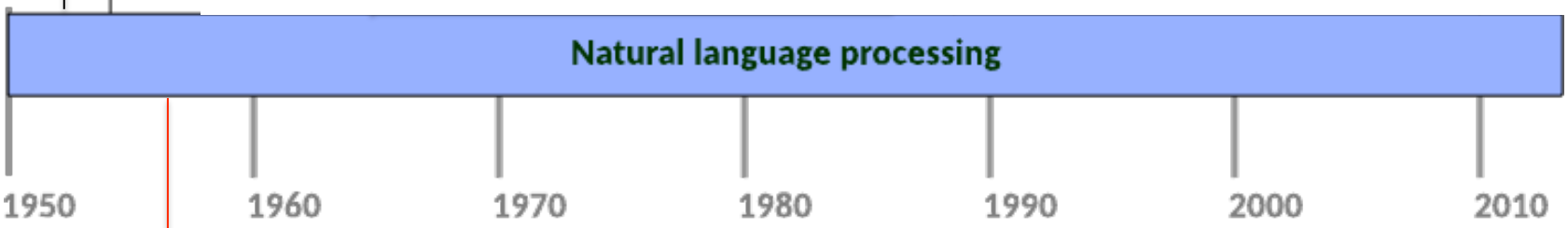
"Recognizing fully, even though necessarily vaguely, the semantic difficulties because of multiple meanings, etc., I have wondered if it were unthinkable to design a computer which would translate…
One … wonders if the problem of translation could conceivably be treated as a problem in cryptography. 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**." --1947

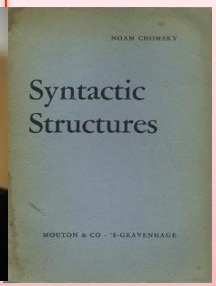
1950 1960 010

Are texts just data?

Warren Weaver (to Norbert Wiener) & "Translation" memo

Yes, because words are just another kind of information.



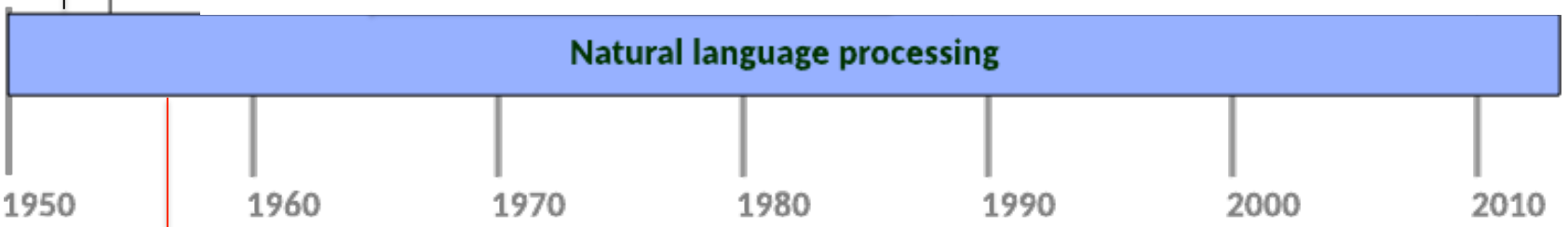


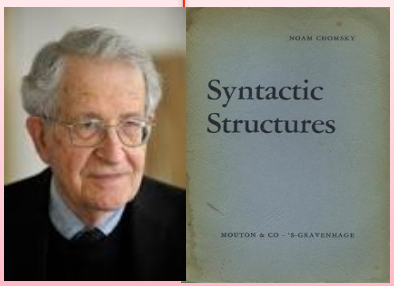
No, because grammar is what makes language intelligible.

Are texts just data?

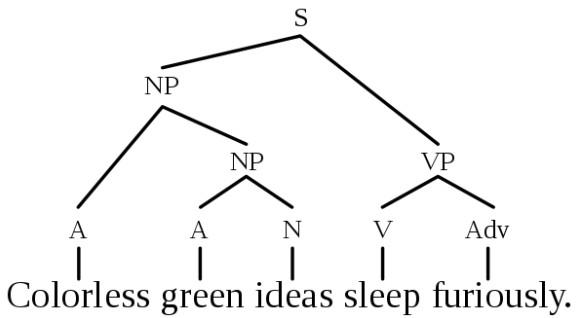
Warren Weaver (to Norbert Wiener) & "Translation" memo

Yes, because words are just another kind of information.





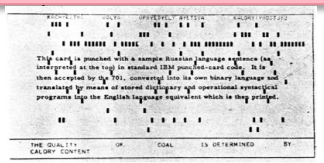
No, because grammar is what makes language intelligible.



Are texts just data?

Warren Weaver (to Norbert Wiener) & "Translation" memo

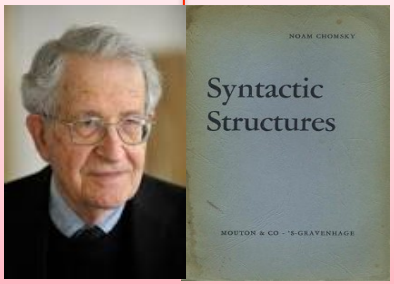
Georgetown Experiment: "Kitty Hawk"
250 words, 60 sentences, 6 syntax rules
Solvable in 3-5 years



Automatic Language Processing Advisory Committee

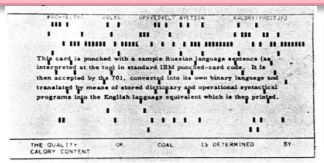
Natural language processing

1950 1960 1970 1980 1990 2000 2010



Warren Weaver (to Norbert Wiener) & "Translation" memo

Georgetown Experiment: "Kitty Hawk"
250 words, 60 sentences, 6 syntax rules
Solvable in 3-5 years



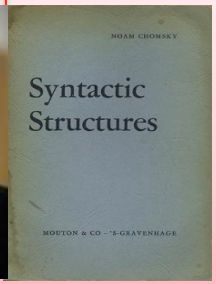
Automatic Language Processing Advisory Committee

IBM's Fred Jelinek:
"Every time I fire a linguist, my system's performance improves."

Are texts just data?

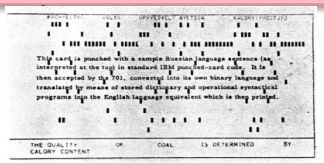
Natural language processing

1950 1960 1970 1980 1990 2000 2010



Warren Weaver (to Norbert Wiener) & "Translation" memo

Georgetown Experiment: "Kitty Hawk"
250 words, 60 sentences, 6 syntax rules
Solvable in 3-5 years



Automatic Language Processing Advisory Committee

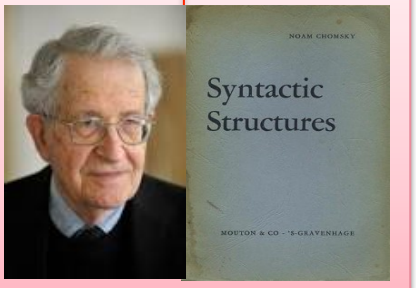
IBM's Fred Jelinek:
"Every time I fire a linguist, my system's performance improves."

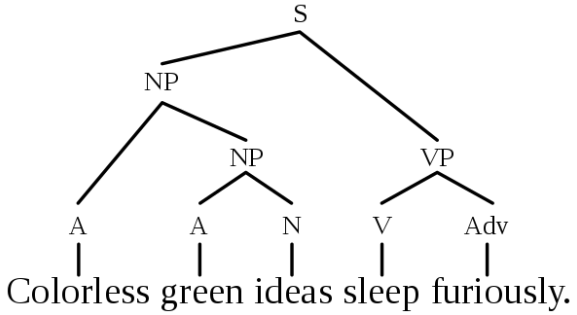
Are texts just data?

IBM's statistical/probabilistic models based on parallel corpora (Parliamentary translations)

Natural language processing

1950 1960 1970 1980 1990 2000 2010





McCulloch & Pitts

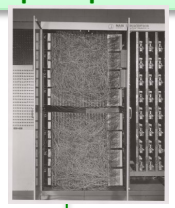
Rosenblatt's perceptron

DEEP LEARNING

Backpropagation

LSTM

Word2Vec

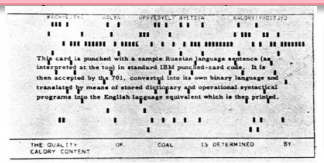


Warren Weaver (to Norbert Wiener) & "Translation" memo

Georgetown Experiment: "Kitty Hawk"
250 words, 60 sentences, 6 syntax rules
Solvable in 3-5 years

IBM's Fred Jelinek:
"Every time I fire a linguist, my system's performance improves."

Are texts just data?

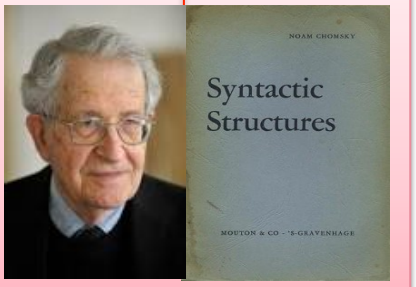


Automatic Language Processing Advisory Committee

IBM's statistical/probabilistic models based on parallel corpora (Parliamentary translations)

Natural language processing

1950 1960 1970 1980 1990 2000 2010



McCulloch & Pitts

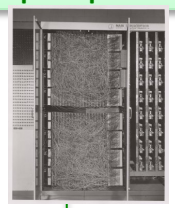
Rosenblatt's perceptron

DEEP LEARNING

Backpropagation

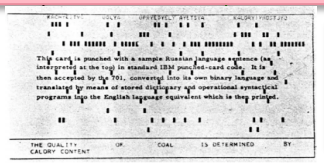
LSTM

Word2Vec



Warren Weaver (to Norbert Wiener) & "Translation" memo

Georgetown Experiment: "Kitty Hawk"
250 words, 60 sentences, 6 syntax rules
Solvable in 3-5 years



Automatic Language Processing Advisory Committee

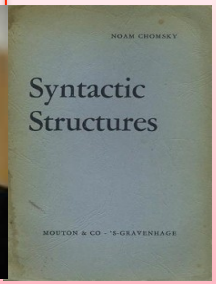
IBM's Fred Jelinek:
"Every time I fire a linguist, my system's performance improves."

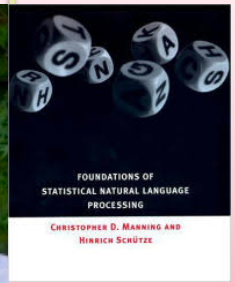
Are texts just data?

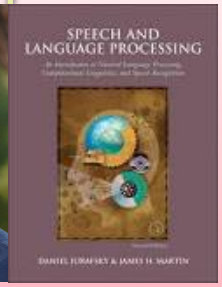
IBM's statistical/probabilistic models based on parallel corpora (Parliamentary translations)

Natural language processing

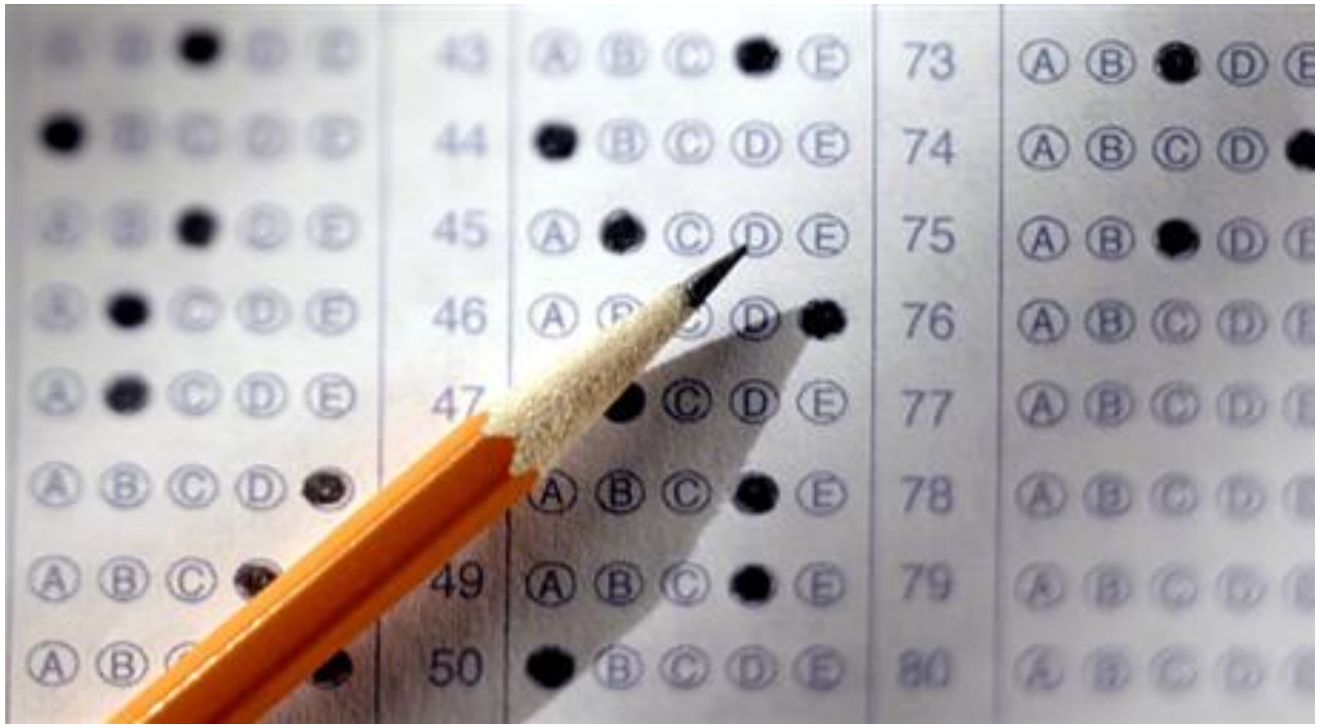
1950 1960 1970 1980 1990 2000 2010







Part 2: My LSAT Experiment



LSAT Logic Game

Questions 1–5

A company employee generates a series of five-digit product codes in accordance with the following rules:

The codes use the digits 0, 1, 2, 3, and 4, and no others.
Each digit occurs exactly once in any code.
The second digit has a value exactly twice that of the first digit.
The value of the third digit is less than the value of the fifth digit.

1. If the last digit of an acceptable product code is 1, it must be true that the

(A) first digit is 2
(B) second digit is 0
(C) third digit is 3
(D) fourth digit is 4
(E) fourth digit is 0

Why LSAT Logic Games?

Solving Logic Puzzles: From Robust Processing to Precise Semantics

Iddo Lev,* Bill MacCartney,* Christopher D. Manning,*† and Roger Levy†

* Department of Computer Science
Stanford University
Stanford, CA 94305-9040, USA
{*iddolev|wcmac|manning*}*@cs.stanford.edu*

† Department of Linguistics
Stanford University
Stanford, CA 94305-2150, USA
rog@stanford.edu

1) Humans are strong readers but weak calculators; computers are weak readers but strong calculators.
2) ordinary and diverse language
3) require Natural Language Understanding and successful inferences
4) clear evaluation metric (multiple choice test)

How to Solve LSAT Logic Games

- ❑ Identify Type
- ❑ Set Up
- ❑ Parse Rules
- ❑ Apply Rules
- ❑ Parse Question
- ❑ Parse Answers
- ❑ Pick Correct Answer

How to Solve LSAT Logic Games

- ✓ Identify Type
- ✓ Set Up
- ❑ Parse Rules
- ✓ Apply Rules
- ✓ Parse Question
- ✓ Parse Answers
- ✓ Pick Correct Answer

How to Solve LSAT Logic Games

- ✓ Identify Game Type
- ✓ Set Up
- ❑ Parse Rules
- ✓ Apply Rules
- ✓ Parse Question
- ✓ Parse Answers
- ✓ Pick Correct Answer

NLP Landscape

Dan Jurafsky

MOSTLY SOLVED

Spam Detection
Part of Speech Tagging
Named Entity Recognition

MAKING GOOD PROGRESS

Sentiment Analysis
Co-reference Resolution
Word Sense Disambiguation
Syntactic Parsing
Machine Translation
Information Extraction

STILL REALLY HARD

Question Answering
Paraphrase
Summarization
Dialog

NLP Landscape

Dan Jurafsky

MOSTLY SOLVED

Spam Detection

Part of Speech Tagging

Named Entity Recognition

MAKING GOOD PROGRESS

Sentiment Analysis

Co-reference Resolution

Word Sense Disambiguation

Syntactic Parsing

Machine Translation

Information Extraction

STILL REALLY HARD

Question Answering

Paraphrase

Summarization

Dialog

NLP Landscape

Dan Jurafsky

MOSTLY SOLVED

Spam Detection
Part of Speech Tagging
Named Entity Recognition

MAKING GOOD PROGRESS

Sentiment Analysis
Co-reference Resolution
Word Sense Disambiguation
Syntactic Parsing
Machine Translation
Information Extraction

STILL REALLY HARD

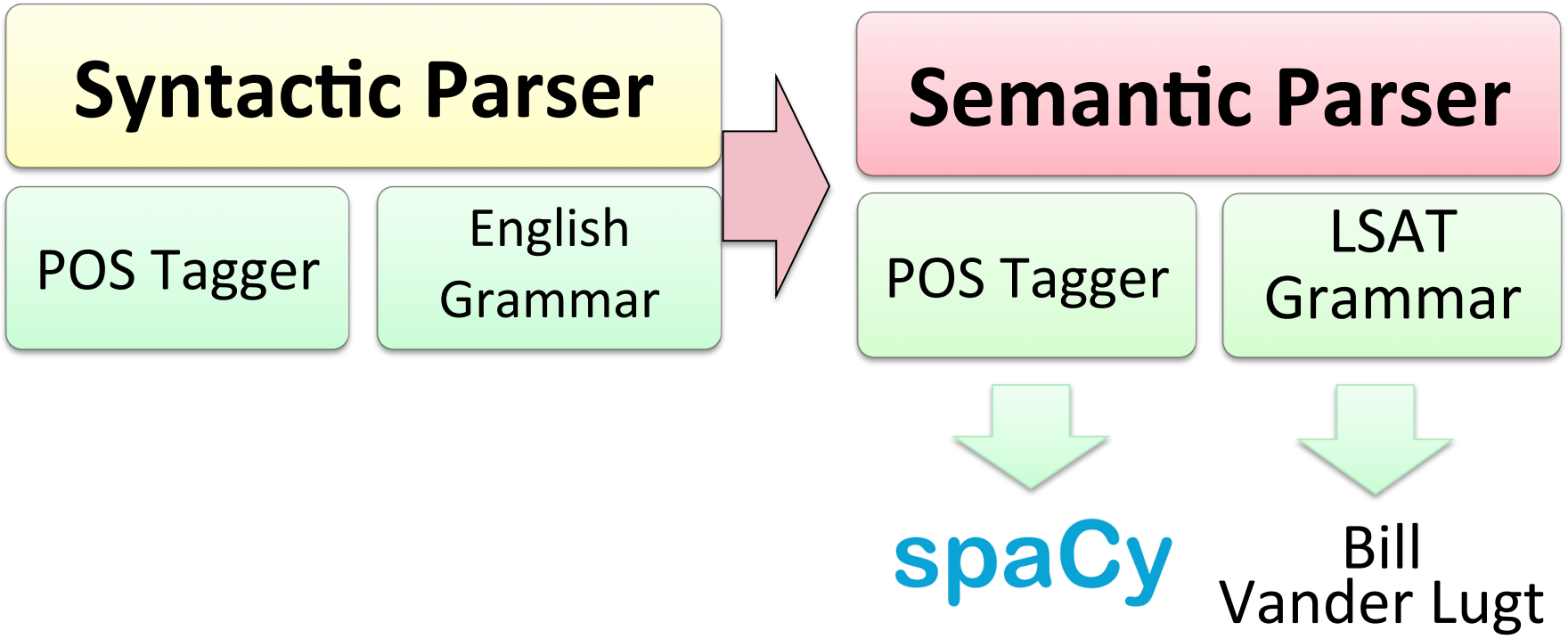
Question Answering
Paraphrase
Summarization
Dialog

Model #1: Parser

Model #2: Translator

Model #1: Parser

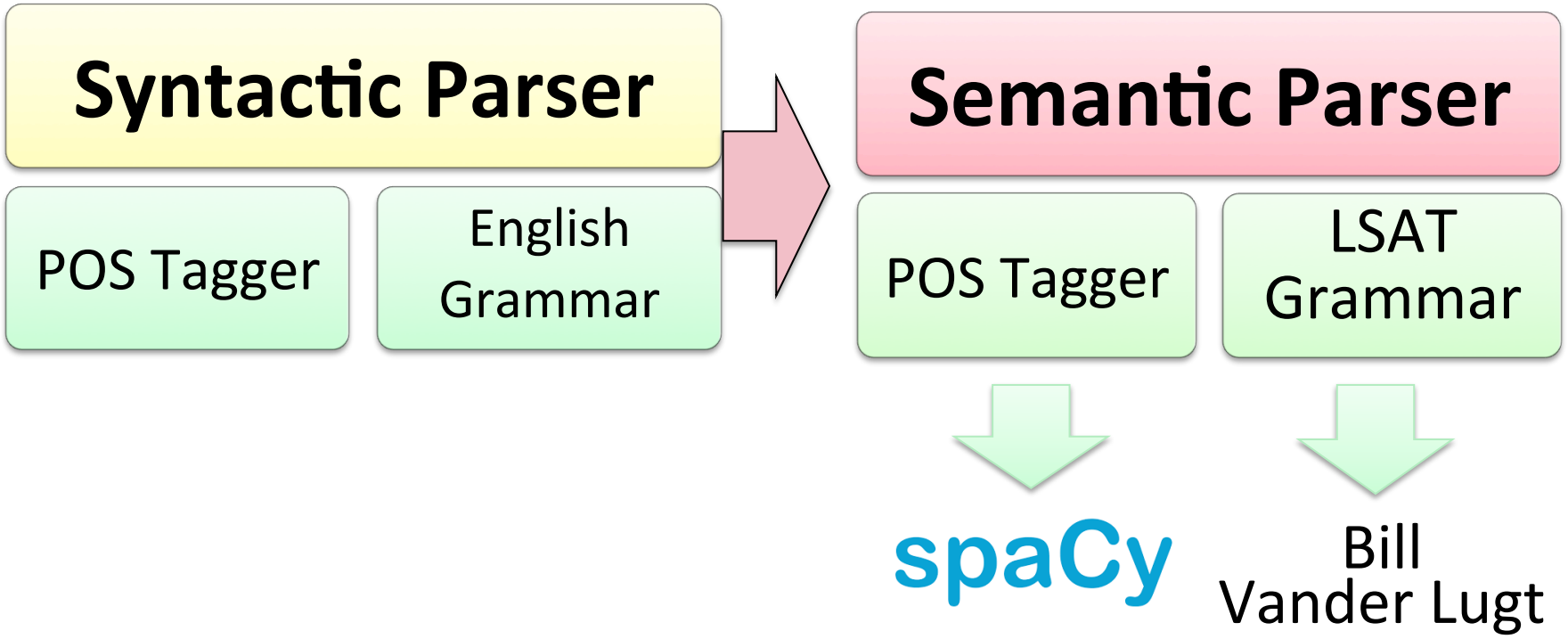
Stanford / Google (2016)



Model #1: Parser

Stanford / Google (2016)

Parsey McParseface



Model #1: Parser
Grammar-Based
Limited Machine Learning

PROS

- Understandable
- Builds on prior knowledge
- Context-Free Grammar infinitely complex
- Solved entire puzzle

CONS

- Fragile
- Faltered on new puzzles

Model #2: Translator
Google's Seq2Seq

(April 11, 2017)

SOURCE Sequence: Das ist mein Haus.

TARGET Sequence: That is my house.

Model #2: Translator

SOURCE language: English

Jennifer lectures first if, but only if,

Reese lectures before Anika.

TARGET language: Python

((if (B<C): (A==0)) and (if (not(B<C)):(not(A==0))))

Model #2: Promising Results Even with a Miniscule Data Set

Training set: 138 examples

Test set: 15 examples

Accuracy: 11-13 of 15 correct (7 runs)

Model #2: Promising Results Even with a Miniscule Data Set

Training set: 138 examples

Test set: 15 examples

Accuracy: 11-13 of 15 correct (7 runs)

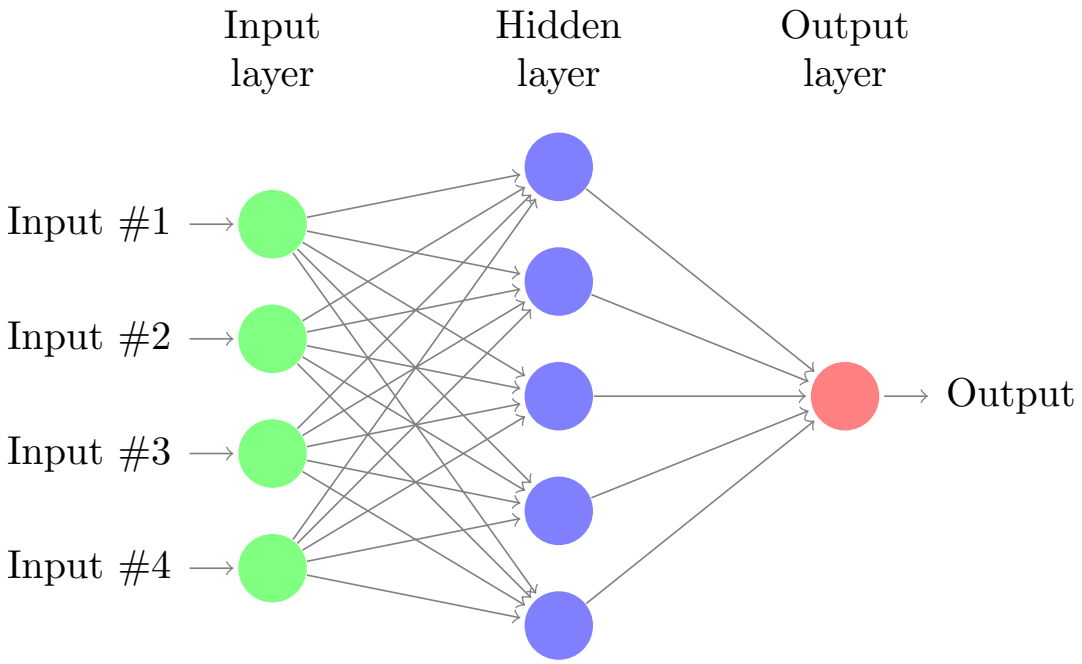
Seq2Seq: (((B-A)==x) or ((A-B)==x))

MY LABEL: (abs(A-B)==x)

Part 3: Architecture of Seq2Seq

Model

Neural Net

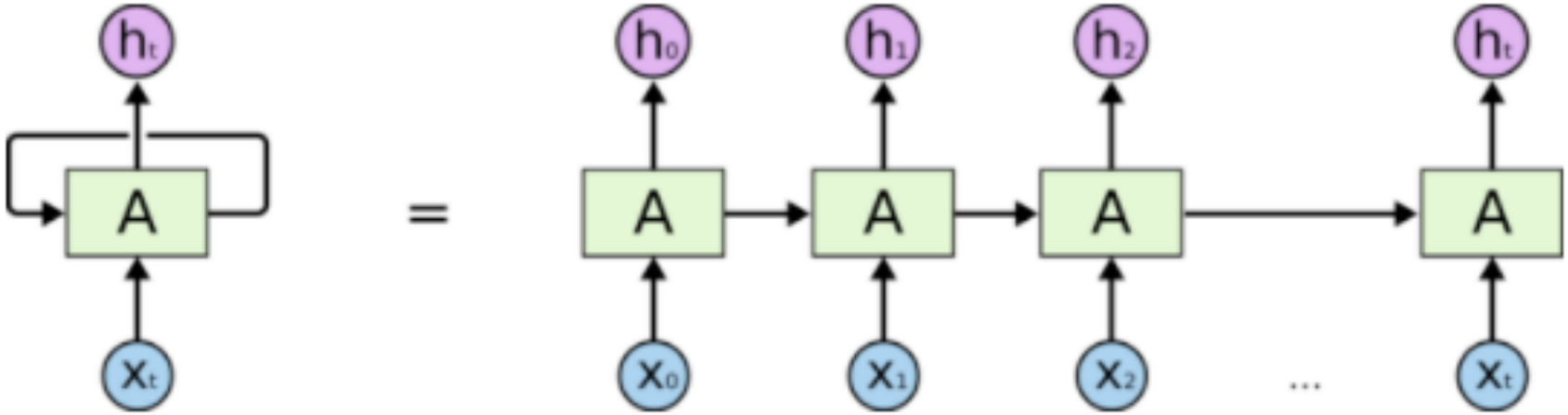


Architecture of Seq2Seq

Model	Problem
Neural Net	can’t do sequences

Architecture of Seq2Seq

Model	Problem
Neural Net	can't do sequences
Recurrent NN	



An unrolled recurrent neural network.

Illustration by Chris Olah

Architecture of Seq2Seq

Model	Problem
Neural Net	can't do sequences
Recurrent NN	reading v. writing
Encoder/Decoder	

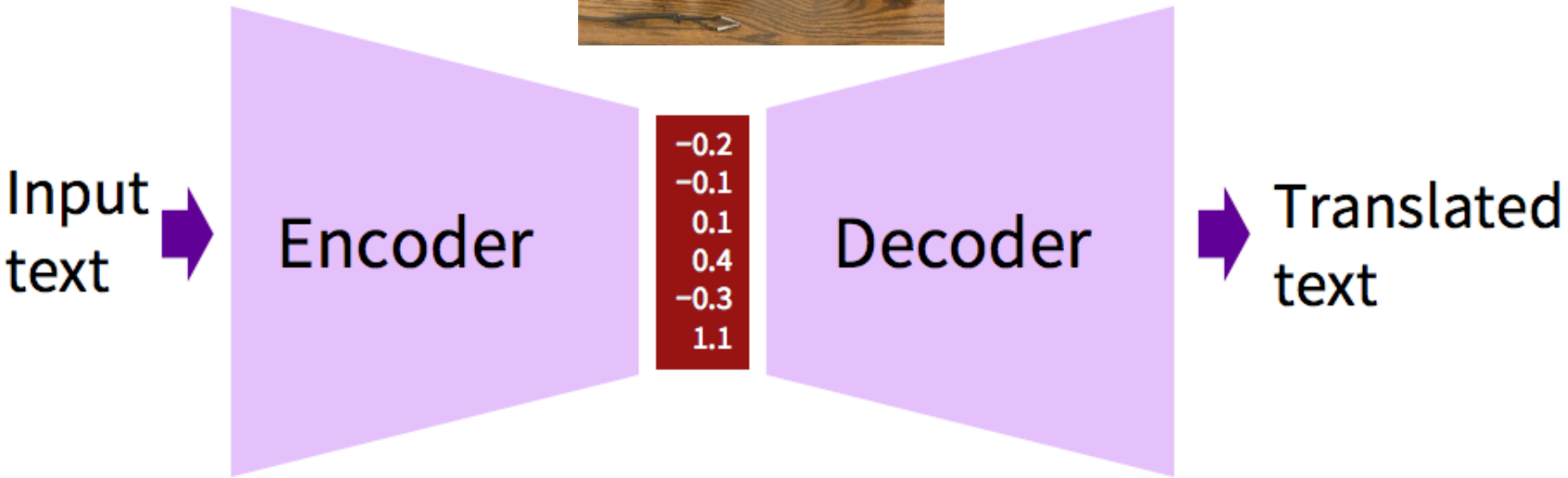


Illustration by Chris Manning

Architecture of Seq2Seq

Model	Problem
Neural Net	can't do sequences
Recurrent NN	reading v. writing
Encoder/Decoder	

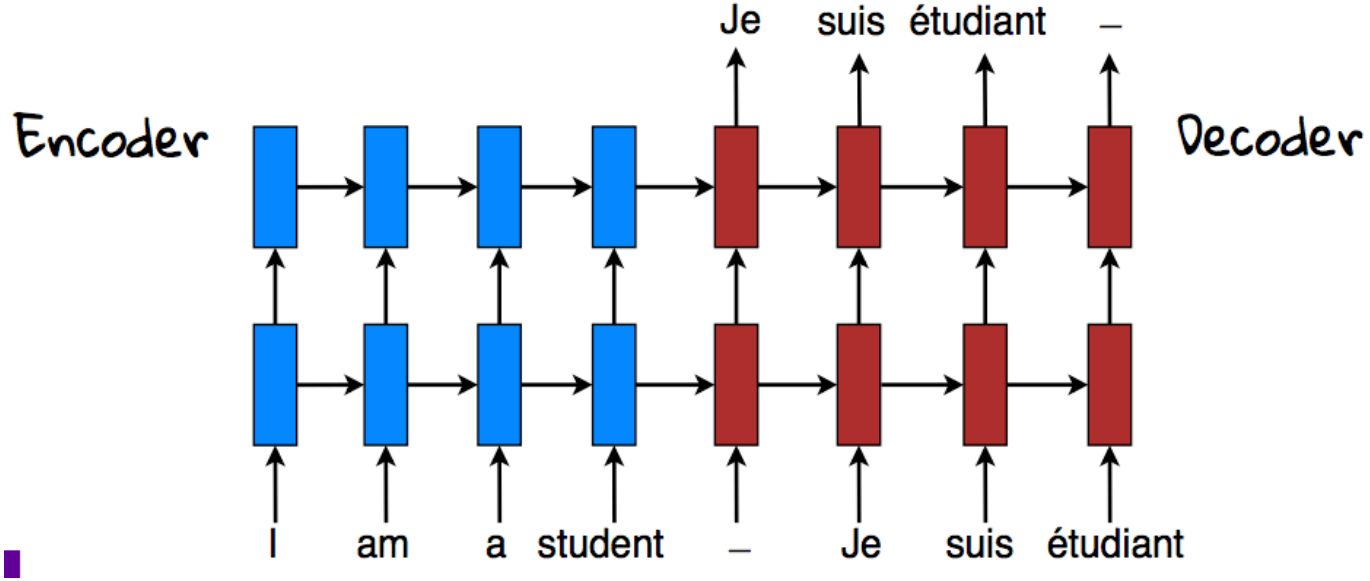
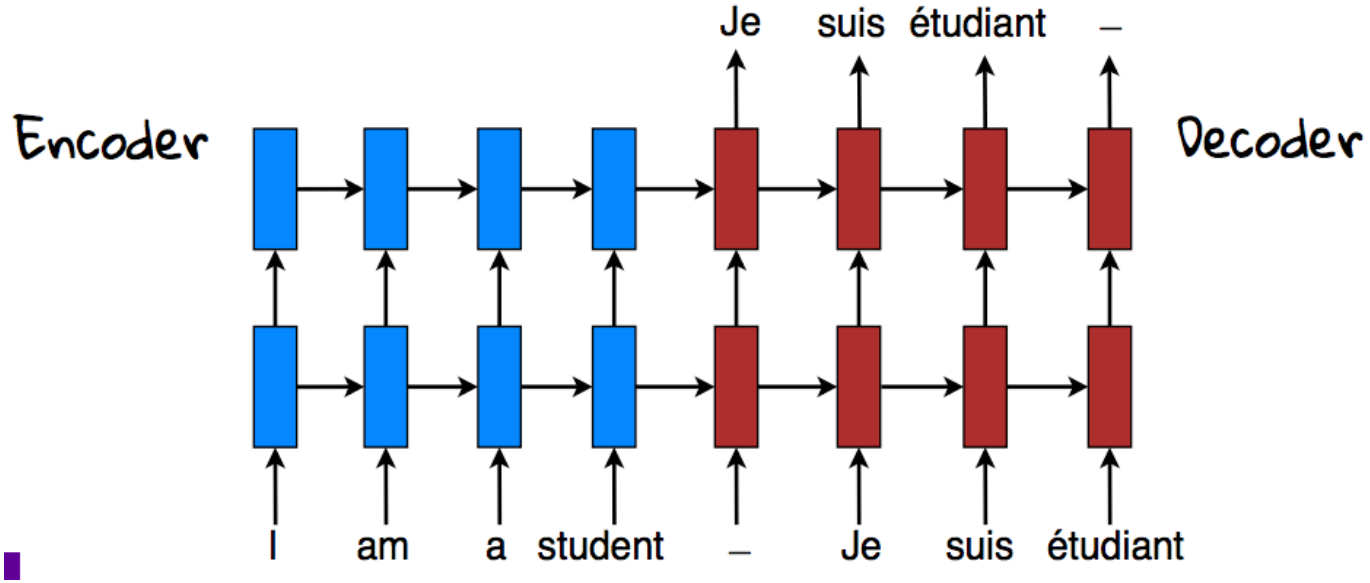


Illustration by Chris Manning

Architecture of Seq2Seq

Model	Problem
Neural Net	can't do sequences
Recurrent NN	reading v. writing
Encoder/Decoder	forgets after 7-ish words



Architecture of Seq2Seq

Model	Problem
Neural Net	can't do sequences
Recurrent NN	reading v. writing
Encoder/Decoder	forgets after 7-ish words
LSTM/GRU	

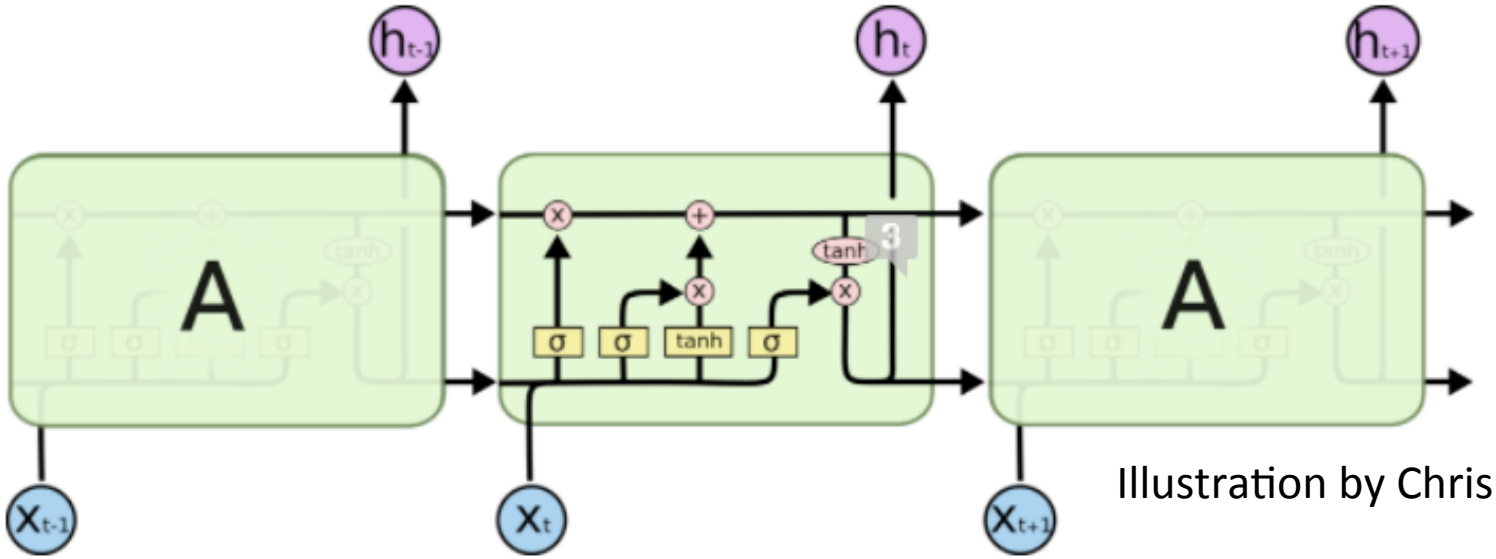


Illustration by Chris Olah

Architecture of Seq2Seq

Model	Problem
Neural Net	can't do sequences
Recurrent NN	reading v. writing
Encoder/Decoder	forgets after 7-ish words
LSTM/GRU	can't look ahead

Between the banks flows _____

water?

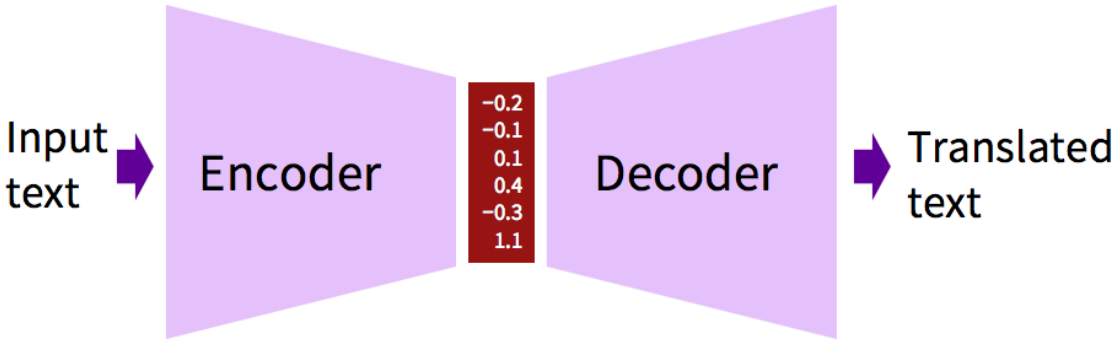
money?

Architecture of Seq2Seq

Model	Problem
Neural Net	can't do sequences
Recurrent NN	reading v. writing
Encoder/Decoder	forgets after ~7 words
LSTM/GRU	can't look ahead
Bidirectional encoding	

Architecture of Seq2Seq

Model	Problem
Neural Net	can't do sequences
Recurrent NN	reading v. writing
Encoder/Decoder	forgets after 7-ish words
LSTM/GRU	can't look ahead
Bidirectional	single, fixed-length vector



Architecture of Seq2Seq

Model	Problem
Neural Net	can't do sequences
Recurrent NN	reading v. writing
Encoder/Decoder	forgets after 7-ish words
LSTM/GRU	can't look ahead
Bidirectional	single, fixed-length vector
Attention	

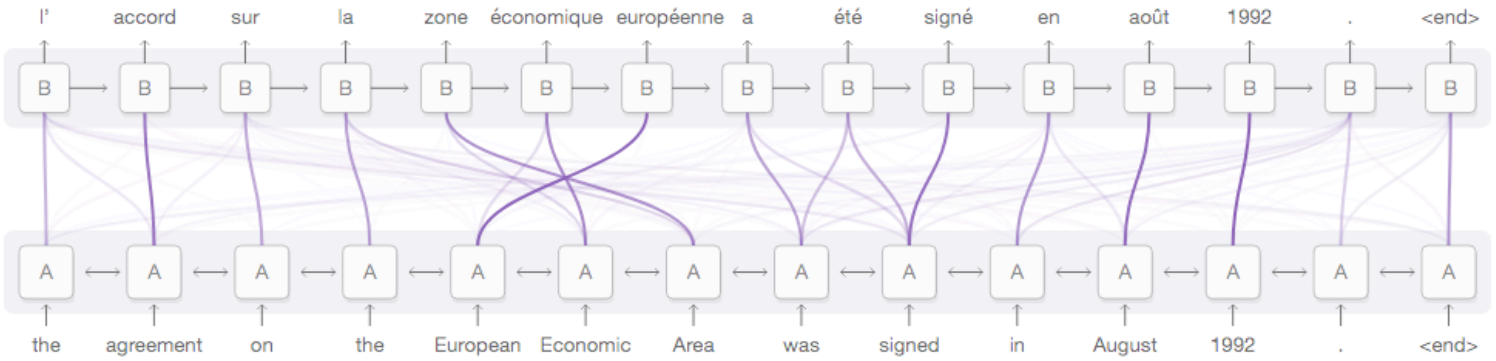
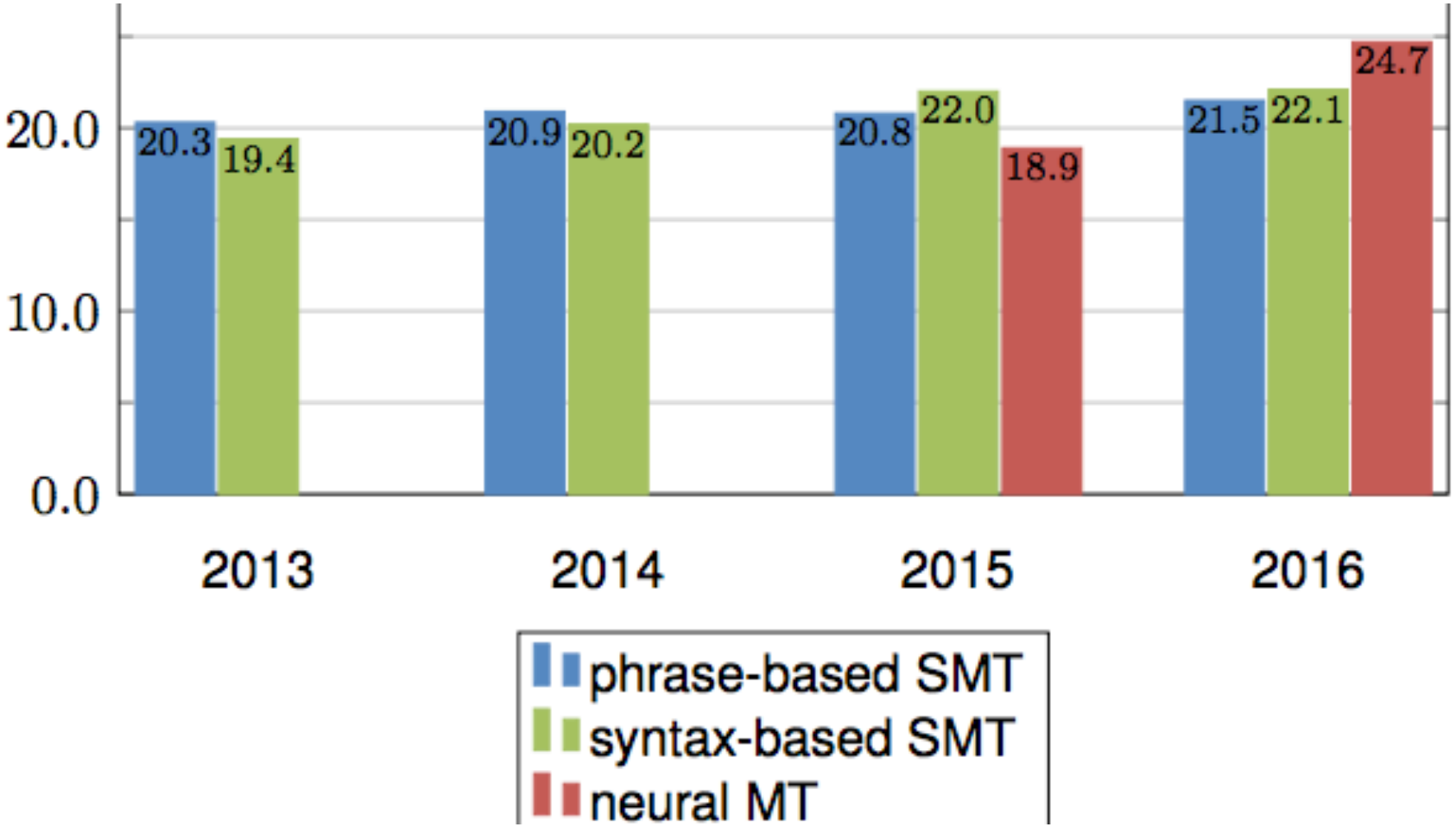


Illustration by Chris Olah

Workshop on Machine Translation (WMT)



Graph by Rico Sennrich

“We are not yet rid of God because we still have faith in grammar.”

—Friedrich Nietzsche

"We are not yet rid of God because
we still have faith in grammar."
—Friedrich Nietzsche

Is Grammar/Linguistics Dead?

"We are not yet rid of God because
we still have faith in grammar."
—Friedrich Nietzsche

Is Grammar/Linguistics Dead?


Chris Manning
Dan Jurafsky
Linguistics & Computer Science

"We are not yet rid of God because
we still have faith in grammar."
—Friedrich Nietzsche

Is Grammar/Linguistics Dead?


Chris Manning
Dan Jurafsky
Linguistics & Computer Science

Google: "No"

SyntaxNet: grammar-based
syntactic parsers for 40 languages.

BILL VANDER LUGT Ph.D. J.D.

DATA SCIENTIST

billvanderlugt@gmail.com

Denver, CO

billvanderlugt

billvanderlugt/LSAT

billvanderlugt.com

LSAT's Context-Free Grammar

LSAT Rule_1 → Boolean (True/False)

LSAT Rule_2 → If Boolean: Boolean

Boolean_1 → Set Compared Set

Boolean_2 → Boolean Conjunction Boolean

Set_1 → Variable(s)

Set_2 → Set Conjunction Set

Conjunction → and, or, xor

Compared_1 → >, <, =, -

Compared_2 → not (Compared)

Model #2: Translator

PREPROCESSING STEPS:

Jennifer lectures first if, but only if,
Reese lectures before Anika.

Jennifer lectures 0 if, but only if, Reese lectures before Anika.

A ~~lectures~~ 0 if, but only if, B ~~lectures~~ before C.

A 0 if but only if B before C

TARGET SEQUENCE:

((if (B<C): (A==0)) and (if (not(B<C)):(not(A==0))))

Seq2Seq Learning Step by Step…

A 0 if but only if B before C

Step 201: (((A)))))))))))))))))))))))))))))))

Seq2Seq Learning Step by Step…

A 0 if but only if B before C

Step 201: (((A)))))))))))))))))))))))))))))))

Step 401 ((AAA<B))(A>B))SEQUENCE_END

(A(A(A)B))SEQUENCE_END

(A(A>B))SEQUENCE_END

SEQUENCE_END

SEQUENCE_END

Seq2Seq Learning Step by Step…

A 0 if but only if B before C

Step 201: (((A)))))))))))))))))))))))))))))))

Step 401 ((AAA<B))(A>B))SEQUENCE_END

(A(A(A)B))SEQUENCE_END

(A(A>B))SEQUENCE_END

SEQUENCE_END

SEQUENCE_END

Step 1101 ((((A-B))(A==0))SEQUENCE_END

Step 1201 (abs((B-B))(A==0))SEQUENCE_END

Seq2Seq Learning Step by Step…

A 1 if but only if B before C

Step 1801 (absabs(B and C))(A==0))SEQUENCE_END
(if (A(A-C)) and (not(A==0))))SEQUENCE_END

Step 2001 (absabs(B-B) and (B==0)) and (if (abs(A-B)))(not(A==0))))SEQUENCE_END

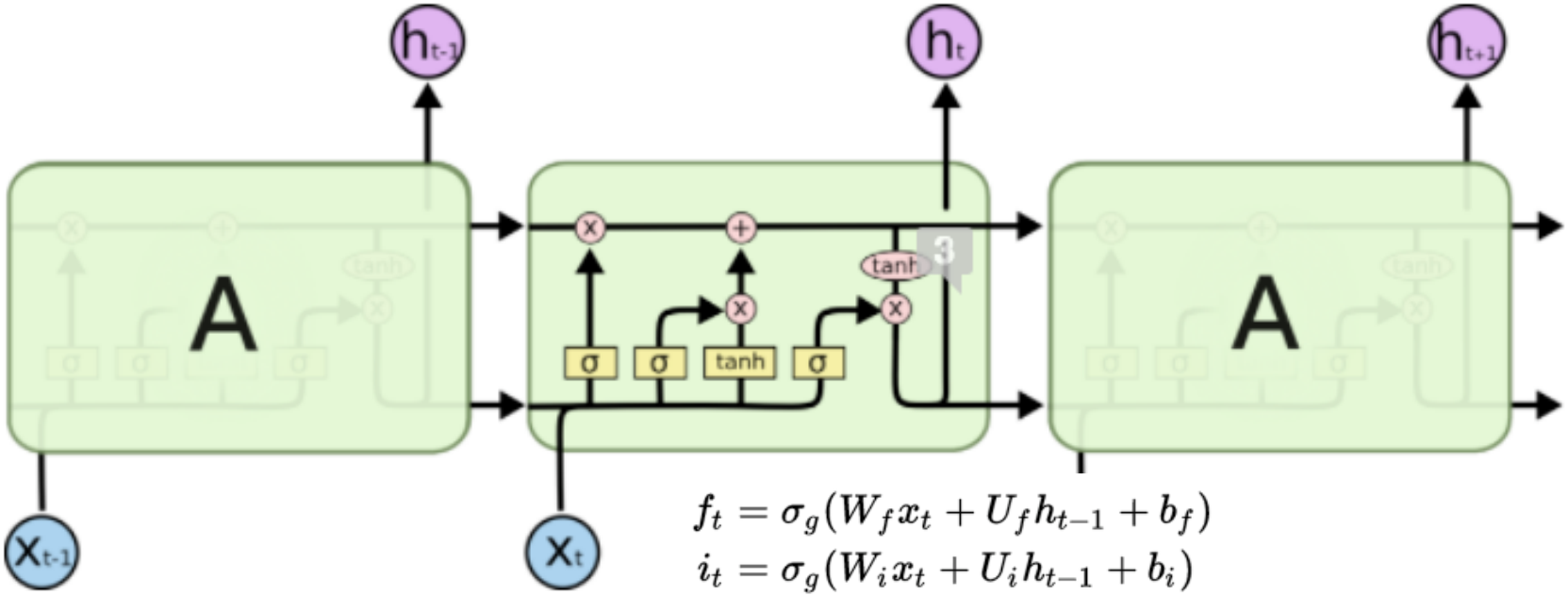
Step 2201 (abs((B-C))(A==0)) and (if (abs(A-C))==(not(A==0))))SEQUENCE_END

Step 2601 ((((B<C))(A==0)) and (if (A(B<C)) and (not(A==0))))SEQUENCE_END

Step 2701 ((((B<C) and (A==0)) and (if (not(B<C)): (not(A==0))))SEQUENCE_END

Step 3301 ((if (B<C): (A==0)) and (if (not(B<C)): (not(A==0))))SEQUENCE_END

CORRECT: ((if (B<C): (A==0)) and (if (not(B<C)): (not(A==0))))SEQUENCE_END

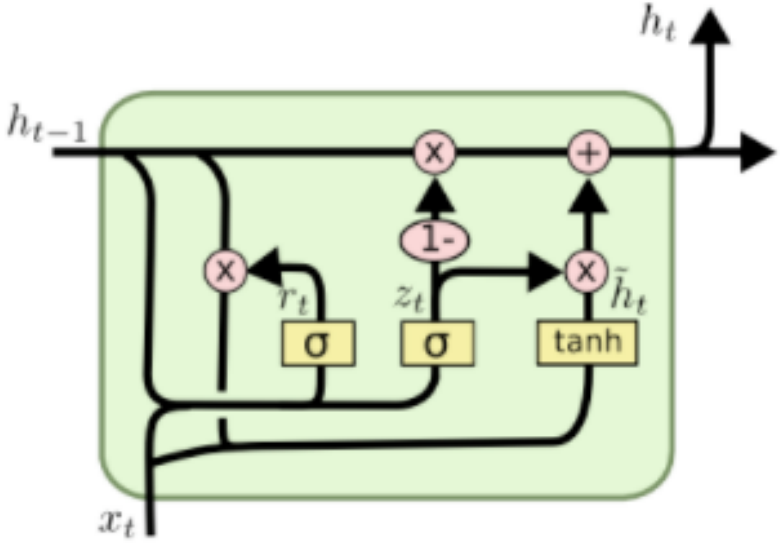


$$
\begin{aligned}
f_t &= \sigma_g(W_f x_t + U_f h_{t-1} + b_f) \\
i_t &= \sigma_g(W_i x_t + U_i h_{t-1} + b_i) \\
o_t &= \sigma_g(W_o x_t + U_o h_{t-1} + b_o) \\
c_t &= f_t \circ c_{t-1} + i_t \circ \sigma_c(W_c x_t + U_c h_{t-1} + b_c) \\
h_t &= o_t \circ \sigma_h(c_t)
\end{aligned}
$$

Variables

- x_t: input vector
- h_t: output vector
- c_t: cell state vector
- W, U and b: parameter matrices and vector
- f_t, i_t and o_t: gate vectors
  - f_t: Forget gate vector. Weight of remembering old information.
  - i_t: Input gate vector. Weight of acquiring new information.
  - o_t: Output gate vector. Output candidate.

Gated Recurrent Unit (GRU)



$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$

$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$

$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

model:AttentionSeq2Seq

encoder.class: seq2seq.encoders.BidirectionalRNNEncoder
cell_class: GRUCell
cell_params:
num_units: 128
dropout_input_keep_prob: 0.8
dropout_output_keep_prob: 1.0
num_layers: 1
decoder.class: seq2seq.decoders.AttentionDecoder
decoder.params:
rnn_cell:
cell_class: GRUCell
cell_params:
num_units: 128
dropout_input_keep_prob: 0.8
dropout_output_keep_prob: 1.0
num_layers: 1
optimizer.name: Adam
optimizer.params:
epsilon: 0.0000008
optimizer.learning_rate: 0.0001
source.max_seq_len: 17
source.reverse: false
target.max_seq_len: 42