Machine Reading, Question Answering & Dialog

AMMI – Deep NLP

Angela Fan, Louis Martin, Antoine Bordes
FAIR – Meta AI

May 29 – June 3, 2022

Who are we?

Angela
angelafan@fb.com
Research Scientist
Meta AI - NYC

Louis
louismartin@fb.com
Research Scientist
Meta AI - London

Antoine
abordes@fb.com
Research Director
Meta AI - Paris

This Class:

- Machine Reading with deep learning
- Open-domain Question answering
- Deep learning for dialogue

Quick schedule (Dakar Time)

- Monday: 9am Lecture [Antoine] + 2pm Q&A [All]
- Tuesday: 9am Lecture [Angela] + 2pm Q&A [All] + 3pm Labs [Louis]
- Wednesday: 9am Lecture [Antoine] + 2pm Q&A [All] + 3pm Labs [Louis]
- Thursday: 10am Career panel [All]
- Friday: 10am Quizz [Louis]

ROBOTS CAN NOW READ BETTER THAN HUMANS, PUTTING MILLIONS OF JOBS AT RISK

BY **ANTHONY CUTHBERTSON** ON 1/15/18 AT 8:00 AM

ROBOTS CAN NOW PATTERN MATCH ON A BENCHMARK DATASET BETTER THAN HUMANS

BY **ANTHONY CUTHBERTSON** ON 1/15/18 AT 8:00 AM

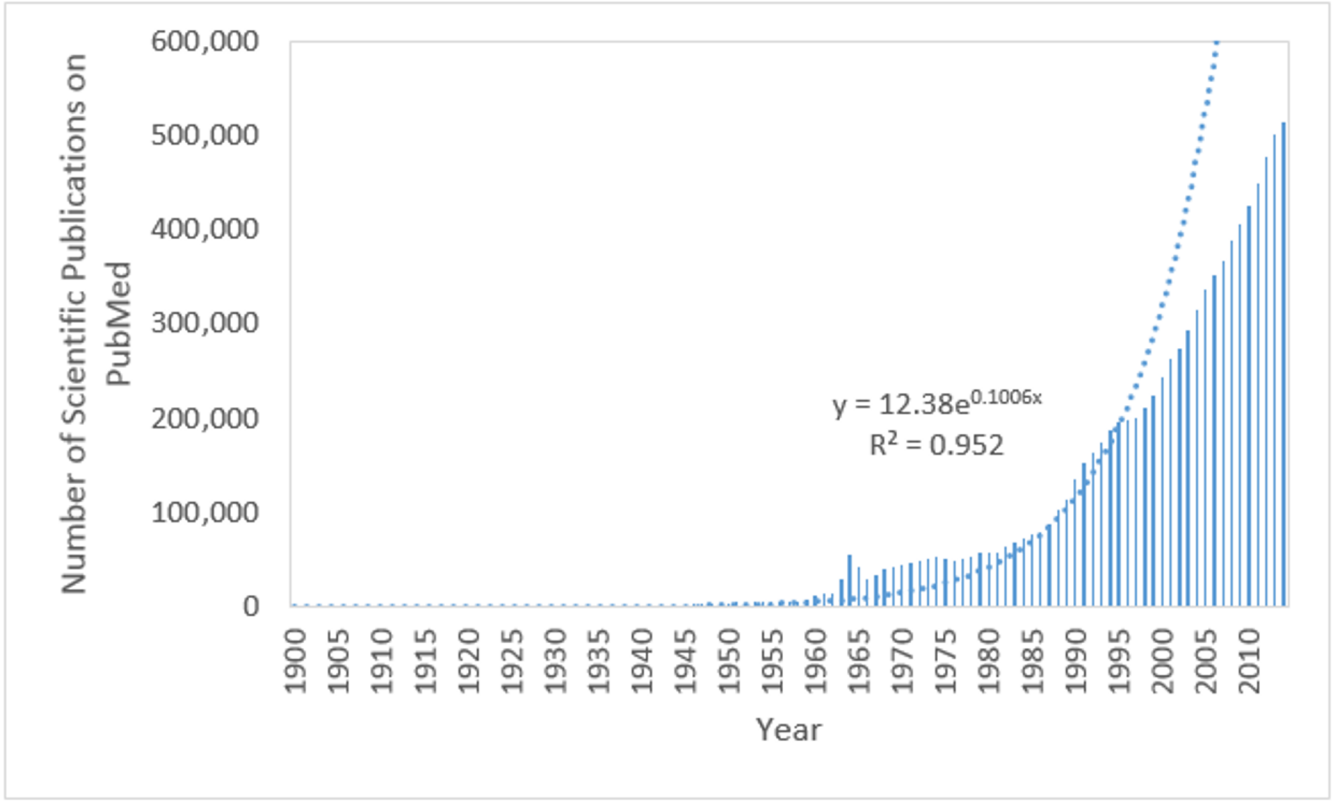
BUT THERE HAS BEEN A LOT OF PROGRESS AND MACHINE READING RESEARCH ACTIVITY HAS SKYROCKETED

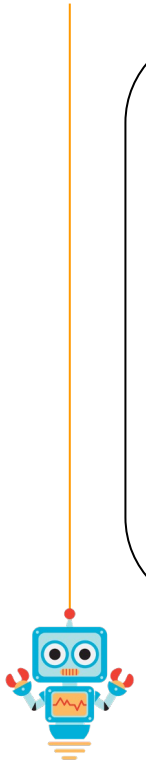
BY **ANTHONY CUTHBERTSON** ON 1/15/18 AT 8:00 AM

a Ăīŧ [illegible]

Machines processing **text** to satisfy an **information need** is long standing goal of AI

Motivation 1: Information Overload



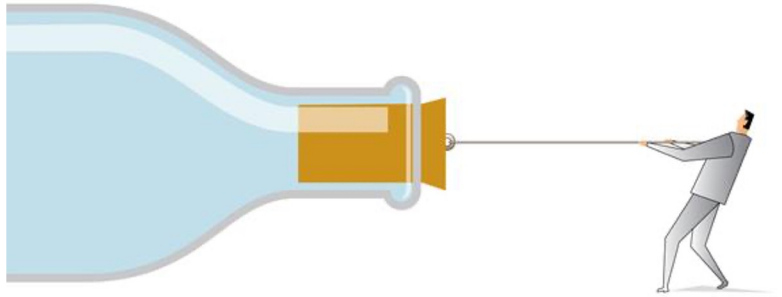


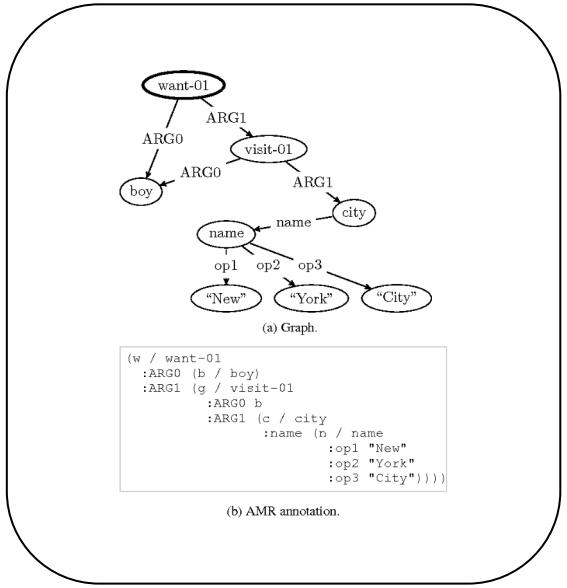
[Information Need]

uses for

Motivation 2: The Knowledge Acquisition Bottleneck

"The problem of knowledge acquisition is the critical bottleneck problem in artificial intelligence."
E. A. Feigenbaum 1984





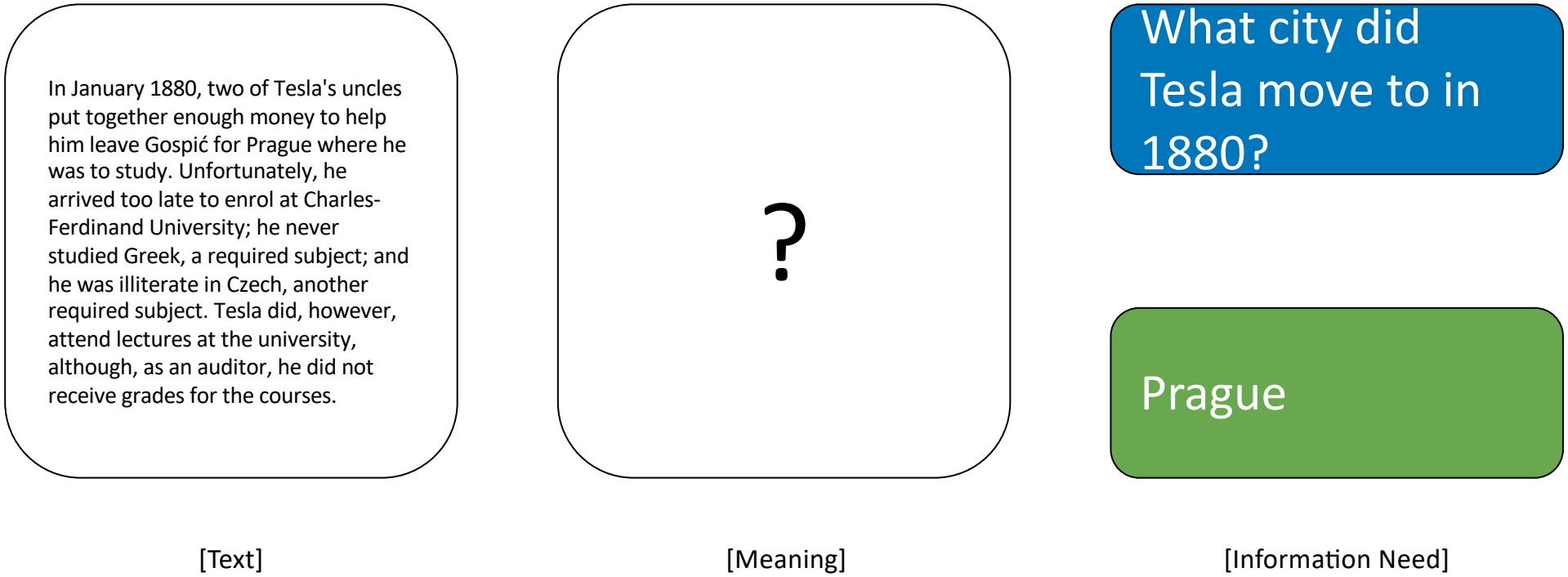
[Knowledge]

[Meaning]

uses for

[Information Need]

Applications: Question Answering



Applications: Support a Molecular Tumor Board

Poon et al., ACL'17

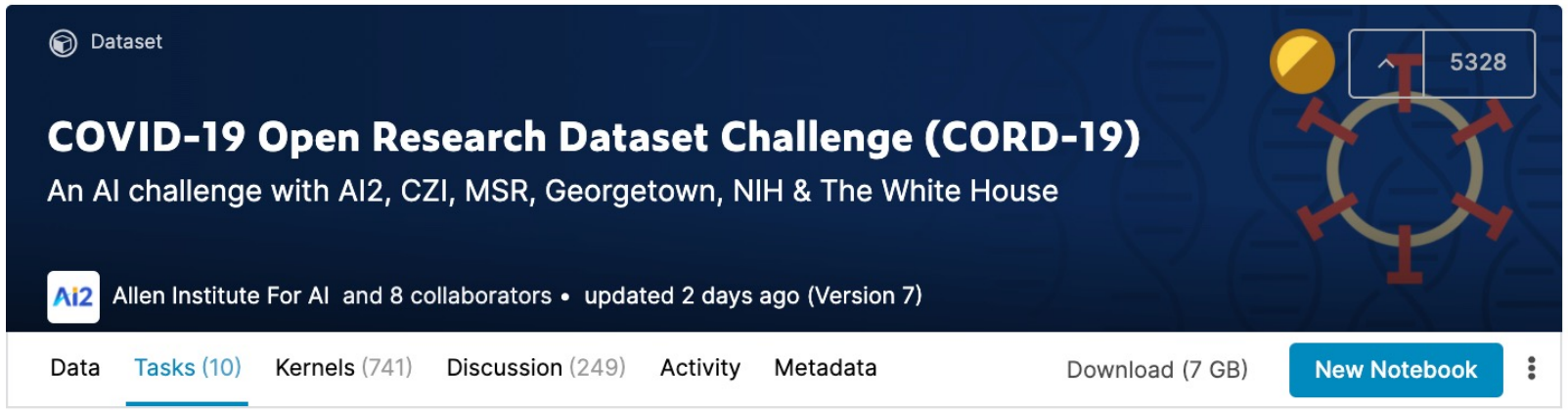
The deletion mutation on exon-19 of EGFR gene was present in 16 patients, while the L858E point mutation on exon-21 was noted in 10. All patients were treated with gefitinib and showed a partial response.

[Text]

?

[Meaning]

[Information Need]



Tasks

What is known about transmission, incubation, and environmental stability? 998
COVID-19 Open Research Dataset Challenge (CORD-19)
Paul Mooney · 84 Submissions

What do we know about COVID-19 risk factors? 318
COVID-19 Open Research Dataset Challenge (CORD-19)
Paul Mooney · 71 Submissions

What do we know about virus genetics, origin, and evolution? 161
COVID-19 Open Research Dataset Challenge (CORD-19)
Paul Mooney · 55 Submissions

What do we know about vaccines and therapeutics? 136
COVID-19 Open Research Dataset Challenge (CORD-19)
Paul Mooney · 53 Submissions

What has been published about medical care? 114
COVID-19 Open Research Dataset Challenge (CORD-19)
Paul Mooney · 42 Submissions

https://www.kaggle.com/allen-institute-for-ai/COR research-challenge/tasks

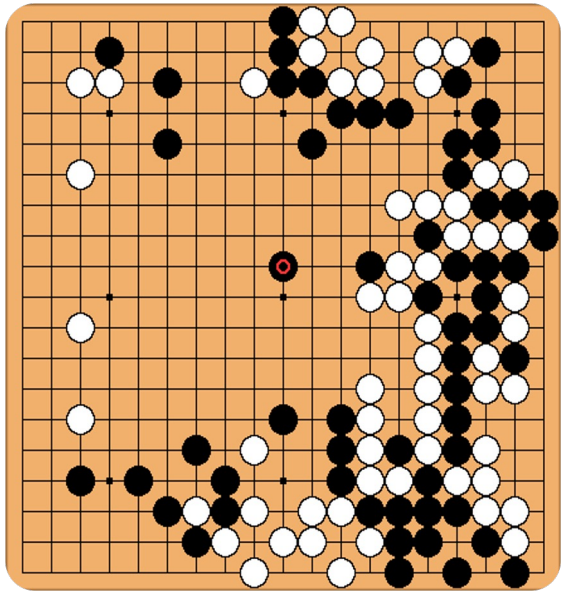
Applications: Helping Agents to learn Faster

A fundamental Go strategy involves keeping stones connected. Connecting a group with one eye to another one-eyed group makes them live together. Connecting individual stones into a single group results in an increase of liberties ...

[Text]

?

[Meaning]



[Information Need]

Machine Reading

Machines understanding text?

Machine Reading

"A machine comprehends a passage of text if, for any question regarding that text that can be answered correctly by a majority of native speakers, that machine can provide a string which those speakers would agree both answers that question, and does not contain information irrelevant to that question."

Towards the Machine Comprehension of Text: An Essay

Christopher J.C. Burges
Microsoft Research
One Microsoft Way
Redmond, WA 98052, USA

December 23, 2013

Machine Reading

A **machine** processes a **passage of text** to satisfy an **information need** (usually answer a question on it)

Machine Reading

In January 1880, two of Tesla's uncles put together enough money to help him leave Gospić for Prague where he was to study. Unfortunately, he arrived too late to enrol at Charles-Ferdinand University; he never studied Greek, a required subject; and he was illiterate in Czech, another required subject. Tesla did, however, attend lectures at the university, although, as an auditor, he did not receive grades for the courses.

[Passage of Text]

uses for

[Information Need]

Machine Reading

In January 1880, two of Tesla's uncles put together enough money to help him leave Gospić for Prague where he was to study. Unfortunately, he arrived too late to enrol at Charles-Ferdinand University; he never studied Greek, a required subject; and he was illiterate in Czech, another required subject. Tesla did, however, attend lectures at the university, although, as an auditor, he did not receive grades for the courses.

?

[Passage of Text]

converts into

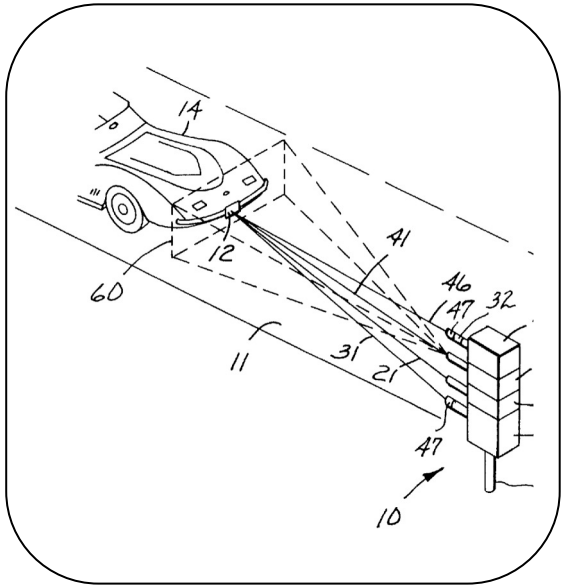
[Meaning]

uses for

[Information Need]

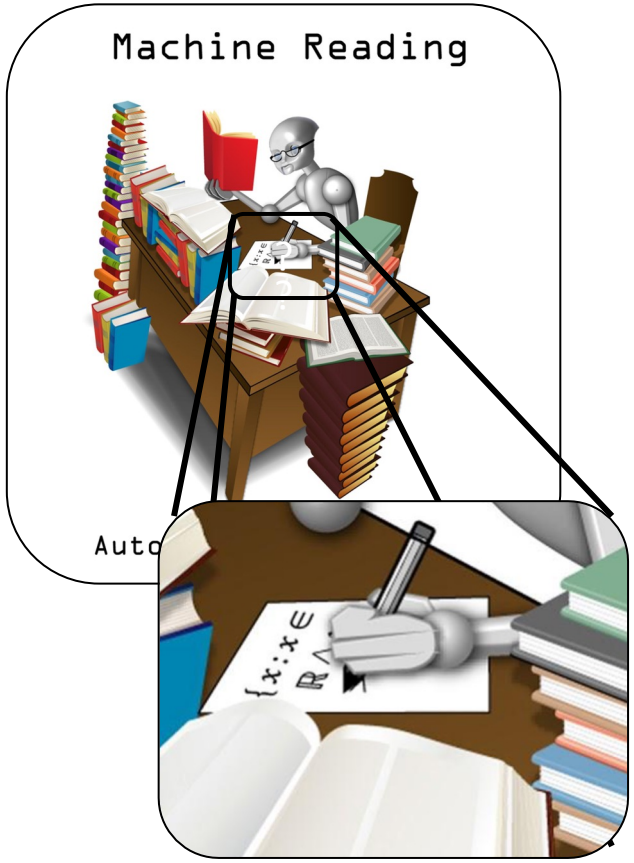
Timeline of Machine Reading

Something else entirely!

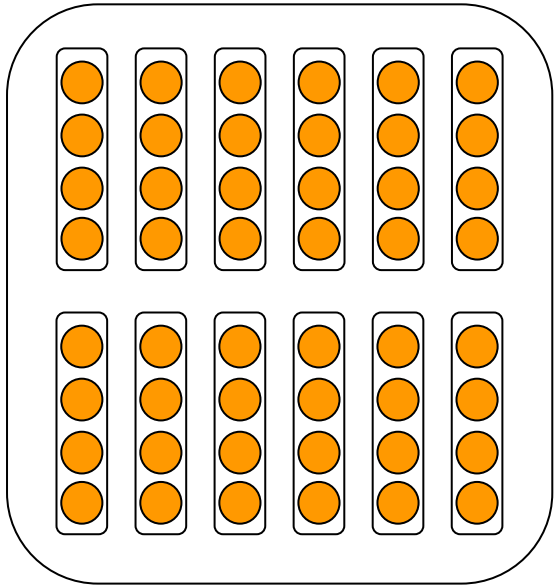


before 2006

Text to Meaning Representations



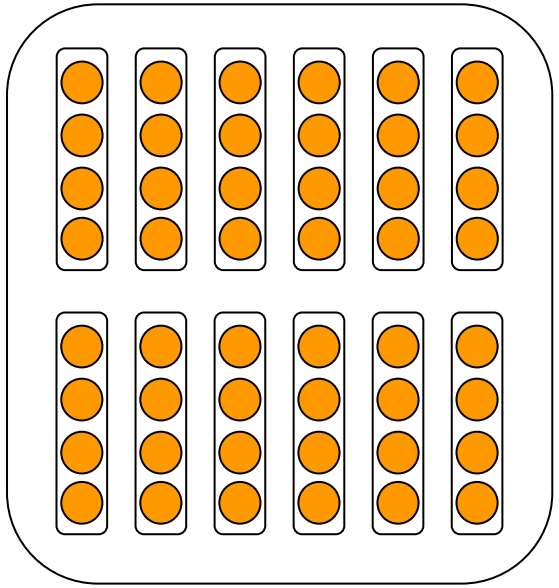
End-to-End Machine Reading



since 2014

Today we cover:

End-to-End
Machine Reading



since 2014

Machine Reading

In January 1880, two of Tesla's uncles put together enough money to help him leave Gospić for Prague where he was to study. Unfortunately, he arrived too late to enrol at Charles-Ferdinand University; he never studied Greek, a required subject; and he was illiterate in Czech, another required subject. Tesla did, however, attend lectures at the university, although, as an auditor, he did not receive grades for the courses.

?

[Passage of Text]

converts into

[Meaning]

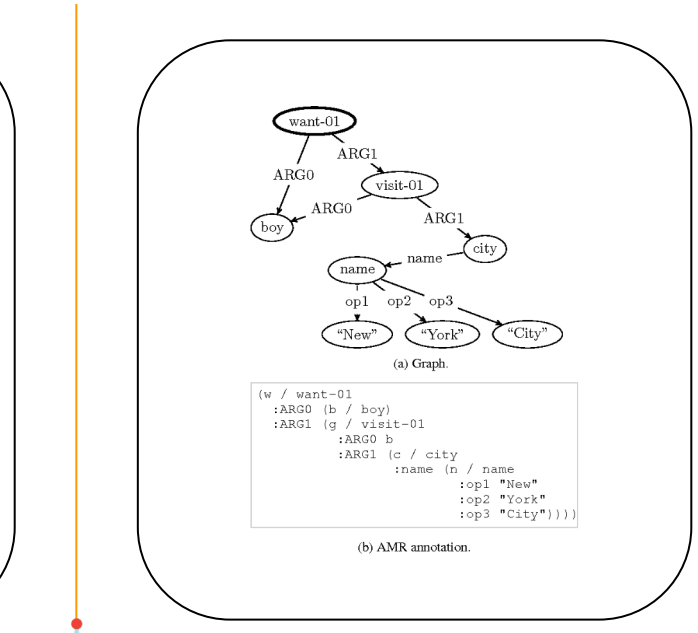
uses for

[Information Need]

Symbolic Approaches (until 2014 or so)

In January 1880, two of Tesla's uncles put together enough money to help him leave Gospić for Prague where he was to study. Unfortunately, he arrived too late to enrol at Charles-Ferdinand University; he never studied Greek, a required subject; and he was illiterate in Czech, another required subject. Tesla did, however, attend lectures at the university, although, as an auditor, he did not receive grades for the courses.

[Passage of Text]



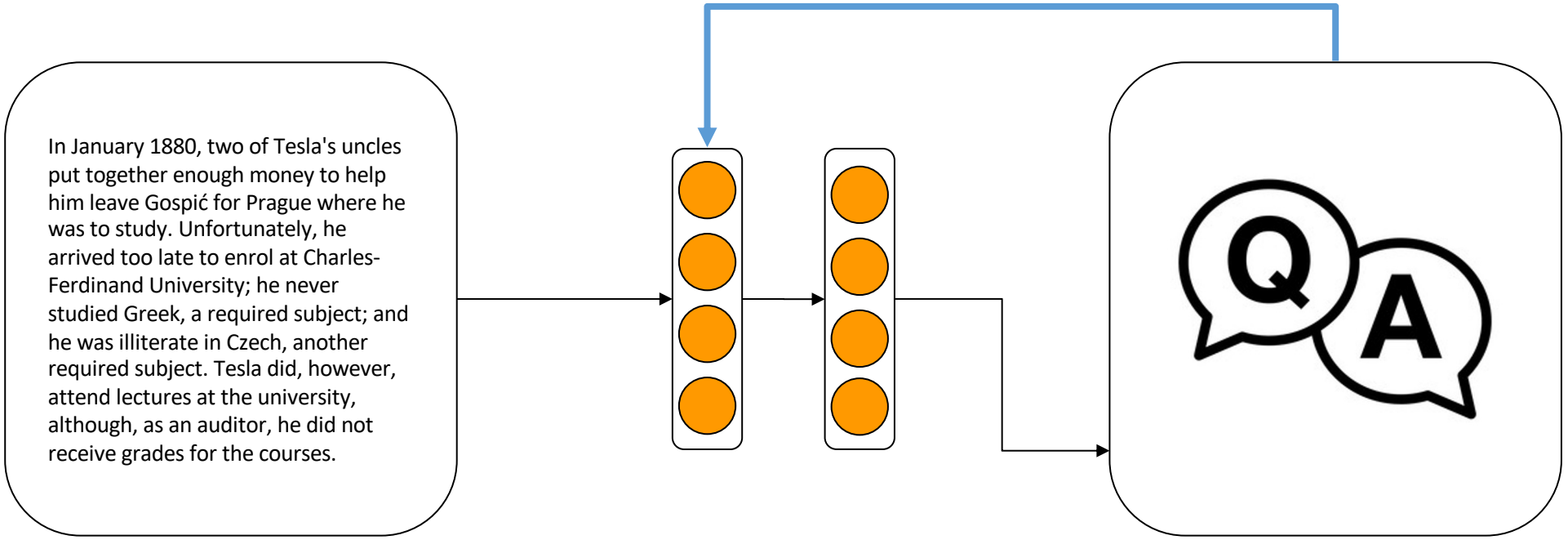
[Meaning]

[Information Need]

converts into

uses for

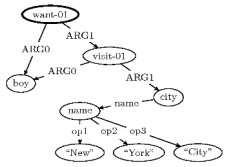
End-to-End Approaches (since 2014 or so)



What do we need from a representation?

In January 1880, two of Tesla's uncles put together enough money to help him leave Gospić for Prague where he was to study. Unfortunately, he arrived too late to enrol at Charles-Ferdinand University; he never studied Greek, a required subject; and he was illiterate in Czech, another required subject. Tesla did, however, attend lectures at the university, although, as an auditor, he did not receive grades for the courses.

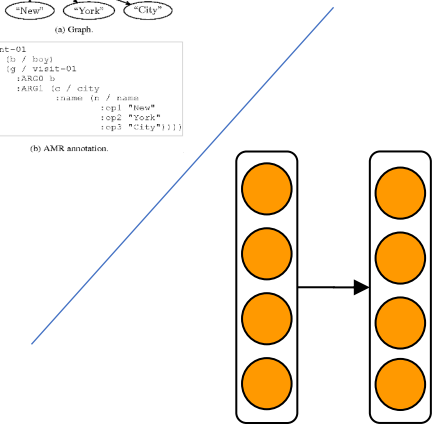
[Passage of Text]



(a) Graph.

```
(w / want-01
  :ARG0 (b / boy)
  :ARG1 (g / visit-01
          :ARG0 b
          :ARG1 (c / city
                  :name (n / name
                          :op1 "New"
                          :op2 "York"
                          :op3 "City"))))
```

(b) AMR annotation.



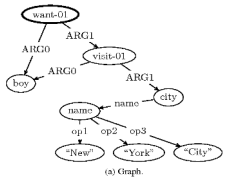
[Meaning]

- Fast Retrieval
- Generalization
- Broad Coverage
- Easy Engineering
- Support Reasoning
- Small Memory Footprint

What are the core challenges?

In January 1880, two of Tesla's uncles put together enough money to help him leave Gospić for Prague where he was to study. Unfortunately, he arrived too late to enrol at Charles-Ferdinand University; he never studied Greek, a required subject; and he was illiterate in Czech, another required subject. Tesla did, however, attend lectures at the university, although, as an auditor, he did not receive grades for the courses.

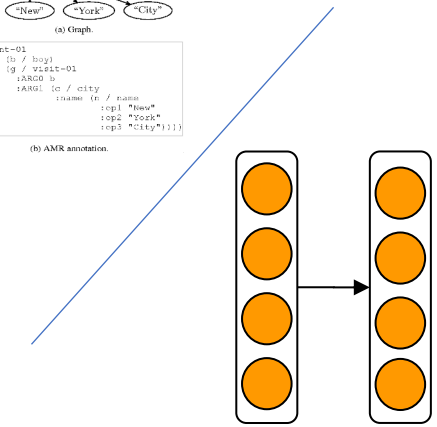
[Passage of Text]



(a) Graph.

```
(w / want-01
  :ARG0 (b / boy)
  :ARG1 (g / visit-01
          :ARG0 b
          :ARG1 (c / city
                  :name (n / name
                          :op1 "New"
                          :op2 "York"
                          :op3 "City"))))
```

(b) AMR annotation.

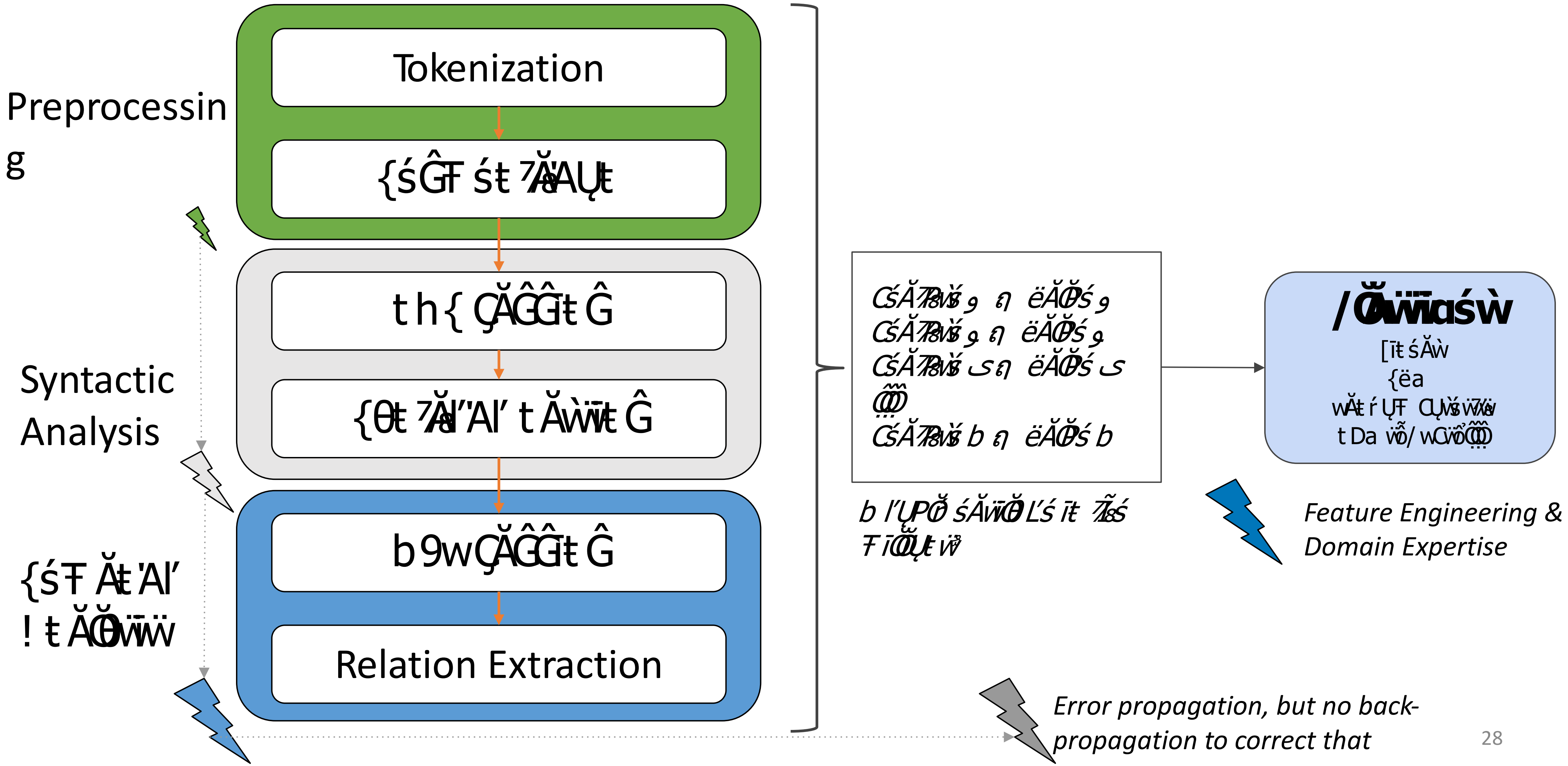


[Meaning]

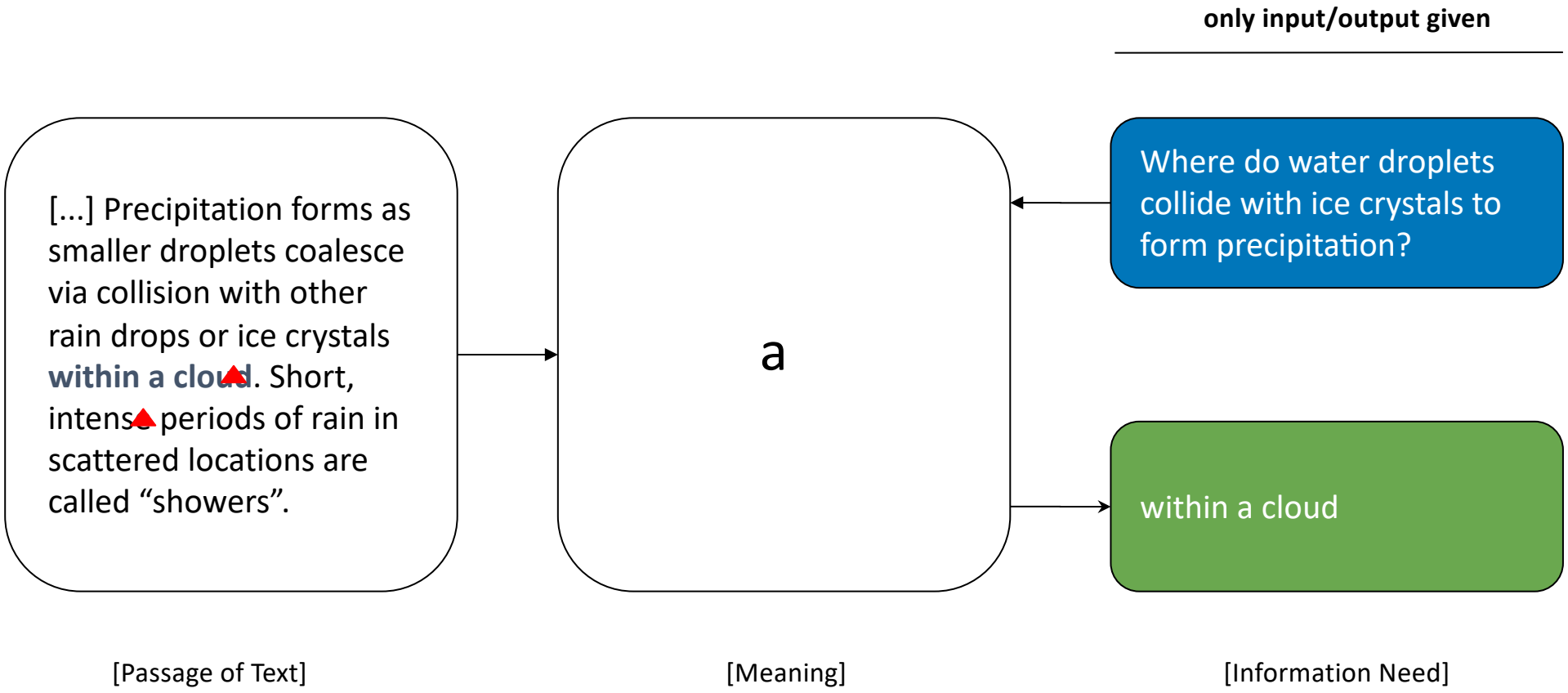
- Ambiguity
- Variation
- Coreference
- Common Sense
- Scale
- …

Unfortunately, he arrived too late to enrol at Charles-Ferdinand University; he never studied Greek, a required subject; and he was illiterate in Czech, another required subject. Tesla did, however, attend lectures at the university, although, as an auditor, he did not receive grades for the courses.

"Traditional" NLP



End-to-end System



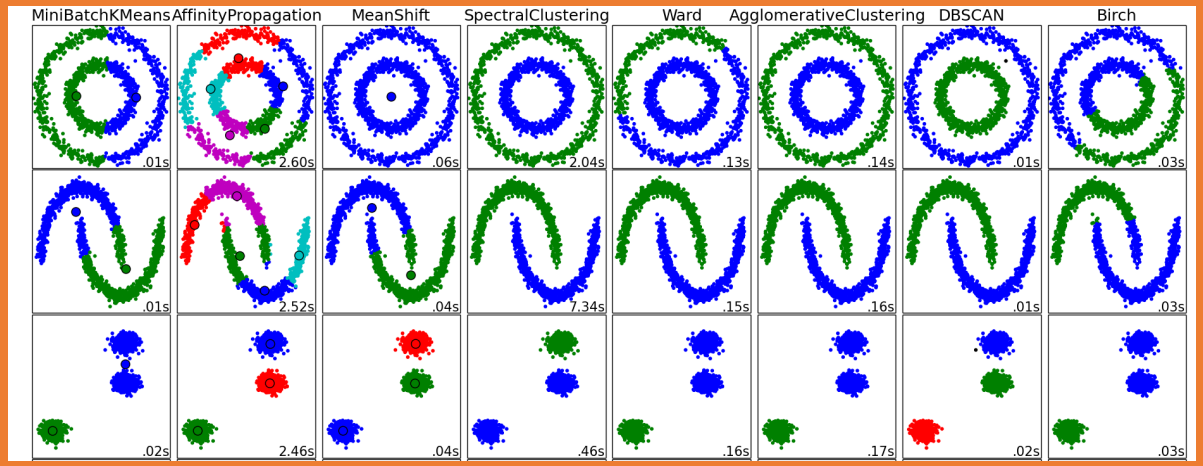
Machine Reading / Data

Limits of Big Model + Big Data

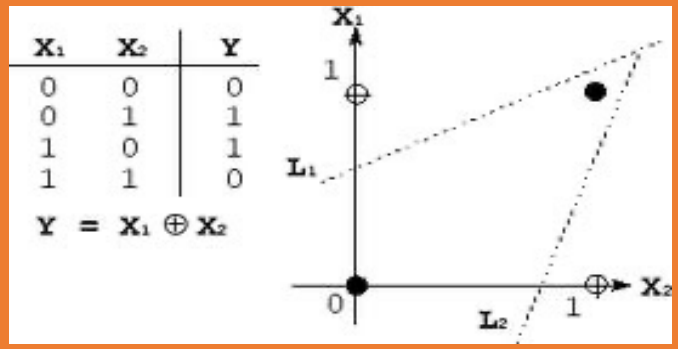
- End-to-end systems need a lot of cleaned data!
- Getting real conditions large-scale data is difficult:
  - Real large-scale data is complex, noisy, unlabeled…
  - Interpretation of success or failure is complex
- Complicates the design of innovative learning systems
- Some recent process with large amounts of unlabeled data [see Tomorrow]

A long history of artificial problems in ML

Two moons and friends (clustering)



XOR (neural networks)

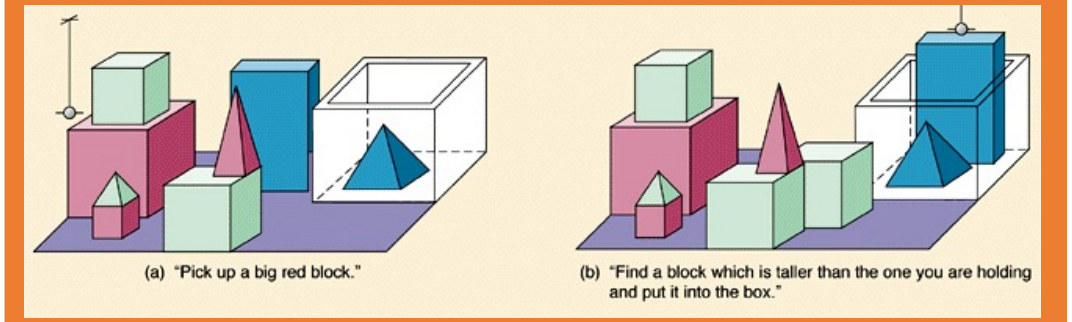


Toy/artificial problems in ML: crucial for demonstrating and assessing the usefulness/efficiency of new algorithms

Many in the UCI repository (regression, classification)

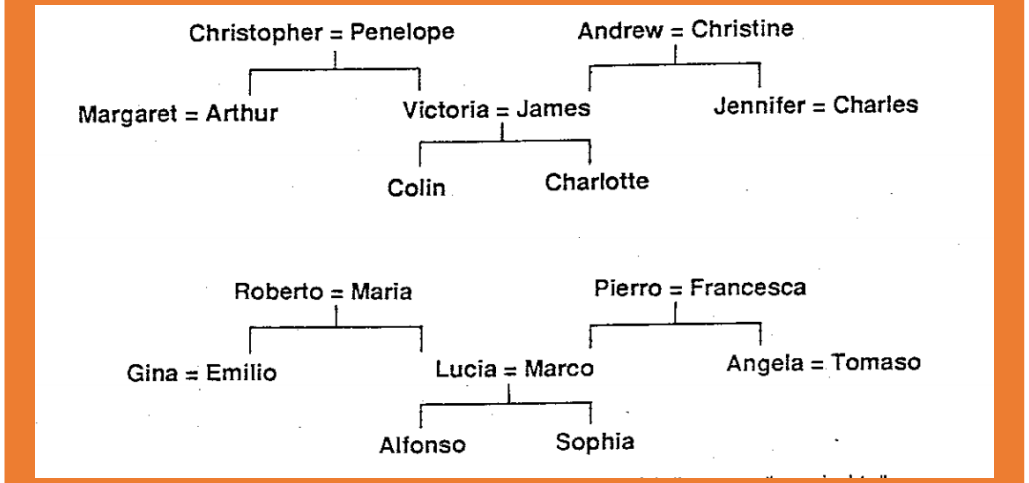
And in early AI too

Block Worlds (Winograd 1971)




```
Person: PICK UP A BIG RED BLOCK.
Computer: OK. (does it)
Person: GRASP THE PYRAMID.
Computer: I DON'T UNDERSTAND WHICH PYRAMID YOU MEAN.
Person: FIND A BLOCK WHICH IS TALLER THAN THE ONE YOU ARE HOLDING AND PUT IT INTO THE BOX.
Computer: BY "IT", I ASSUME YOU MEAN THE BLOCK WHICH IS TALLER THAN THE ONE I AM HOLDING.
Computer: OK. (does it)
Person: WHAT DOES THE BOX CONTAIN?
Computer: THE BLUE PYRAMID AND THE BLUE BLOCK.
Person: WHAT IS THE PYRAMID SUPPORTED BY?
Computer: THE BOX.
Person: HOW MANY BLOCKS ARE NOT IN THE BOX?
Computer: FOUR OF THEM.
Person: IS AT LEAST ONE OF THEM NARROWER THAN THE ONE WHICH I TOLD YOU TO PICK UP?
Computer: YES, THE RED CUBE.
Person: IS IT SUPPORTED?
Computer: YES, BY THE TABLE.
Person: CAN THE TABLE PICK UP BLOCKS?
Computer: NO.
Person: CAN A PYRAMID BE SUPPORTED BY A BLOCK?
Computer: YES.
Person: CAN A PYRAMID SUPPORT A PYRAMID?
Computer: I DON'T KNOW.
Person: STACK UP TWO PYRAMIDS.
```

Family trees (Hinton CogSci'86)



bAbI Tasks

Weston et al., ICLR'15

- 20 artificial tasks for machine reading
- Answer has to be picked among a set of candidates
- Each task checks one skill that a reasoning system should have.
- Goal: develop systems able to solve all tasks *with no task specific engineering*.

A postulate is that *performing well on all of them is a pre-requisite for any system aiming at understanding language and able to reason*.

Simulation commands

- go <place>
- get <object>
- get <object1> from <object2>
- put <object1> in/on <object2>
- give <object> to <person>
- drop <object>
- look
- inventory
- examine <object>

+ 2 commands for "gods" (superusers):

- create <object>
- set <obj1> <relation> <obj2>

Example

Command format

```
jason go kitchen
jason get milk
jason go office
jason drop milk
jason go bathroom
where is milk ?   A: office
where is jason? A: bathroom
```

Story

Jason went to the kitchen.

Jason picked up the milk.

Jason travelled to the office.

Jason left the milk there.

Jason went to the bathroom.

Where is the milk now? A: office

Where is Jason? A: bathroom

(T1) Single supporting fact "where is actor"

- A single supporting fact, previously given, provides the answer.
- Simplest case of this: asking for the location of a person.

John is in the playground.
Bob is in the office.
Where is John? A:playground

SUPPORTING FACT

(T2) Two supporting facts "where is actor+object"

- Harder task: two supporting statements have to be chained to answer

John is in the playground.
Bob is in the office.
John picked up the football.
Bob went to the kitchen.
Where is the football? A:playground
Where was Bob before the kitchen? A:office

SUPPORTING FACT

SUPPORTING FACT

- To answer the first question *Where is the football?* both John picked up the football and John is in the playground are supporting

(T3) Three supporting facts

- Similarly, one can make a task with three supporting facts:

John picked up the apple.
John went to the office.
John went to the kitchen.
John dropped the apple.
Where was the apple before the kitchen? A:office

- The first three statements are all required to answer this.

(T4) Two argument relations: subj vs. obj.

- To answer questions the ability to differentiate and recognize subjects and objects is crucial
- Extreme case - sentences feature re-ordered words:

The office is north of the bedroom.
The bedroom is north of the bathroom.
What is north of the bedroom? A:office
What is the bedroom north of? A:bathroom

- The two questions above have exactly the same words, but in a different order, and different answers.
- So a bag-of-words will not work.

(T6) Yes/No questions

- This task tests, in the simplest case possible (with a single supporting fact) the ability of a model to answer true/false type questions:

John is in the playground.
Daniel picks up the milk.
Is John in the classroom? A:no
Does Daniel have the milk? A:yes

(T7) Counting

- This task tests the ability of the QA system to perform simple counting operations, by asking about the number of objects with a certain property:

Daniel picked up the football.
Daniel dropped the football.
Daniel got the milk.
Daniel took the apple.
How many objects is Daniel holding? A:two

(T17) Positional reasoning

- This task tests spatial reasoning:

> The triangle is to the right of the blue square.
> The red square is on top of the blue square.
> The red sphere is to the right of the blue square.
> Is the red sphere to the right of the blue square? A:yes
> Is the red square to the left of the triangle? A:yes

- Close from block worlds, with no vision input.
- The Yes/No task (6) is a prerequisite.

(T18) Reasoning about size

- This task requires reasoning about relative size of objects :

The football fits in the suitcase.
The suitcase fits in the cupboard.
The box of chocolates is smaller than the football.
Will the box of chocolates fit in the suitcase? A:yes

- Inspired by the commonsense reasoning examples of the Winograd schema challenge
- Tasks 3 (three supporting facts) and 6 (Yes/No) are prerequisites.

Winograd Schema

Levesque, AAAI'11

Definition: *A Winograd schema is a pair of sentences that differ in only one or two words and that contain an ambiguity that is resolved in opposite ways in the two sentences and requires the use of world knowledge and reasoning for its resolution.*

The **trophy** would not fit in the brown **suitcase** because **it** was too ***big***.

The **trophy** would not fit in the brown **suitcase** because **it** was too ***small***.

it = **trophy** or **suitcase** ?

More schemas here: https://cs.nyu.edu/faculty/davise/papers/WinogradSchemas/WSCollection.html

(T19) Path finding

- In this task the goal is to find the path between locations:

> The kitchen is north of the hallway.
> The den is east of the hallway.
> How do you go from den to kitchen? A:west,north

- This task is difficult because it effectively involves search.

Dashboard

Training on 1k stories

Weak supervised

Fully supervised

TASK	N-grams	LSTMs	StructSVM + COREF + SRL	Attention model
T1. Single supporting fact	36	50	PASS	PASS
T2. Two supporting facts	2	20	74	PASS
T3. Three supporting facts	7	20	17	PASS

Rank	Method	Accuracy (trained on 10k)	Accuracy (trained on 1k)	Mean Error Rate	Paper Title	Year	Paper	Code
1	QRN	99.7%	90.1%	0.3%	Query-Reduction Networks for Question Answering	2016		
2	EntNet	99.5%	89.1%	9.7%	Tracking the World State with Recurrent Entity Networks	2016		

Source: https://paperswithcode.com/sota/question-answering-babi

T11. Basic coreference				PASS
T12. Conjunction	9	74	PASS	PASS
T13. Compound coreference	26	PASS	PASS	PASS
T14. Time reasoning	19	27	PASS	PASS
T15. Basic deduction	20	21	PASS	PASS
T16. Basic induction	43	23	24	PASS
T17. Positional reasoning	46	51	61	48
T18. Size reasoning	52	52	62	68
T19. Path finding	0	8	49	4
T20. Agent's motivation	76	91	PASS	PASS

Artificial tasks for Machine Reading

- Advantages:
  - Total control on the complexity of the tasks/reasoning
  - Clear interpretation of results
  - Small-ish scale so easy to prototype on them
- Challenges:
  - How do we know that artificial data models the right problem?
  - By creating the tasks that we are solving, aren't we fooling ourselves?
  - How transfer from artificial to real conditions?

Other Machine Reading & QA Datasets

Leaderboards

Add a Result

TREND	DATASET	BEST METHOD	PAPER TITLE	PAPER	CODE	COMPARE
	SQuAD2.0	Retro-Reader on ALBERT (ensemble)	Retrospective Reader for Machine Reading Comprehension			See all
	SQuAD1.1	XLNet (single model)	XLNet: Generalized Autoregressive Pretraining for Language Understanding			See all
	SQuAD1.1 dev	T5-11B	Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer			See all
	WikiQA	TANDA-RoBERTa (ASNQ, WikiQA)	TANDA: Transfer and Adapt Pre-Trained Transformer Models for Answer Sentence Selection			See all
	CNN / Daily Mail	GA+MAGE (32)	Linguistic Knowledge as Memory for Recurrent Neural Networks			See all
	Quora Question Pairs	ALBERT	ALBERT: A Lite BERT for Self-supervised Learning of Language Representations			See all
	SQuAD2.0 dev	XLNet+DSC	Dice Loss for Data-imbalanced NLP			See all
	bAbi	STM	Self-Attentive Associative Memory			See all
	NarrativeQA	(NarrativeQA + MS MARCO)	Generative Reading Comprehension			See all
	Children's Book Test	GPT-2	Language Models are Unsupervised Multitask Learners			See all
	CoQA	BERT Large Augmented (single model)	BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding			See all
	QASent	Attentive LSTM	Neural Variational Inference for Text Processing			See all
	YahooCQA	sMIM (1024) +	SentenceMIM: A Latent Variable Language Model			See all

Many more (60+) on: paperswithcode.com/task/question-answering

Stanford Question Answering Dataset (SQuAD)

Rajpurkar et. al., EMNLP'16

- **Dataset size**: 107,702 samples
- Widely used benchmark dataset
- **Task:** Extractive Question Answering
  - System has to predict the start and end position of the answer in the passage of text

Stanford Question Answering Dataset (SQuAD)

Text Passage

[...] Precipitation
smaller droplets
via collision with
rain drops or ice
within a cloud. Sl
intense periods o
scattered locatio
called “showers”.

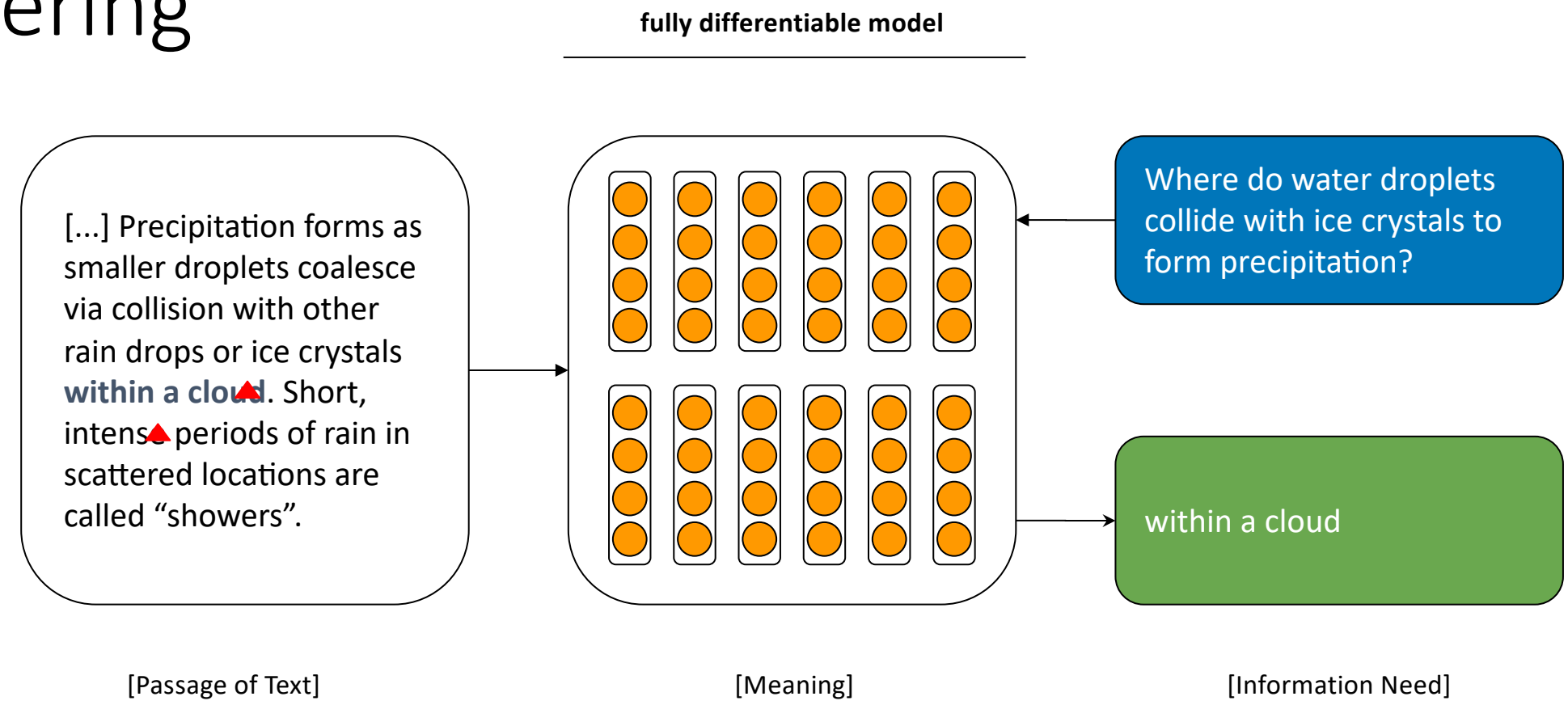
er droplets
crystals to
tion?

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Mar 05, 2019	BERT + N-Gram Masking + Synthetic Self-Training (ensemble) *Google AI Language* https://github.com/google-research/bert	**86.673**	**89.147**
2 Mar 05, 2019	BERT + N-Gram Masking + Synthetic Self-Training (single model) *Google AI Language* https://github.com/google-research/bert	85.150	87.715
3 Jan 15, 2019	BERT + MMFT + ADA (ensemble) *Microsoft Research Asia*	85.082	87.615
4 Jan 10, 201	BERT + Synthetic Self-Training (ensemble)	84.292	86.967
5 Dec 16, 201			043
5 Dec 21, 2018	PAML+BERT (ensemble model) *PINGAN GammaLab*	83.457	86.122

Very popular leaderboard!
https://stanford-qa.com

Machine Reading / Models

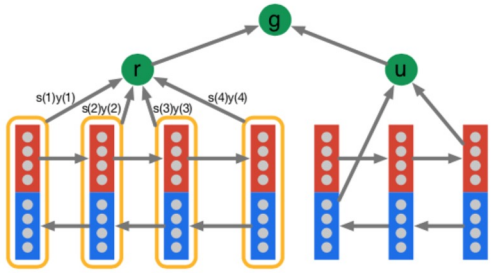
End-to-end Machine Reading for Question Answering



End-to-end Machine Reading for Question Answering

“Simple” Architecture

[...] Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain in scattered locations are called “showers”.



Hermann et al., 2015

Where do water droplets collide with ice crystals to form precipitation?

within a cloud

[Passage of Text]

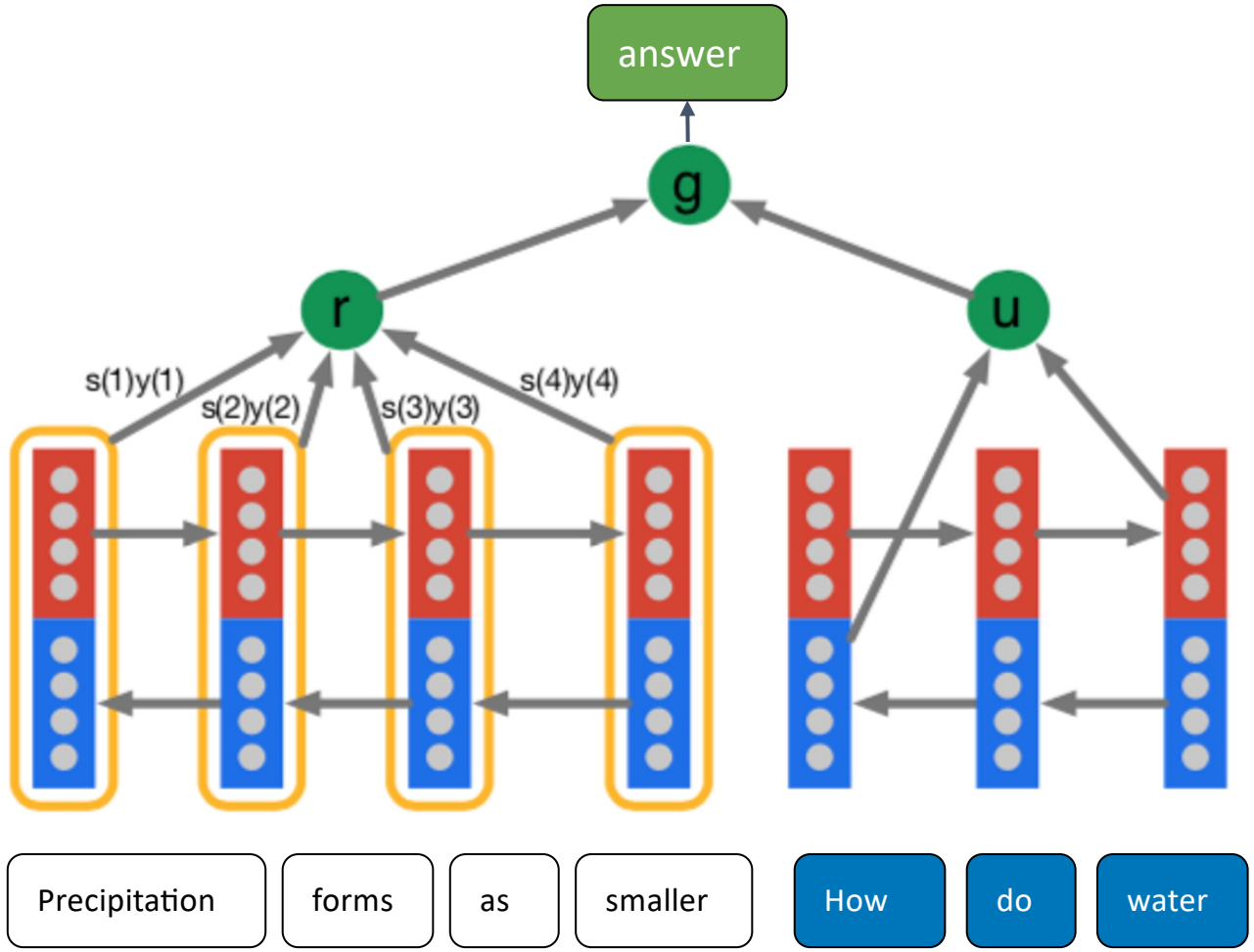
[Meaning]

[Information Need]

The Attentive Reader Model: Overview

Hermann et al., NIPS'15

- 'early' neural model for Machine Reading
- main components reused in many other models

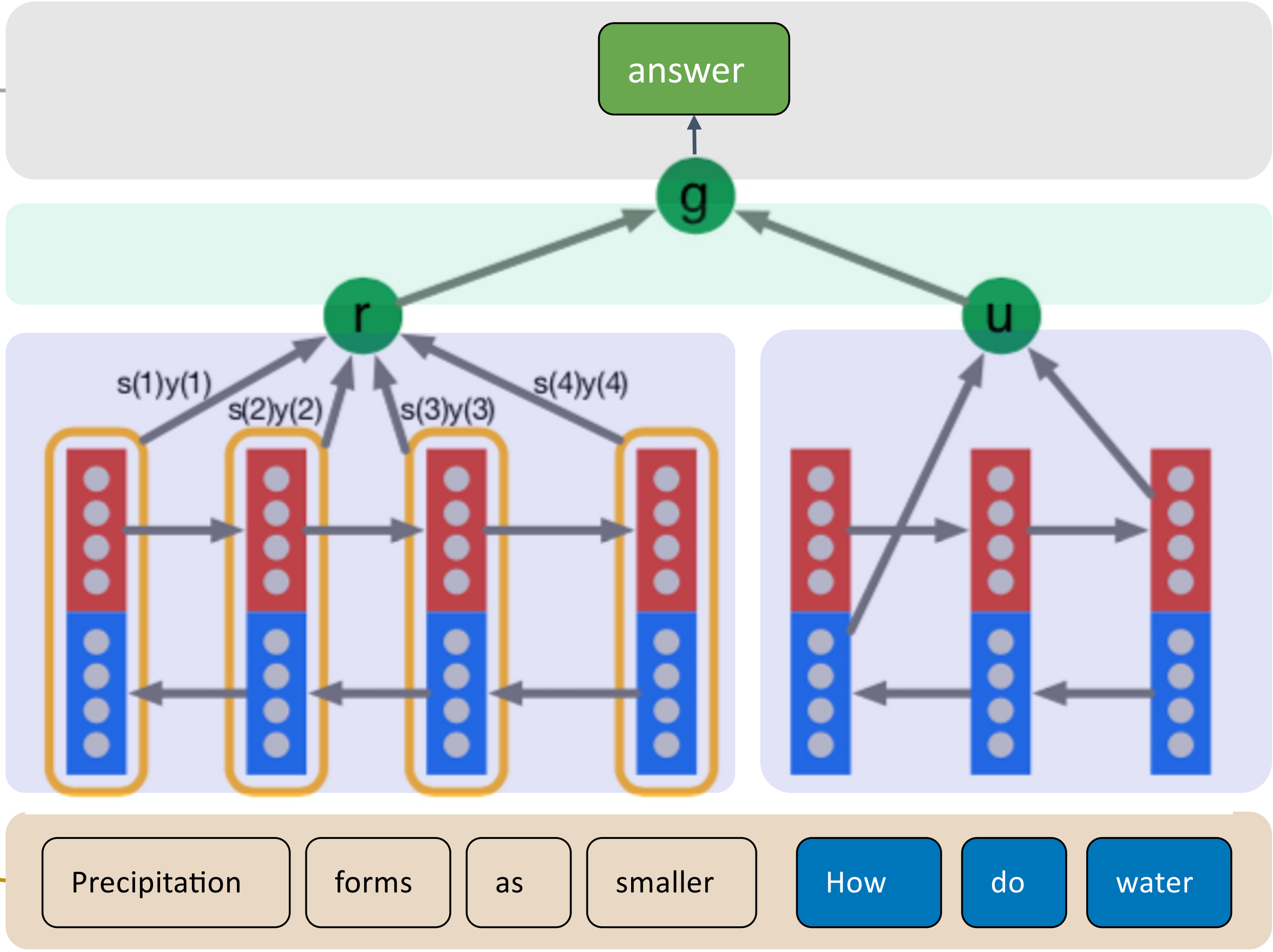


Modified visualization from Hermann et. al. NIPS'15

The Attentive Reader Model: Overview

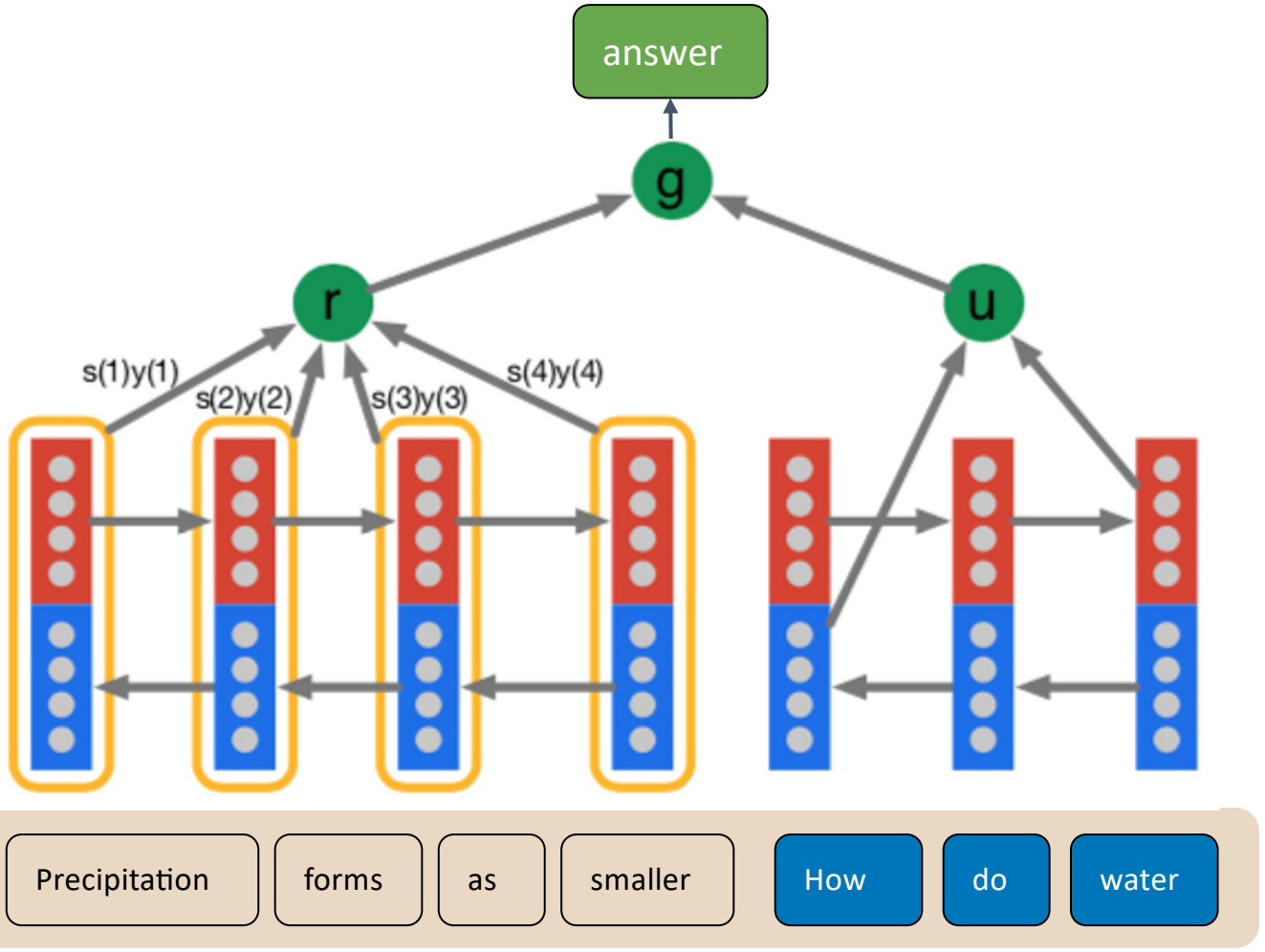
Sequence Interaction:
Matching text with question

Composition:
incorporating context around words



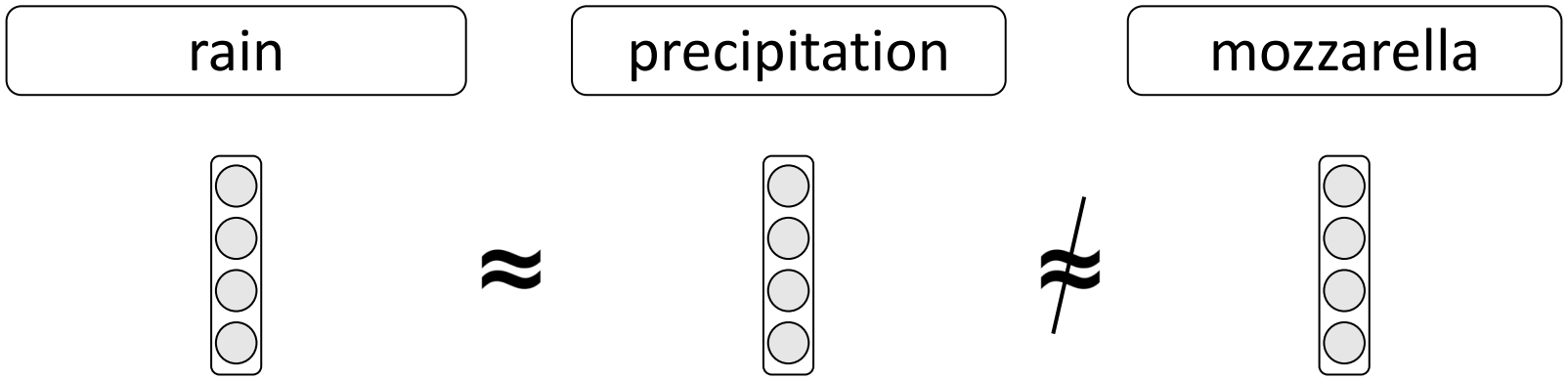
Modified visualization from Hermann et. al. NIPS’15

The Attentive Reader Model: Overview



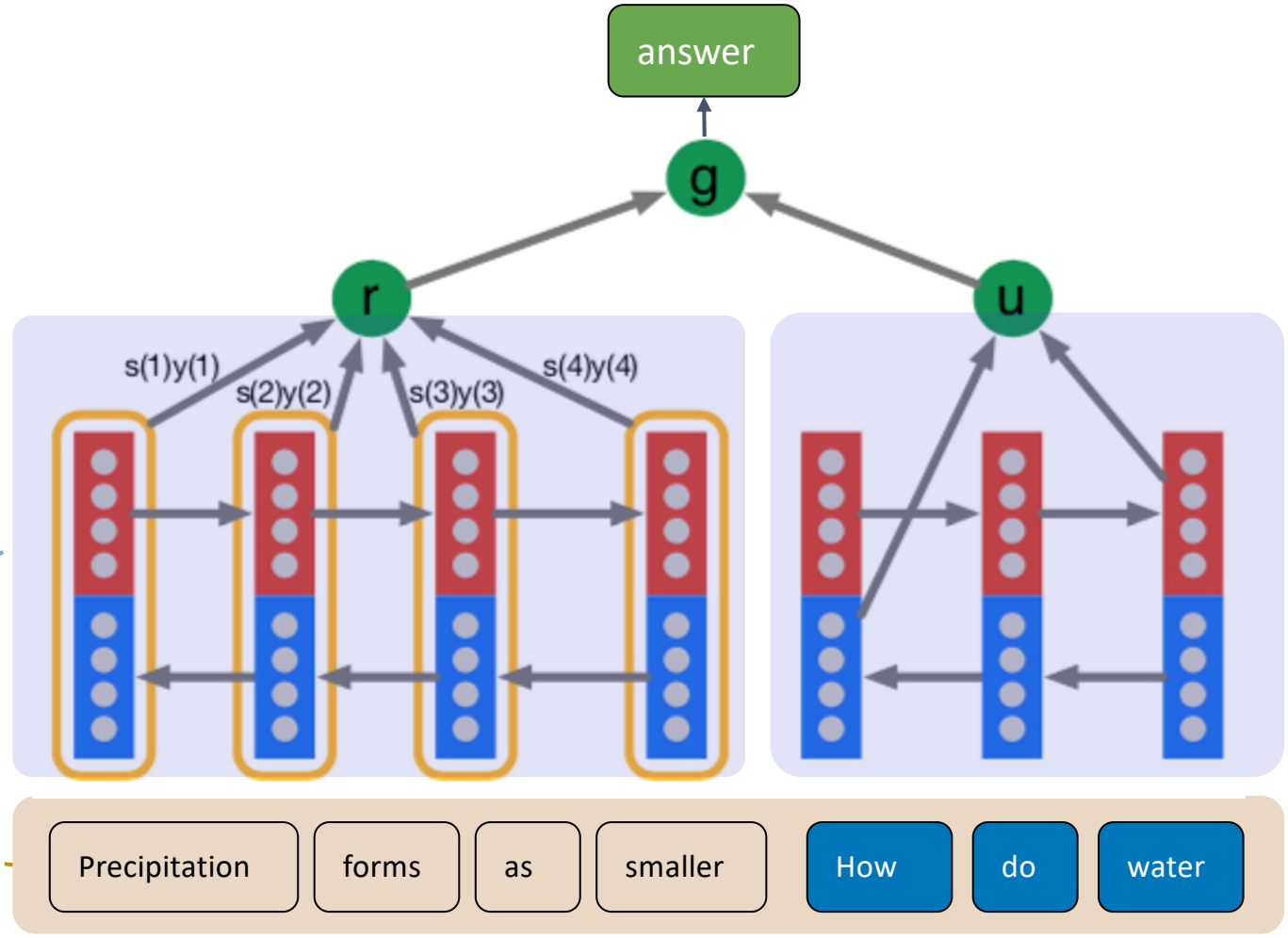
Modified visualization from Hermann et. al. NIPS'15

Representations for words: Embeddings



Similar meaning of words → similar vector representations – see previous lectures!

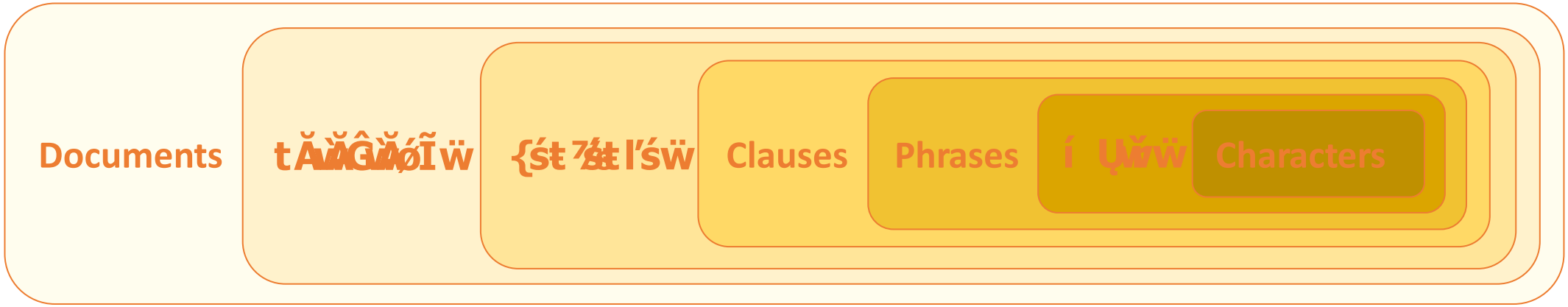
The Attentive Reader Model: Overview



Composition: incorporating context around words

Modified visualization from Hermann et. al. NIPS'15

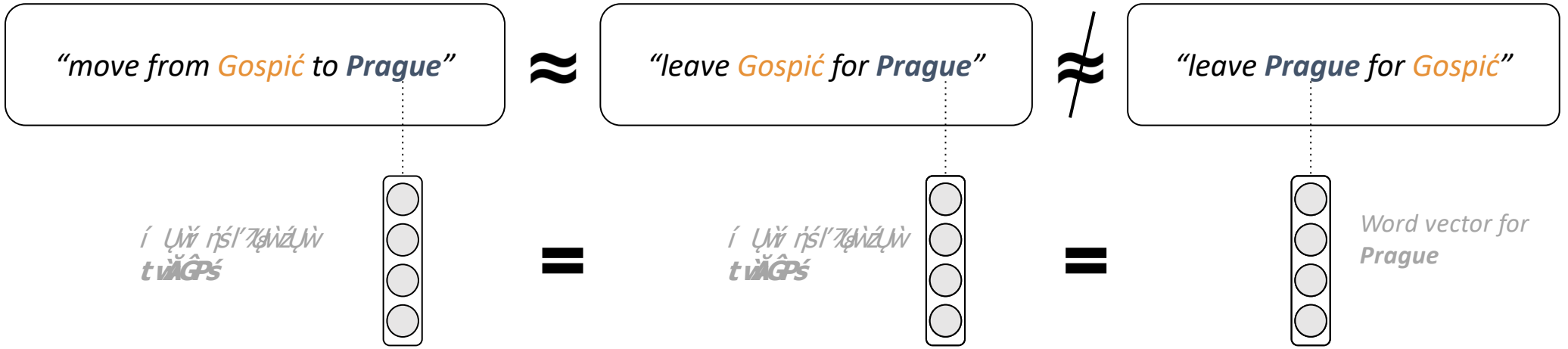
Language is compositional



Challenges

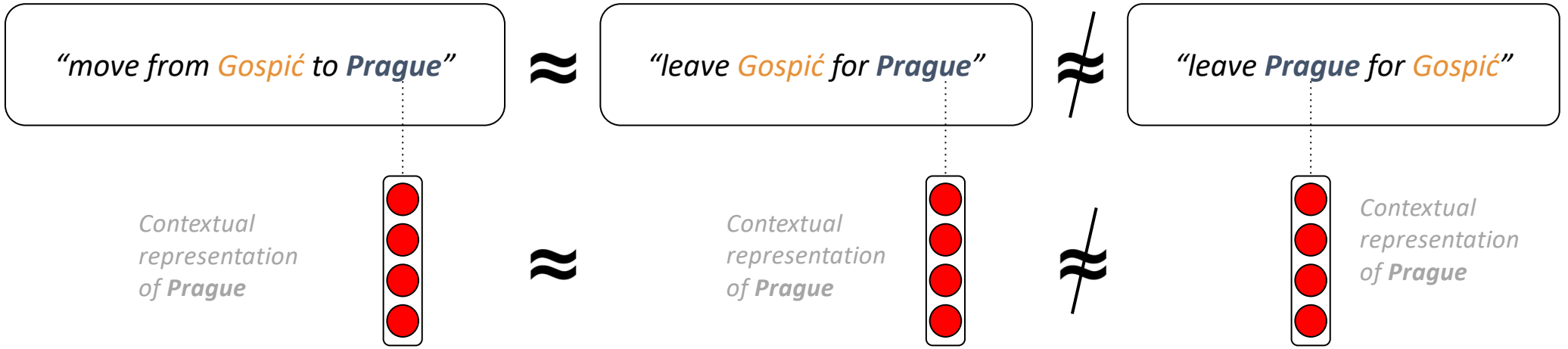
- Inductive bias: which composition function to use?
  - sequence, tree or more general graph structures?
  - varies for different levels
- Capturing long-range dependencies
  - co-reference (tracking entities)
  - effective information flow: ease of learning

Representing Words in Context



- Word representations should vary depending on context

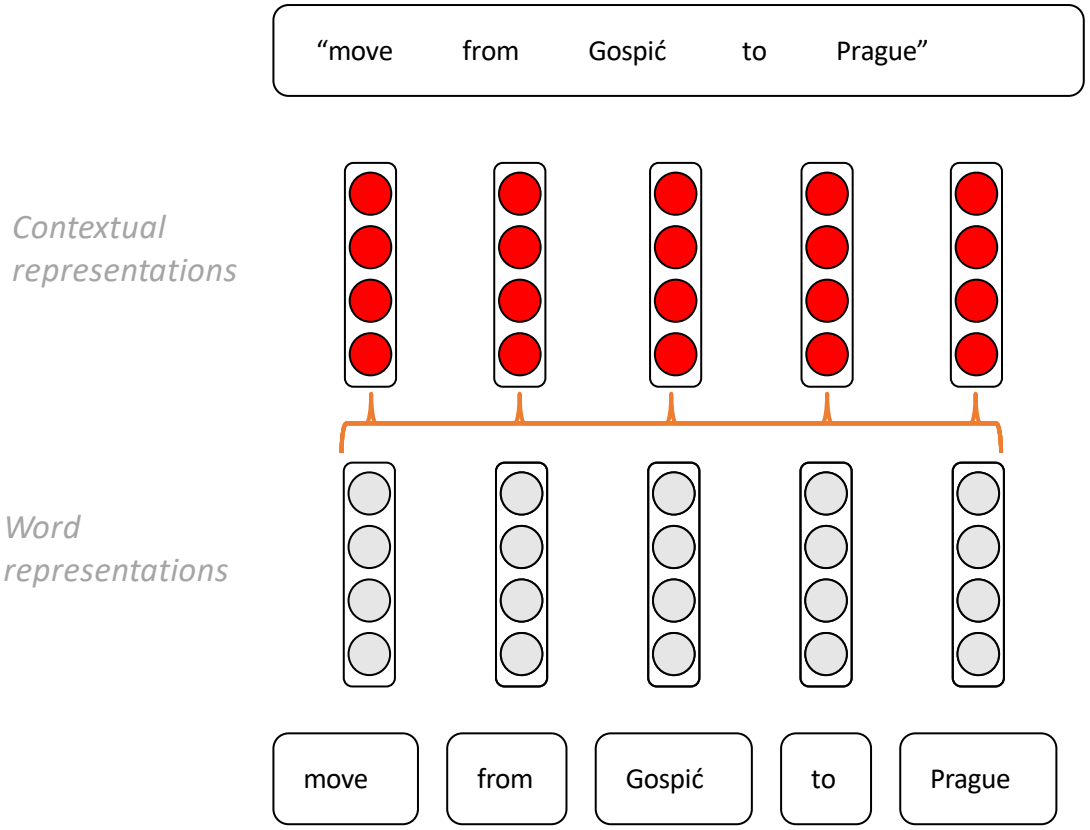
Representing Words in Context



- Word representations should vary depending on context
- **Contextual word representation:**
  - a word representation, computed conditionally on the given context

Representing Words in Context

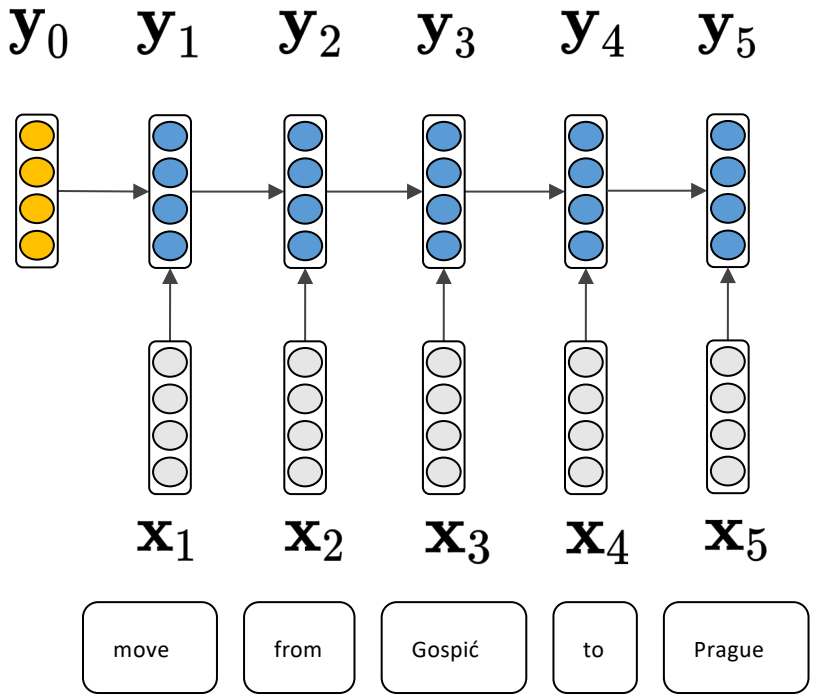
- composition of word vectors into contextualized word representations
- use vector composition function



Recurrent Neural Networks

- **Idea**: text as sequence
- Prominent ~~types:~~ *LSTM, GRU*
- **Inductive bias:** Recency
  - more recent symbols have bigger impact on hidden state
- **Advantages**
  - everything is connected
  - easy to train and robust in practice
- **Disadvantages**
  - Slow → computation time linear in length of text
  - not good for (very) long range dependencies

- *Good for:* sentences, small paragraphs



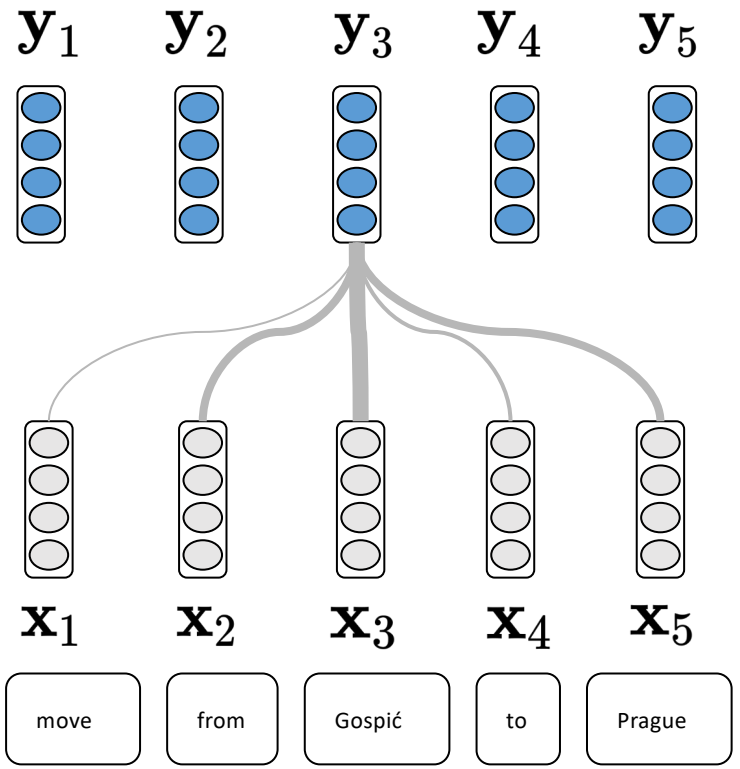
$$\mathbf{y}_t = f(\mathbf{x}_t, \mathbf{y}_{t-1})$$

Tree-variants:

- TreeLSTM (Tai et al., SCL'15)
- RNN Grammars (Dyer et al. NAACL'16)

Self-Attention Layer

- **Idea**: latent graph on text
- **Inductive bias:**
  - relationships between word pairs
- compute *K* separate weighted word representation(s) of the context for each word *t*
- **Advantages**
  - can capture long-range dependencies
  - Parallelizable and fast
- **Disadvantages**
  - careful setup of hyper-parameters
  - potentially memory intensive computation of attention weights for large contexts, *O(T * T * K)*



$$\mathbf{y}_t = f(\mathbf{x}_1, \ldots, \mathbf{x}_T)$$

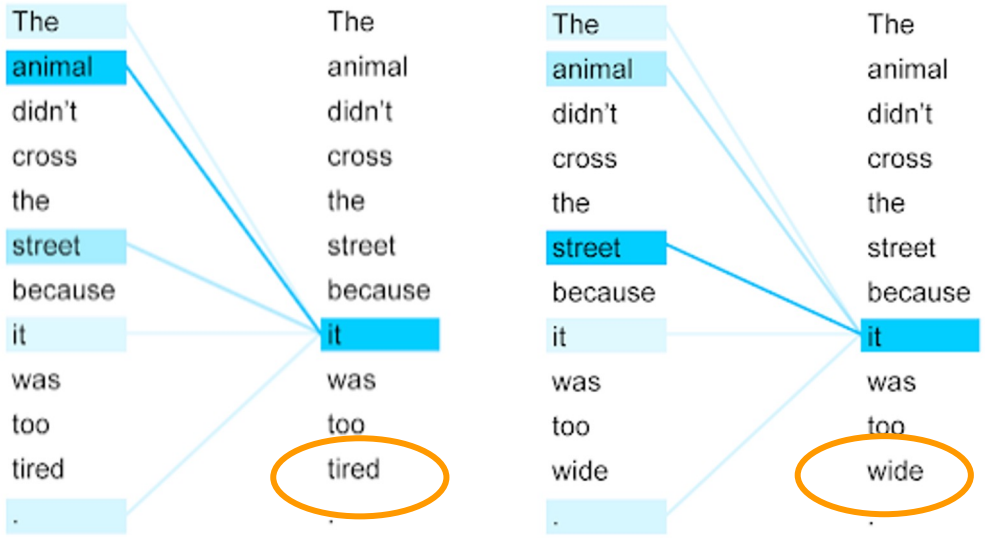
$$\tilde{\mathbf{x}}_t^k = \sum_{j=1}^{T} \alpha_{j,t}^k \mathbf{x}_j \qquad k = 1, \ldots, K$$

$$f(\mathbf{x}_1, \ldots, \mathbf{x}_T) = \text{nonlinear}(\tilde{\mathbf{x}}_t^1, \ldots, \tilde{\mathbf{x}}_t^K)$$

$\alpha_t^k : k^{th}$ self-attention weights for token t

Self-Attention Layer

- **Graph with weighted edges** of *K* types
- Can capture:
  - coreference chains
  - syntactic dependency structure in text



Transformer Self-Attention Coreference Visualization
https://ai.googleblog.com/2017/08/transformer-novel-neural-network.html

Transformer

Vaswani et al., NIPS'17

- Residual connections before and after multi-head attention
- Decoder uses both self attention and encoder attention

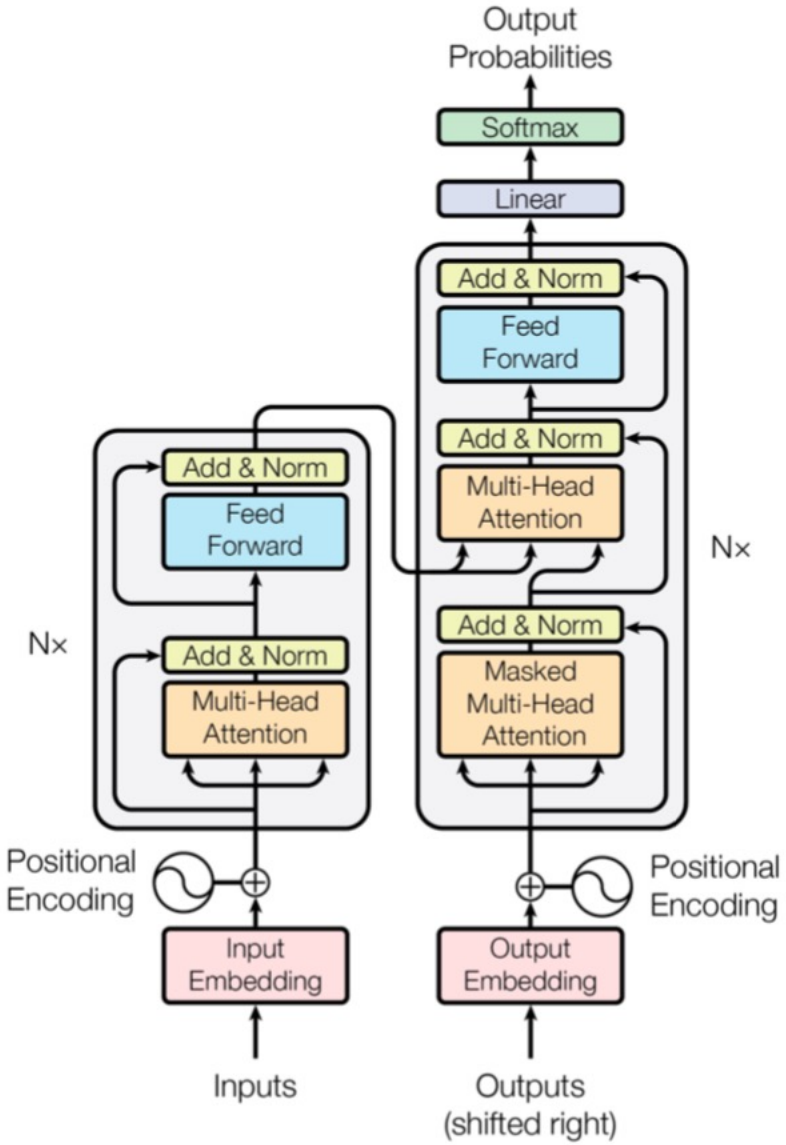
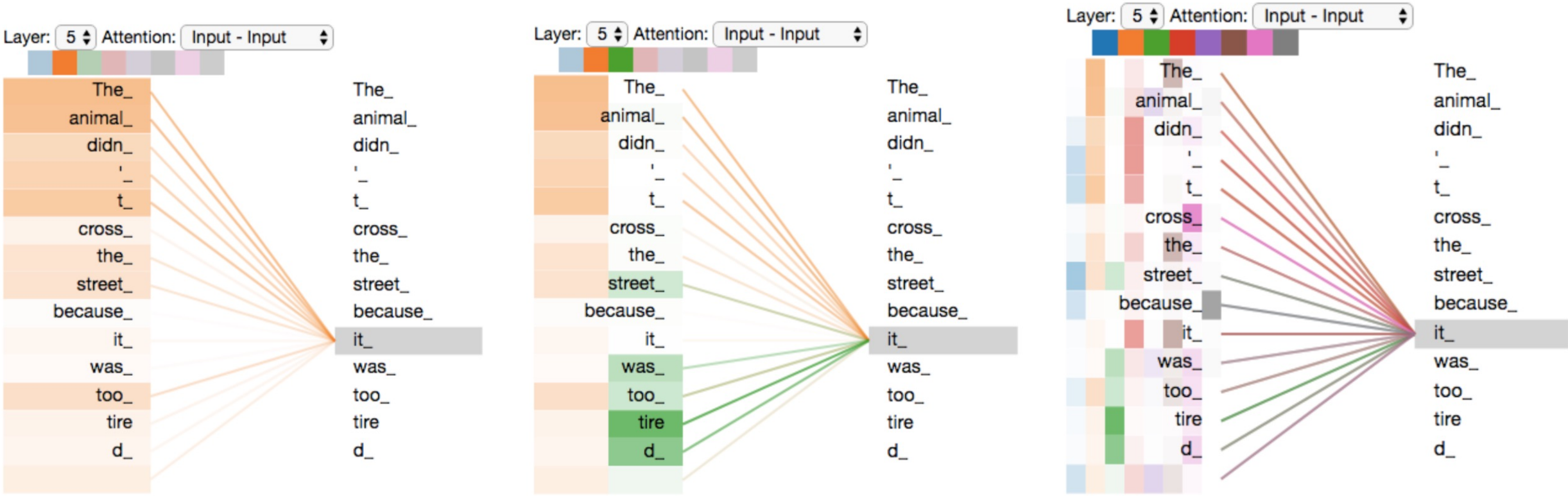


Figure from Vaswani et al., NIPS'17

Multi-head attention

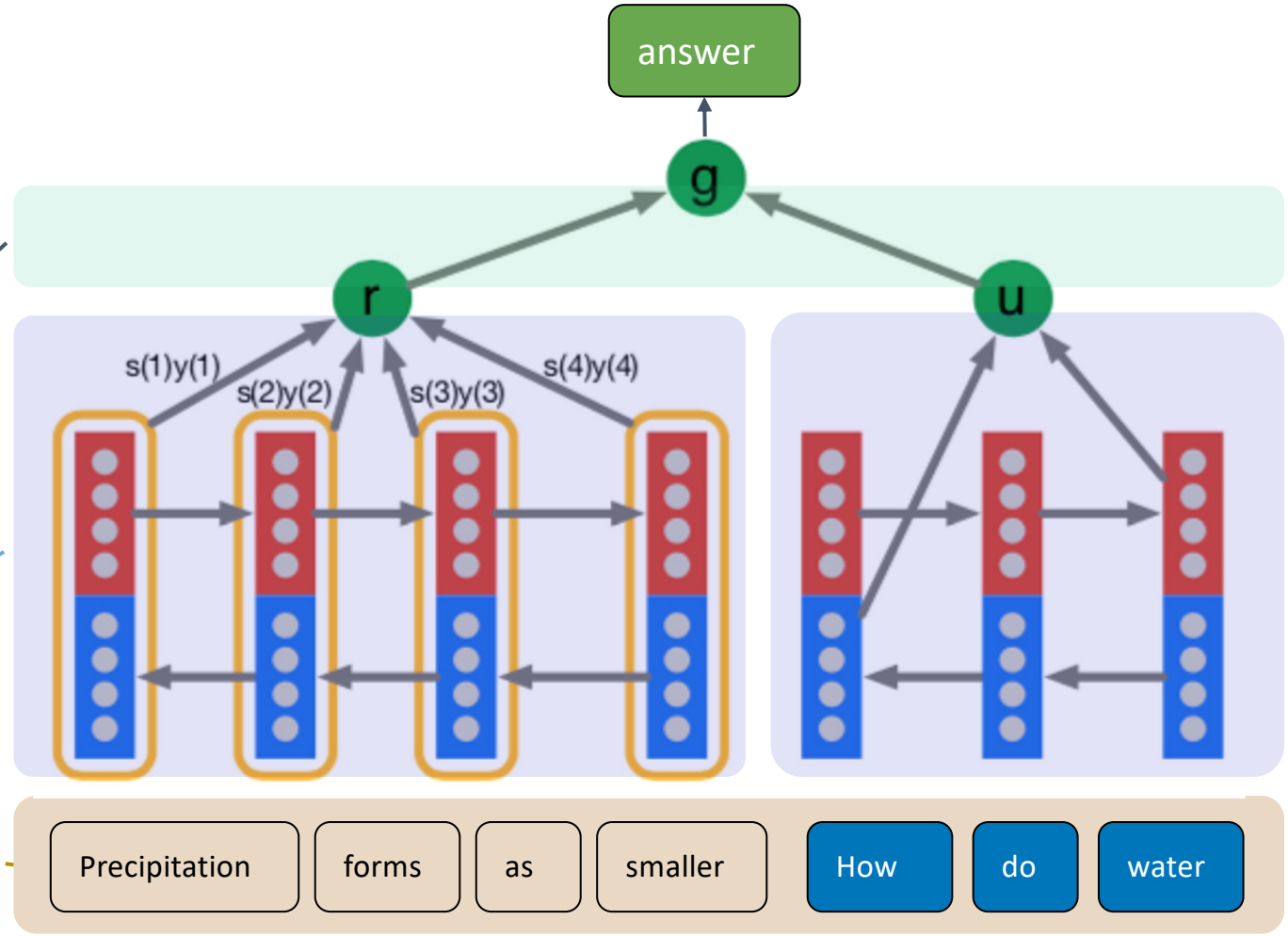


Figures from http://jalammar.github.io/illustrated-transformer/

The Attentive Reader Model: Overview

Sequence Interaction:
Matching text with question

Composition: incorporating context around words



Modified visualization from Hermann et. al. NIPS'15

Modelling sequence interactions

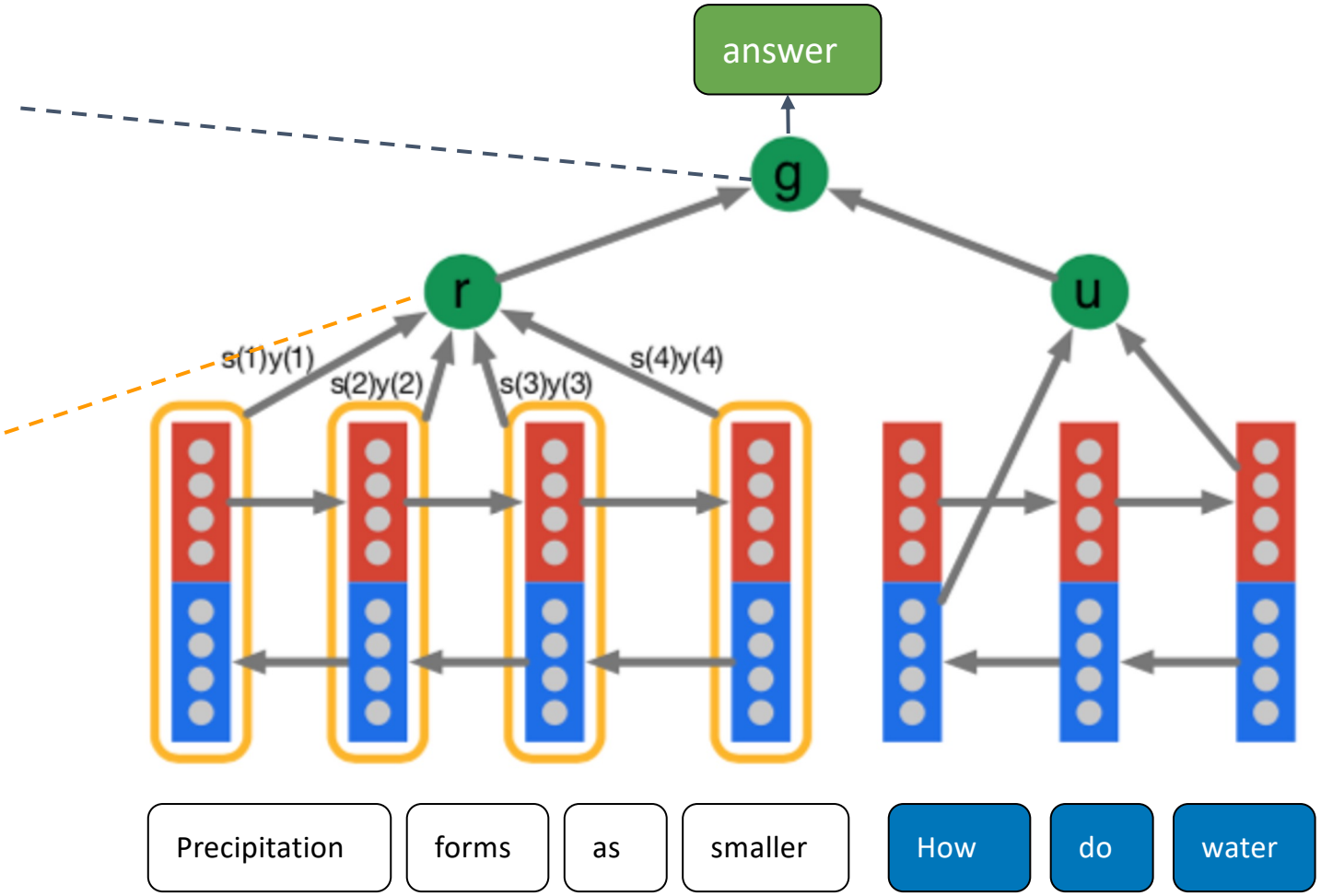
- **Why?** QA requires matching between question and text.
  - condition text representation on question (and vice versa)

- **“Naive approach”**: concatenation
  - append question after text, use RNN with longer sequence

- **Problem with naive approach:**
  - Long range dependencies: Many recurrent steps between answer and question → dilution of signal

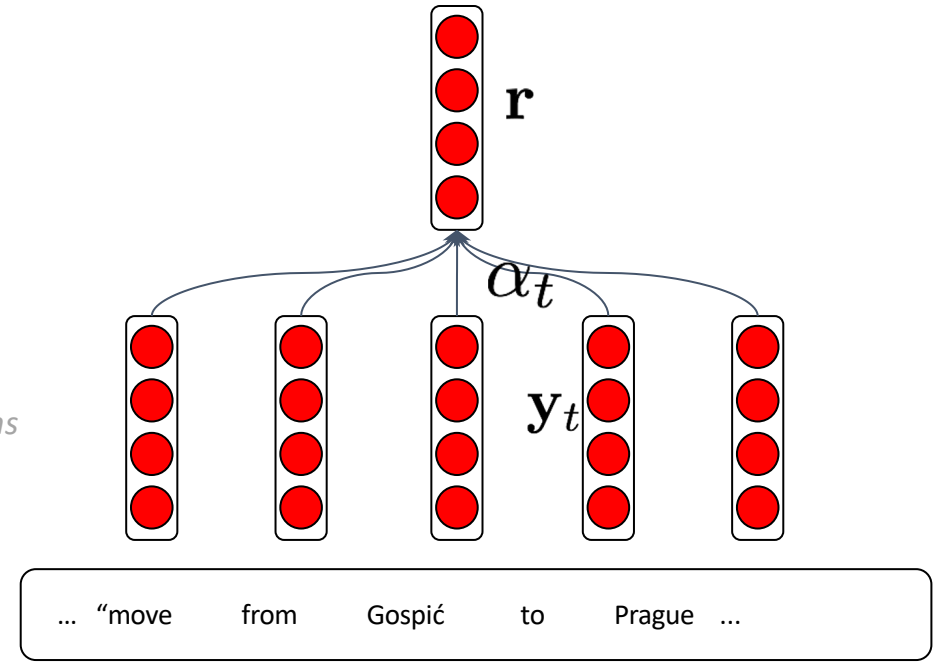
Modelling sequence interactions

Combination of question and text representation



Modelling sequence interactions: Attention

- **Attention:** relevance-weighted pooling of vectors across sequence
  - ! ΔśťĂŲţ ŦĂẅŋľŲŦ ǿP⁷⁄₈ŕ ľĂţ ĽśľŲţŕīĂŲţĂČŲţ ŴPśẅĂŲţ Ăţŕ ⁷⁄₈ŋ⁷⁄₈
  - 5ś⁷⁄₈ẁŦ īţśẅẅ₈ǤŋĄţľś Ųź⁷⁄₈ŋśţẅźŲẁĂţẅ śẁţĜ⁷⁄₈ś ŴPśẅĂŲţ

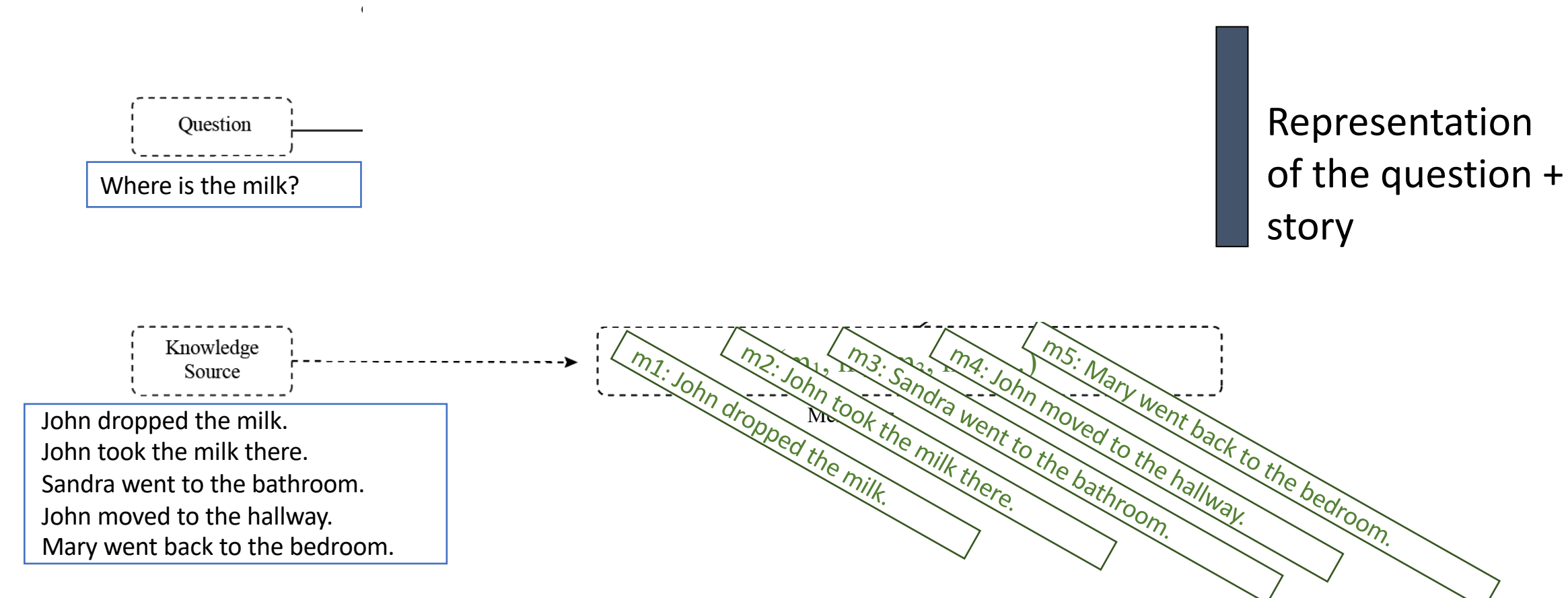


$$\mathbf{r} = \sum_{t=1}^{T} \alpha_t \mathbf{y}_t$$

$$\sum_{t=1}^{T} \alpha_t = 1; \quad \alpha_t \in [0, 1]$$

Modelling sequence interactions: Memory Networks

Sukhbaatar et al., NIPS'15 / Miller et al., EMNLP'16



Example: Learned attention patterns

Story (1: 1 supporting fact)	Support	Hop 1	Hop 2	Hop 3
Daniel went to the bathroom.		0.00	0.00	0.03
Mary travelled to the hallway.		0.00	0.00	0.00
John went to the bedroom.		0.37	0.02	0.00
John travelled to the bathroom.	yes	0.60	0.98	0.96
Mary went to the office.		0.01	0.00	0.00
Where is John? Answer: bathroom Prediction: bathroom				

Story (2: 2 supporting facts)	Support	Hop 1	Hop 2	Hop 3
John dropped the milk.		0.06	0.00	0.00
John took the milk there.	yes	0.88	1.00	0.00
Sandra went back to the bathroom.		0.00	0.00	0.00
John moved to the hallway.	yes	0.00	0.00	1.00
Mary went back to the bedroom.		0.00	0.00	0.00
Where is the milk? Answer: hallway Prediction: hallway				

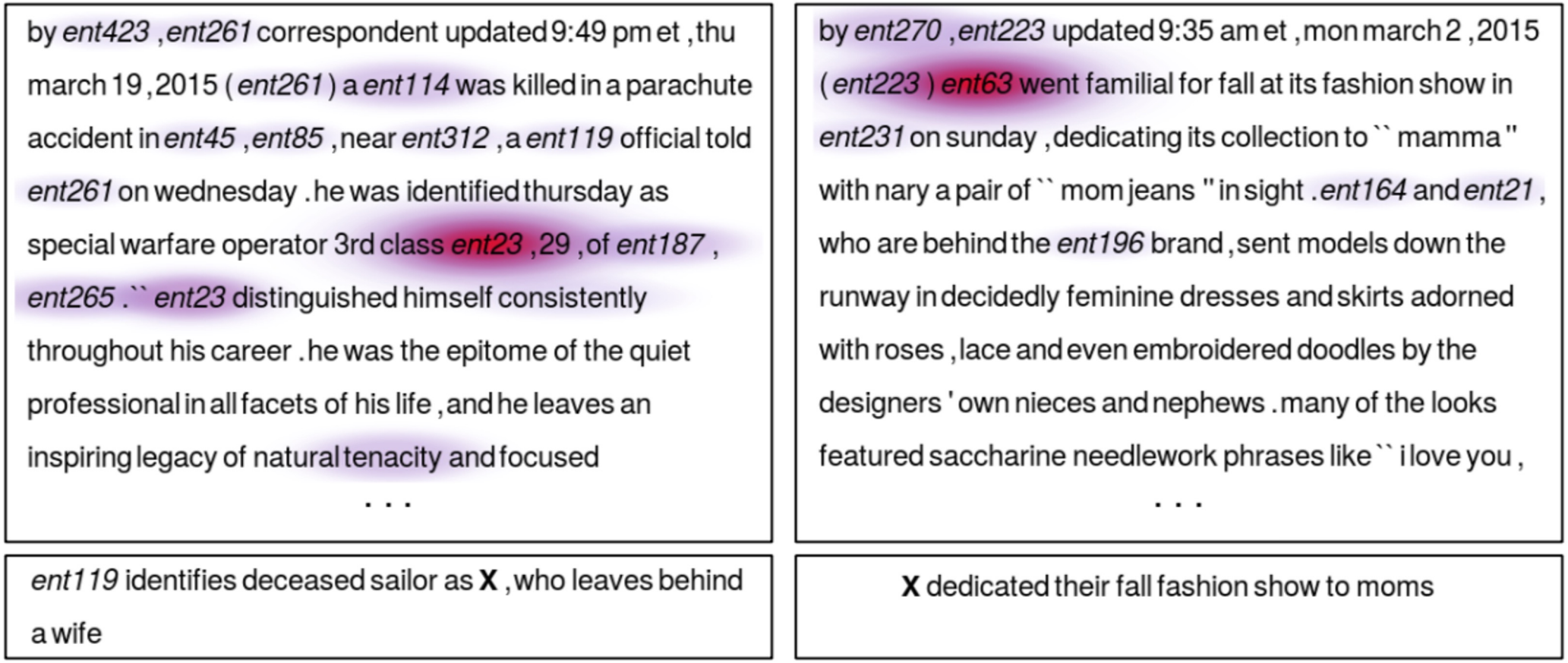
Story (16: basic induction)	Support	Hop 1	Hop 2	Hop 3
Brian is a frog.	yes	0.00	0.98	0.00
Lily is gray.		0.07	0.00	0.00
Brian is yellow.	yes	0.07	0.00	1.00
Julius is green.		0.06	0.00	0.00
Greg is a frog.	yes	0.76	0.02	0.00
What color is Greg? Answer: yellow Prediction: yellow				

Story (18: size reasoning)	Support	Hop 1	Hop 2	Hop 3
The suitcase is bigger than the chest.	yes	0.00	0.88	0.00
The box is bigger than the chocolate.		0.04	0.05	0.10
The chest is bigger than the chocolate.	yes	0.17	0.07	0.90
The chest fits inside the container.		0.00	0.00	0.00
The chest fits inside the box.		0.00	0.00	0.00
Does the suitcase fit in the chocolate? Answer: no Prediction: no				

End-to-end Memory Networks on bAbI tasks

Visualization from Sukhbaatar et. al., NIPS'15

Example: Learned attention patterns



Attentive Reader on QACNN/DailyMail dataset

Visualization from Hermann et. al. NIPS'15

Using self-attention

- Self-attention models like Transformers are now default to model sequence interactions.

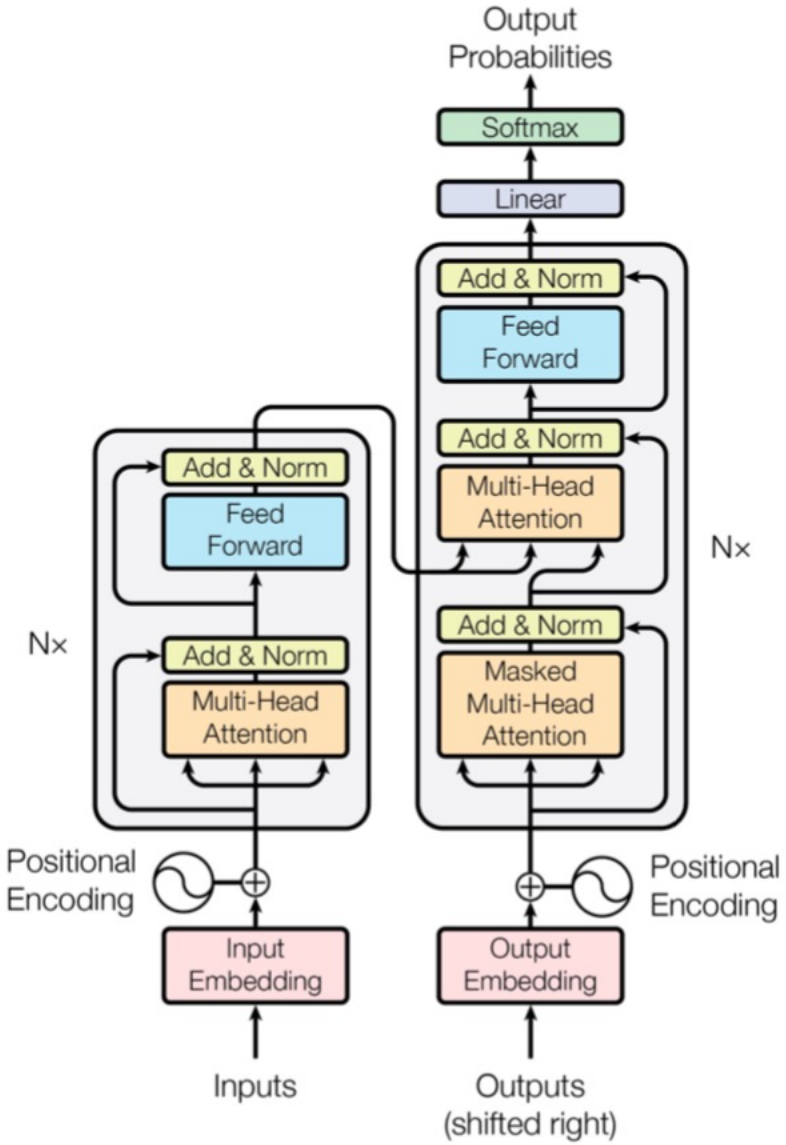
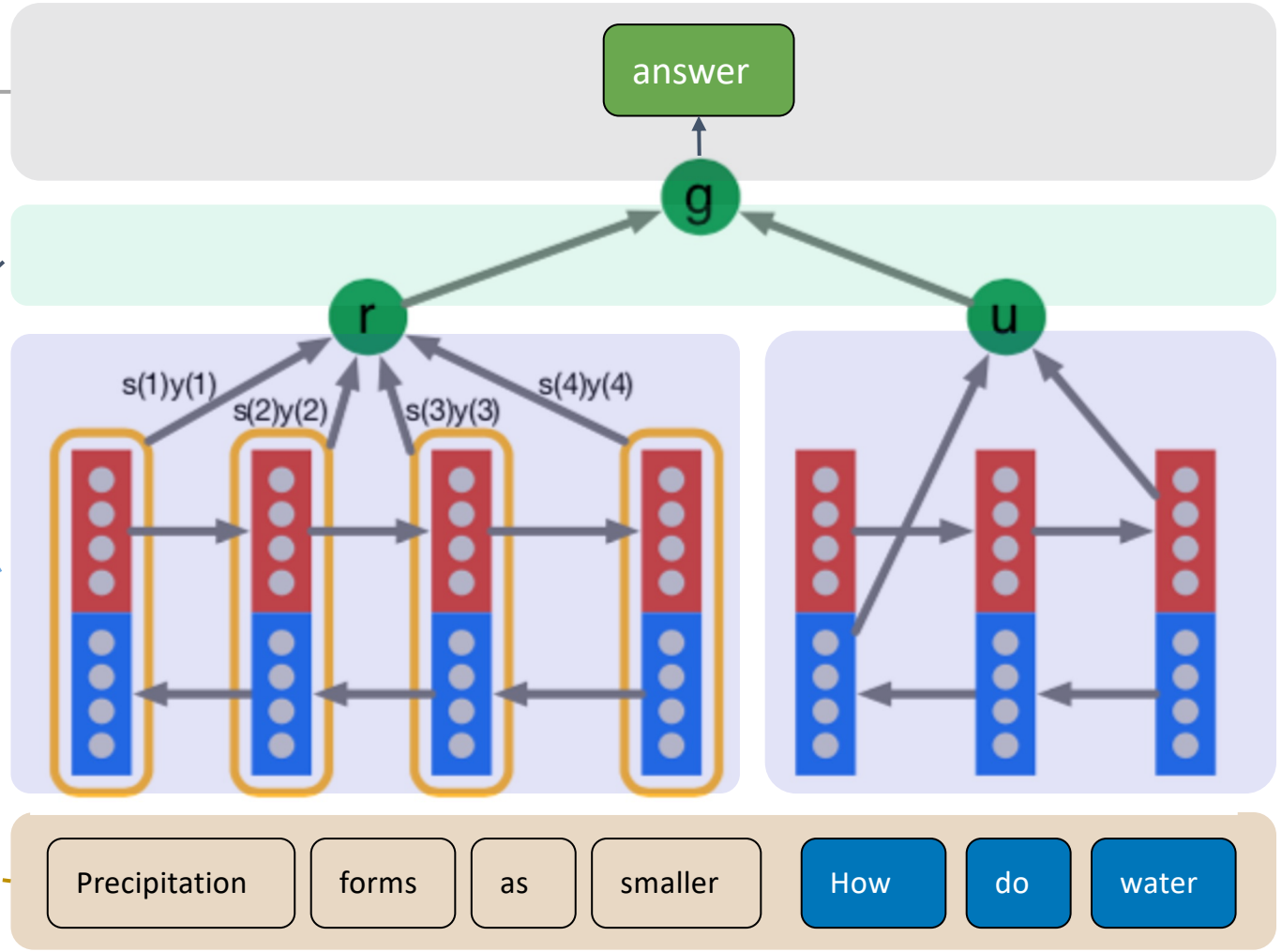


Figure from Vaswani et al., NIPS'17

The Attentive Reader Model: Overview

Answer Selection:
answer prediction

Sequence Interaction:
Matching text with question



Modified visualization from Hermann et. al. NIPS'15

Answer prediction

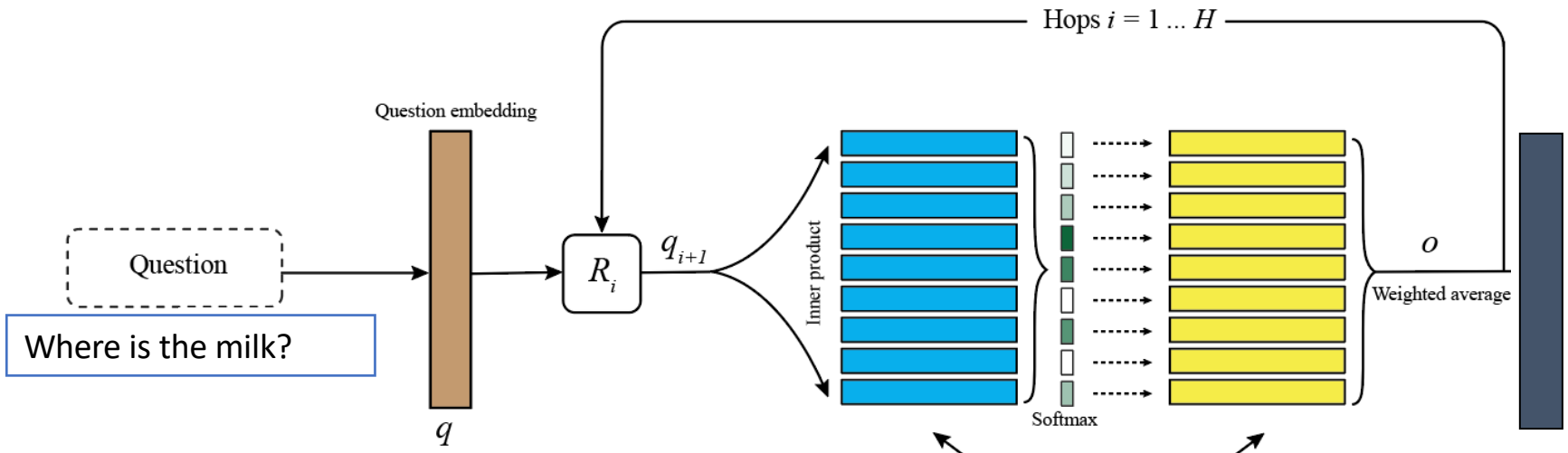
- Usually linear projection

- **Probability distribution over different answer options**
- Multiple choices: candidates (as in bAbI)
  - Spans in text -- distribution over positions for beginning and end (as in SQuAD)

- **Training:**
- Cross-entropy loss
- Ranking loss

Answer selection: Ranking (Memory Networks)

Sukhbaatar et al., NIPS'15 / Miller et al., EMNLP'16



Knowledge Source

John dropped the milk.
John took the milk there.
Sandra went to the bathroom.
John moved to the hallway.
Mary went back to the bedroom.

m1: John dropped the milk.
m2: John took the milk there.
m3: Sandra went to the bathroom.
m4: John moved to the hallway.
m5: Mary went back to the bedroom.

milk
hallway
office
...

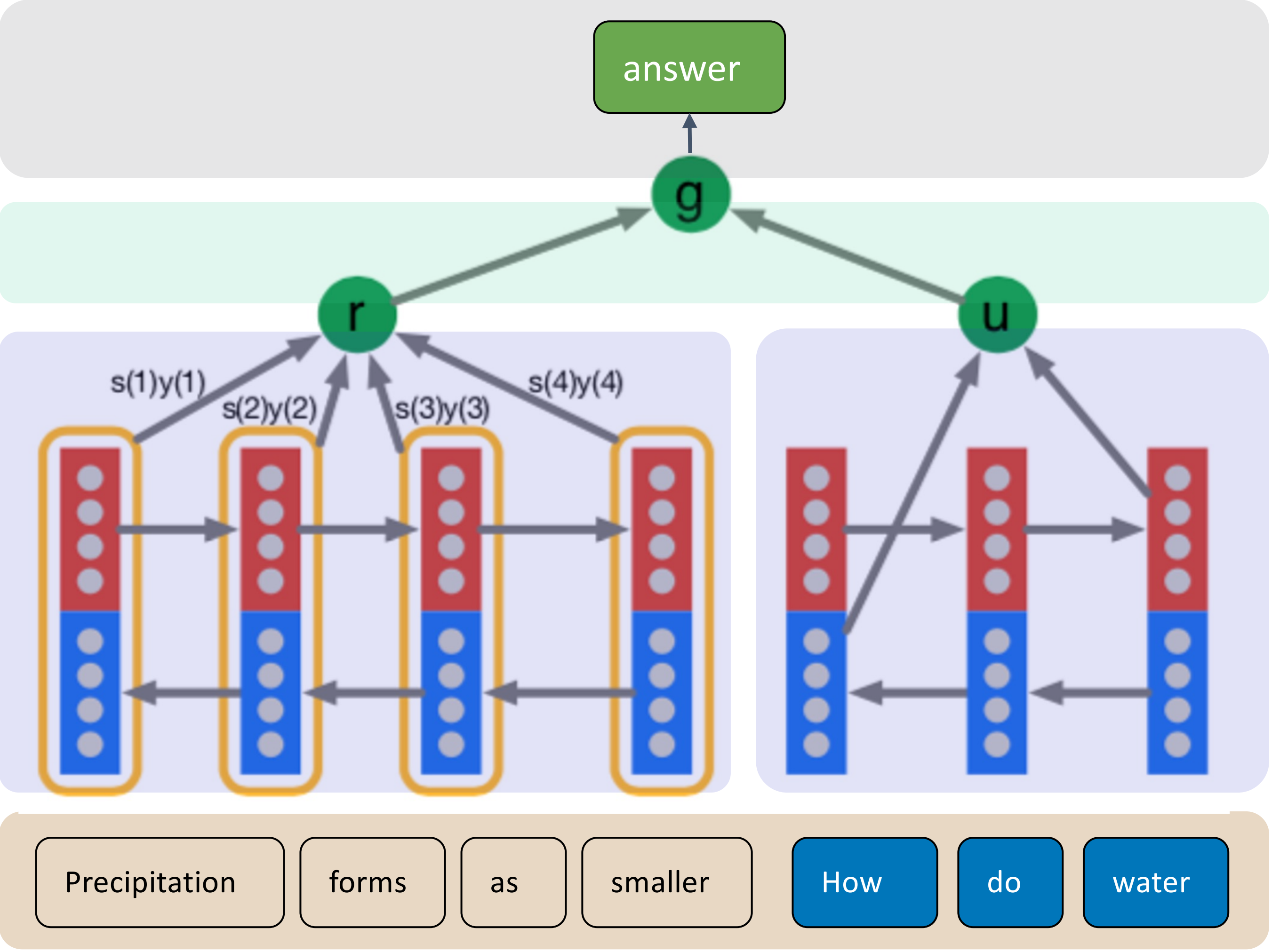
The Attentive Reader Model: Overview

Answer Selection:
answer prediction

{ŚẀPś ŀ l'ś Ŀ ẮśẮ'AŲ ő
a ĂẮĨ ŀ Ĝ ẮśŋẮ ĩẮ
ẀPśẅAŲ

/ŲF øŲẅAŲ ő
ŀ l'ŲẃŲẅĂ Ĝl'Ų ẮśŋẮ
ĂẁŲP ŕ į Ųẃ ẅ

Input: Representing symbols as vectors



Modified visualization from Hermann et. al. NIPS'15

Conclusion

- We gathered all ingredients to build state-of-the-art supervised Machine Reading systems!

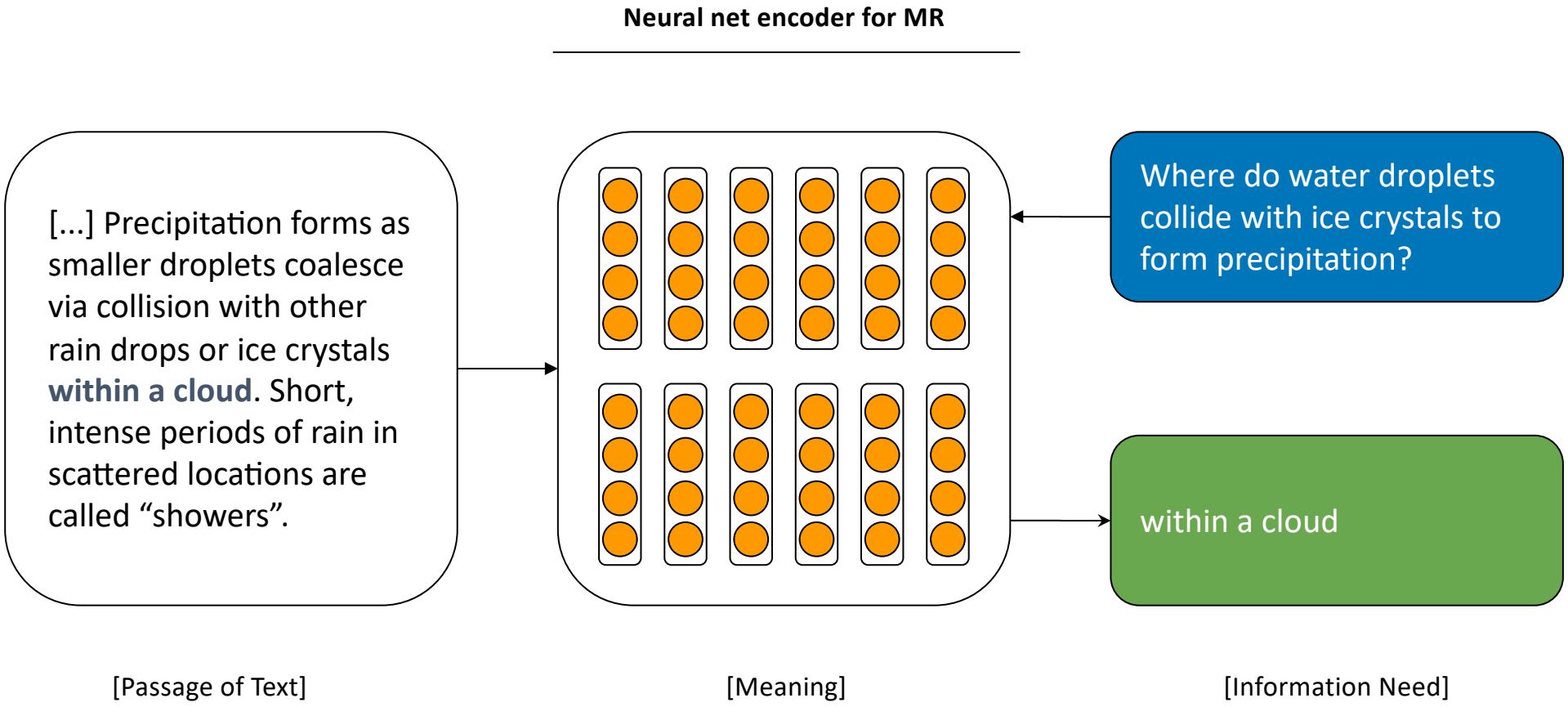
- architectures work well in practice
  ... as long as we stay in-domain and questions are simple

- We covered only extractive and multiple choice questions settings but there is also generative machine reading

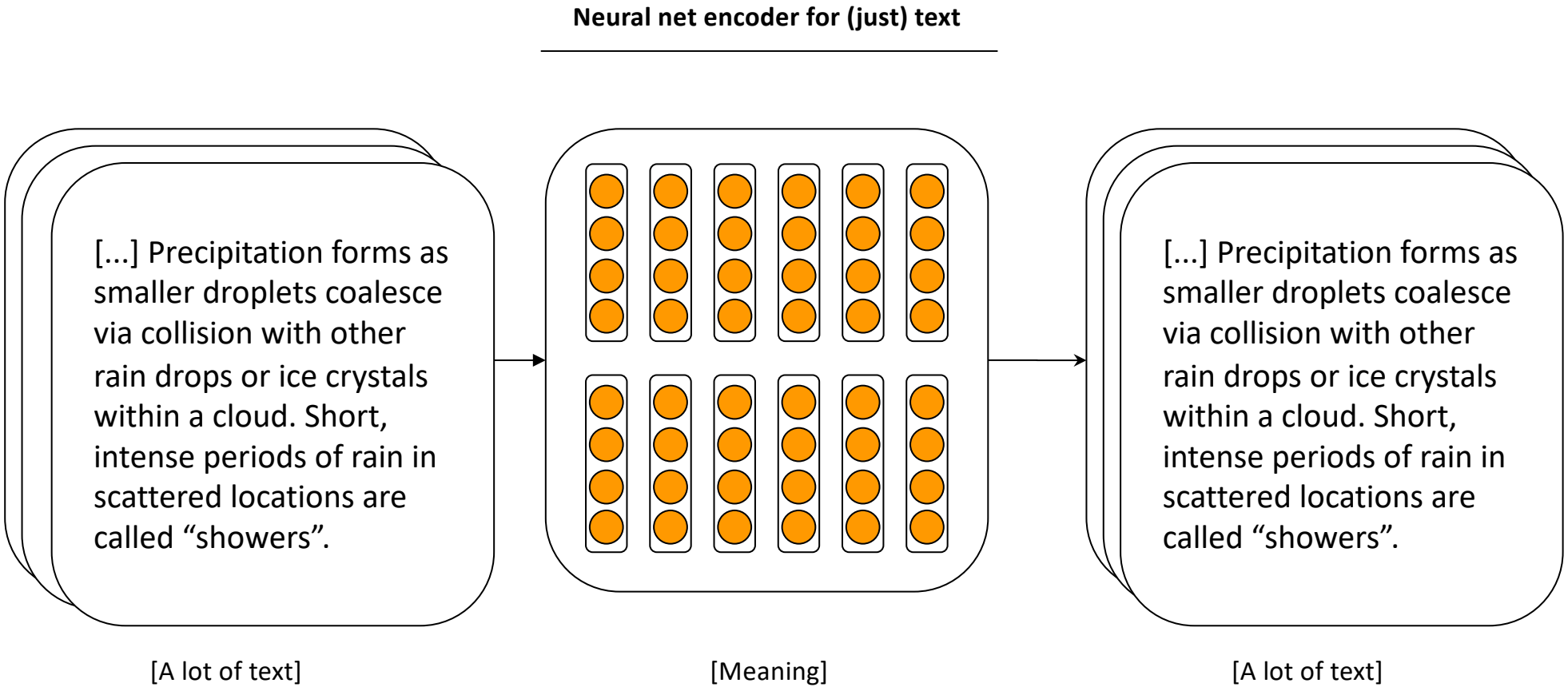
- Practice in Labs on bAbI!

Machine Reading / Current Trend

Supervised training



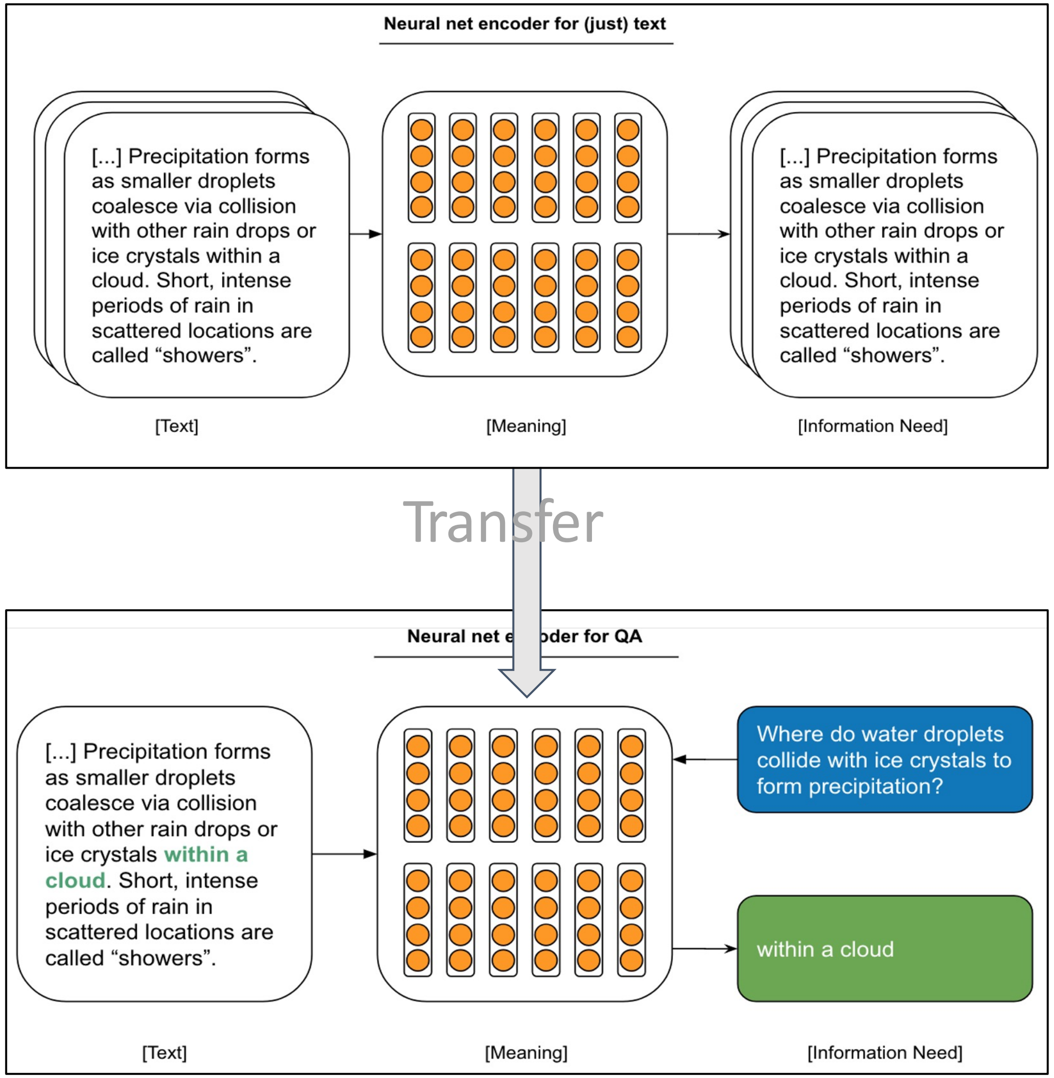
Unsupervised pretrained representations



Lifting over pretrained representations

tvš7ǍĂtśŕ
[ĂtĜPĂĜś a Ųŕ śŎ

Machine Reading



How is this different from pretrained word embeddings?

Pretrained Word Embeddings (word2vec)

- Predicting co-occurring of words
- Independent of other context

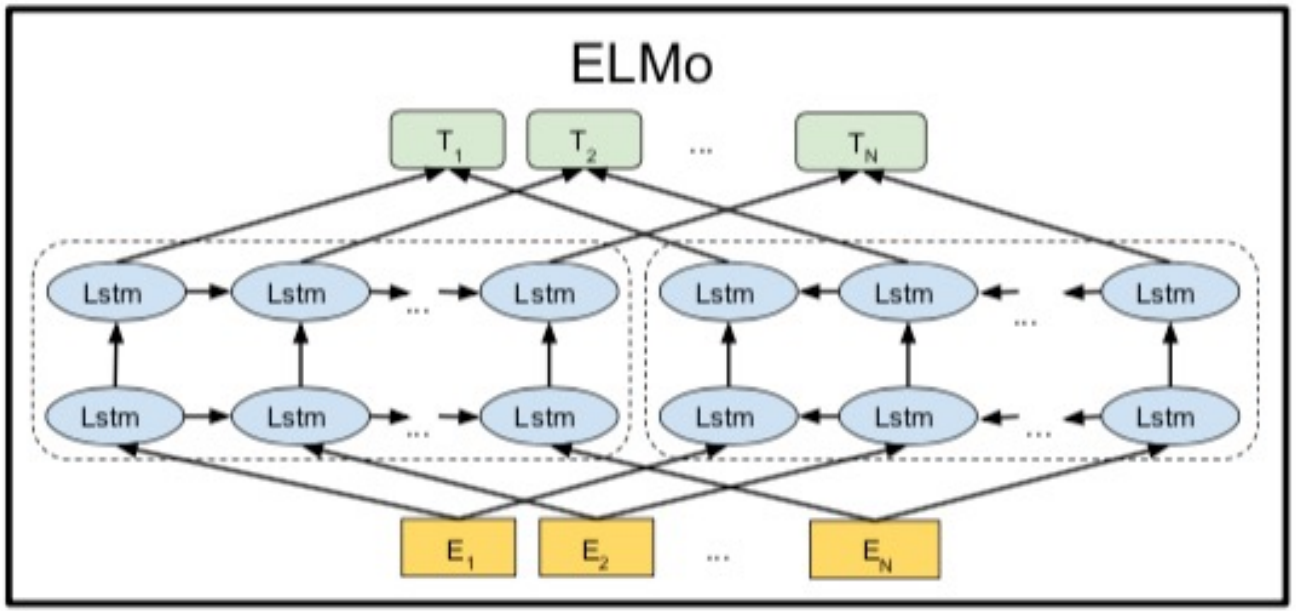
Pretrained Contextualized Embeddings (e.g. ELMo, BERT)

- Predicting whole text (using LSTM, or Self-Attention)
- Full dependence on other context

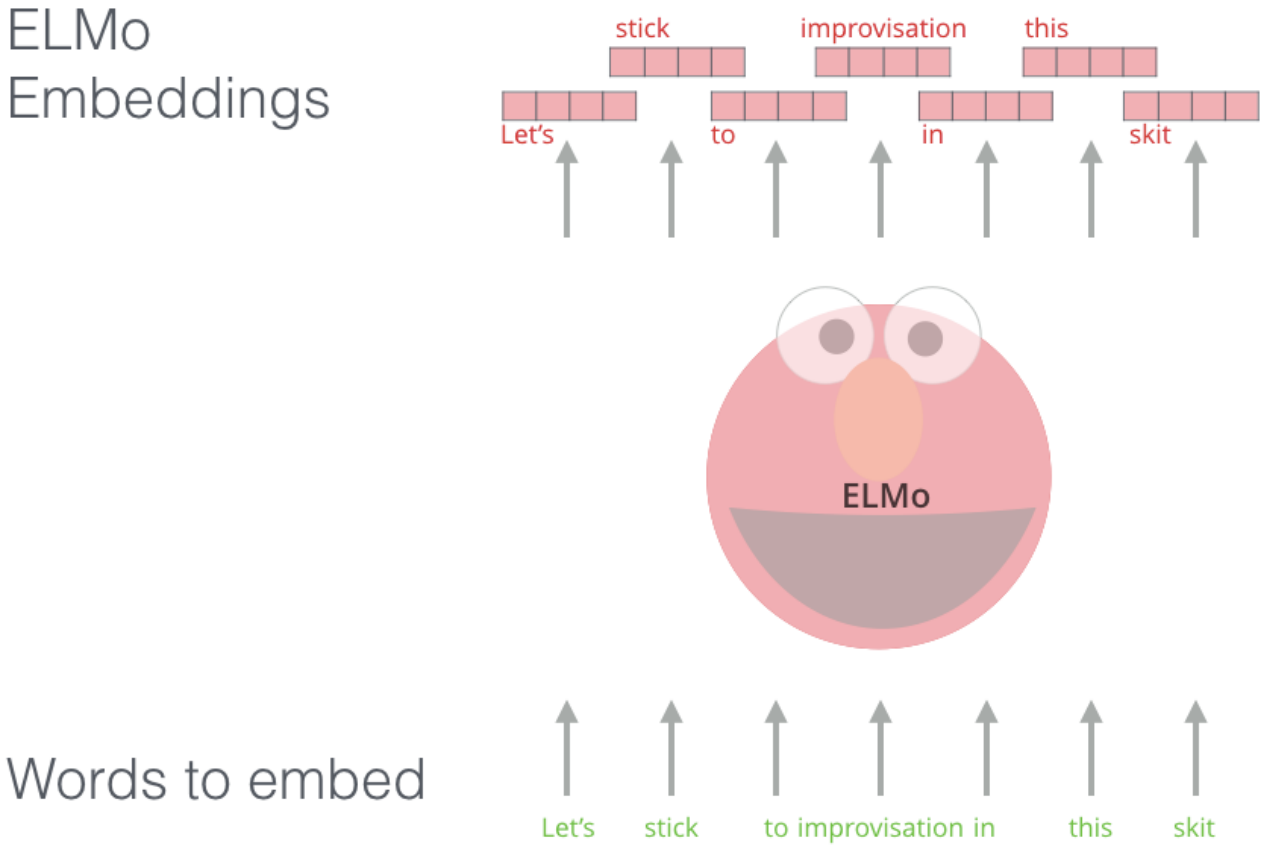
ELMo: Embeddings from Language Models

Peters et al., NAACL'18

- Train a BiLSTM for Bidirectional language modeling on a large dataset
- Run the sentence to encode through both forward and backward LSTMs
- Combine fc al contextual embedding

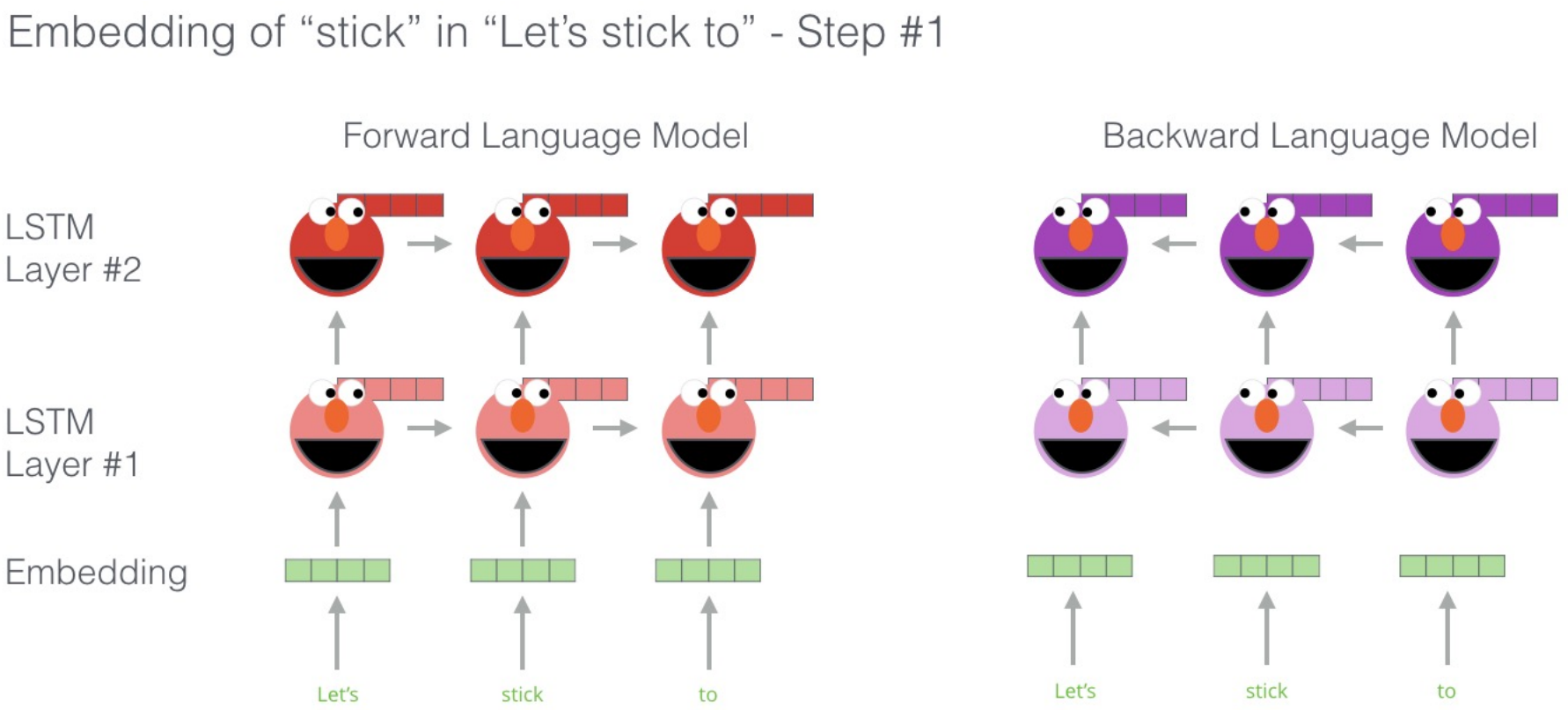


ELMo: **E**mbeddings from **L**anguage **Mo**dels



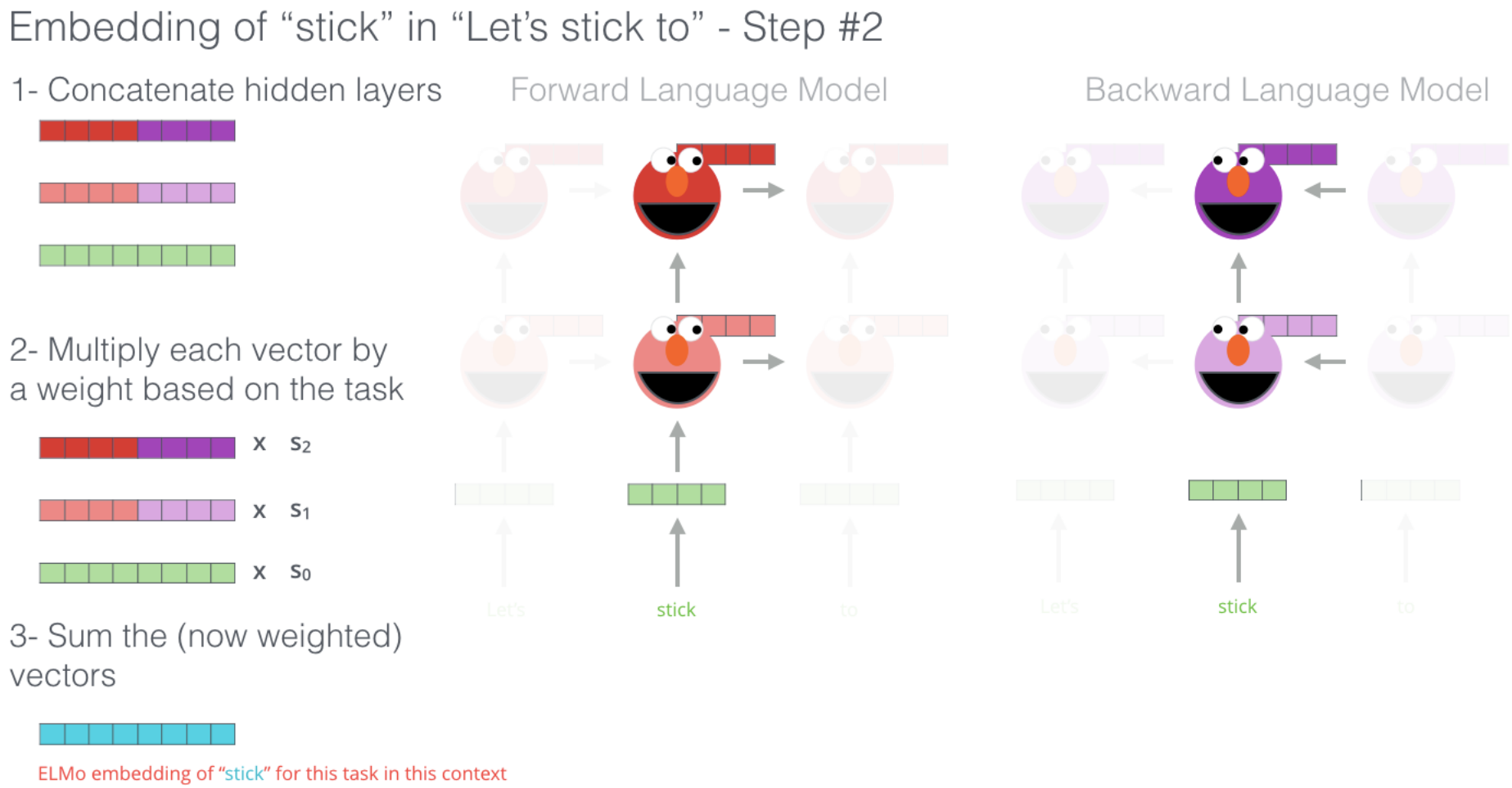
Figures from http://jalammar.github.io/illustrated-bert/

ELMo: **E**mbeddings from **L**anguage **Mo**dels



Figures from http://jalammar.github.io/illustrated-bert/

ELMo: Embeddings from Language Models



Figures from http://jalammar.github.io/illustrated-bert/

ELMo performance

	Task	Previous SOTA		Our baseline	ELMo + baseline	Increase (absolute/ relative)
Machine Reading-	SQuAD	Liu et al. (2017)	84.4	81.1	85.8	4.7 / 24.9%
Textual Entailment -	SNLI	Chen et al. (2017)	88.6	88.0	88.7 ± 0.17	0.7 / 5.8%
Semantic Labeling -	SRL	He et al. (2017)	81.7	81.4	84.6	3.2 / 17.2%
Coreference Resolution -	Coref	Lee et al. (2017)	67.2	67.2	70.4	3.2 / 9.8%
Entity Extraction -	NER	Peters et al. (2017)	91.93 ± 0.19	90.15	92.22 ± 0.10	2.06 / 21%
Sentiment Analysis -	SST-5	McCann et al. (2017)	53.7	51.4	54.7 ± 0.5	3.3 / 6.8%

What is ELMo learning ?

- Meaning of words in context
  - POS, word sense, etc.

	Source	Nearest Neighbors
GloVe	play	playing, game, games, played, players, plays, player, Play, football, multiplayer
biLM	Chico Ruiz made a spectacular play on Alusik 's grounder {...}	Kieffer , the only junior in the group , was commended for his ability to hit in the clutch , as well as his all-round excellent play .
	Olivia De Havilland signed to do a Broadway play for Garson {...}	{...} they were actors who had been handed fat roles in a successful play , and had talent enough to fill the roles competently , with nice understatement .

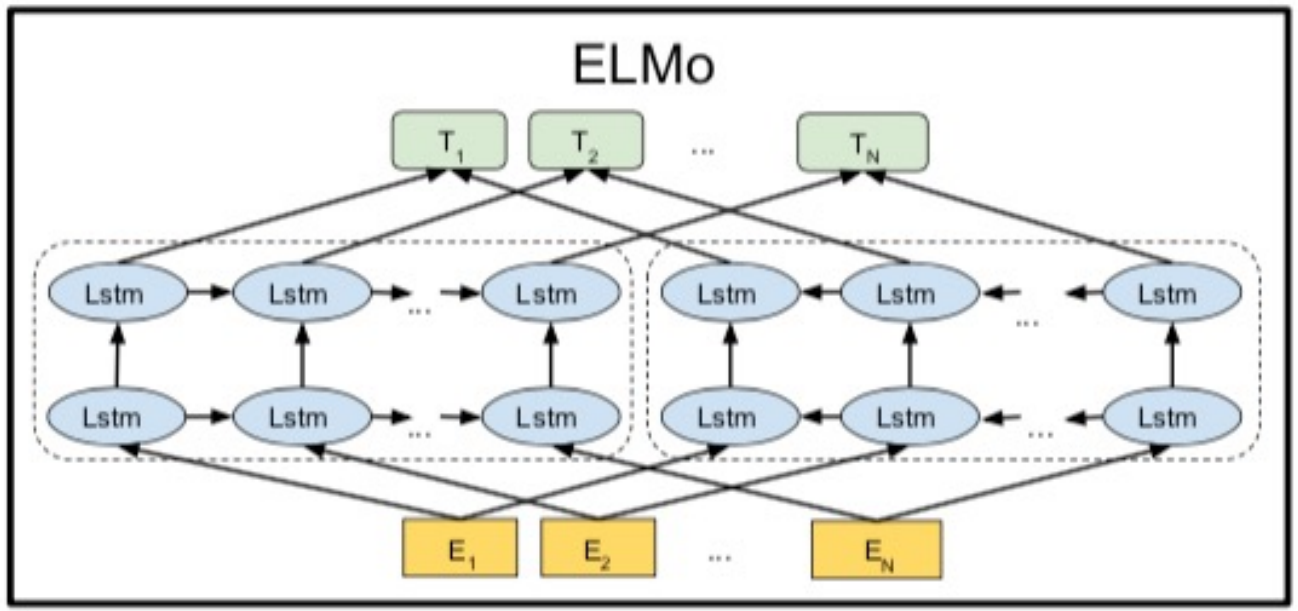
Problems with ELMo

- Need to use different architectures for different tasks
- Retraining models is slow, transfer learning is fast
- Need to deal with long term dependencies in LSTMs!

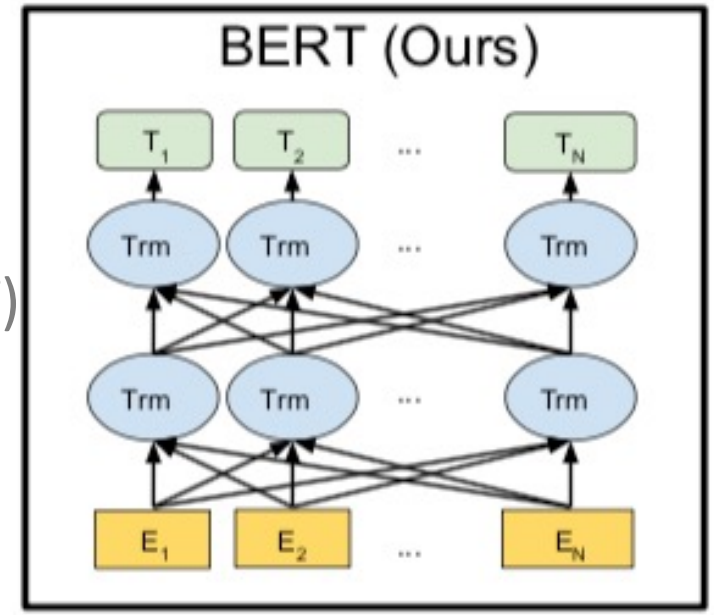
BERT - **B**idirectional **E**ncoder **R**epresentations from **T**ransformers

Devlin et al., NAACL'19

Solutions: use Transformer + encoder layers instead of decoder layers



(OpenAI GPT)

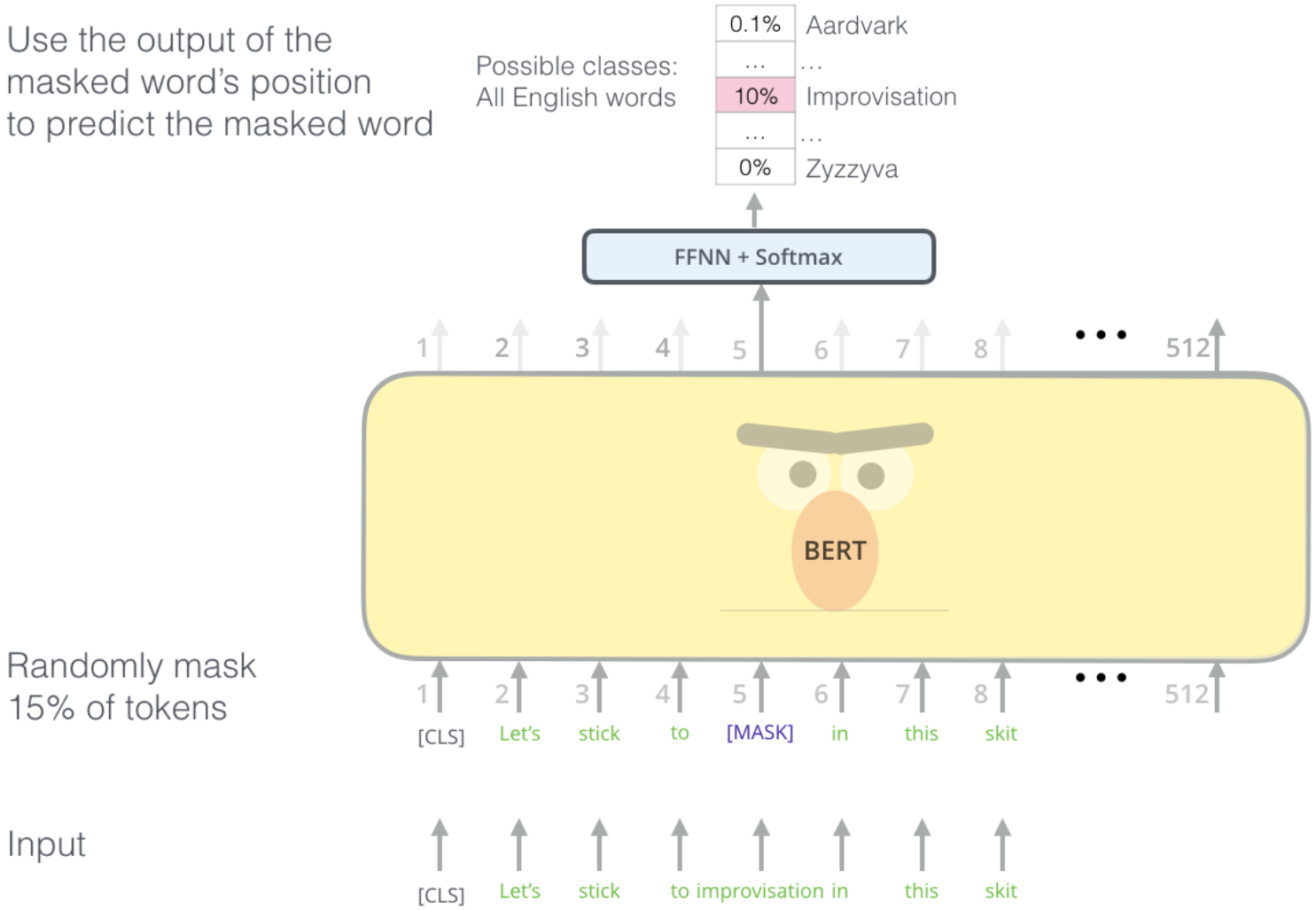


Innovation with multiple pretraining tasks

BERT – Pretraining 1: masked language modeling

- Given a sentence with some words masked at random, can we predict them?
- Randomly select 15% of tokens to be replaced with “<MASK>”

BERT – Pretraining 1: masked language modeling

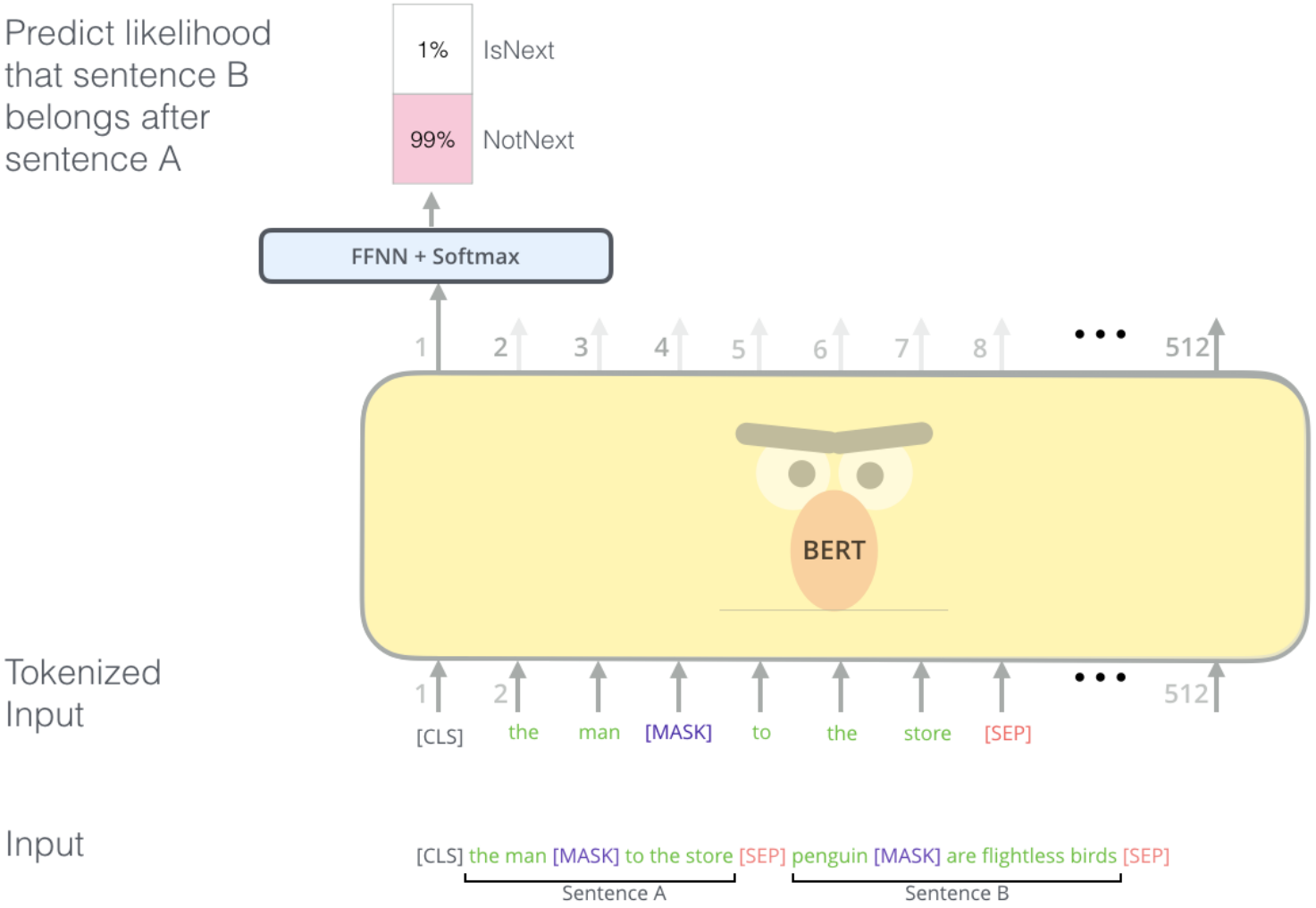


Figures from http://jalammar.github.io/illustrated-bert/

BERT – Pretraining 2: next sentence prediction

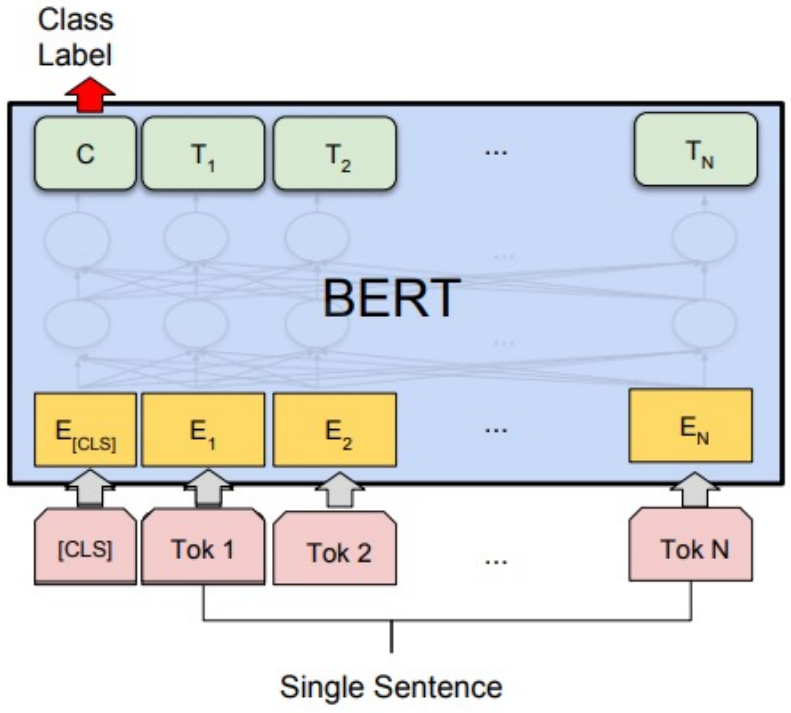
- Given two sentences, does the first follow the second?
- Teaches BERT about relationship between two sentences
- 50% of the time the actual next sentence, 50% random

BERT – Pretraining 2: next sentence prediction



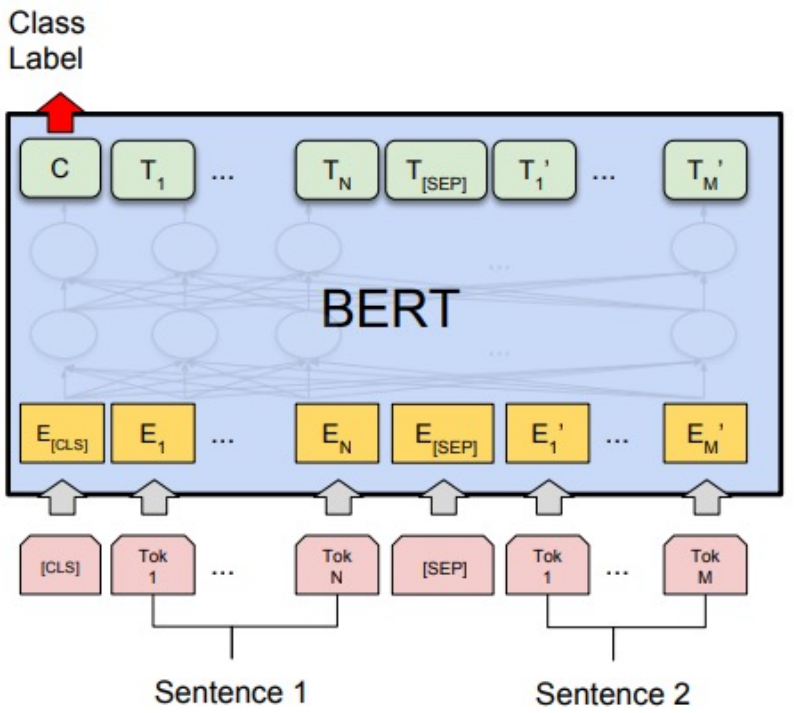
Figures from http://jalammar.github.io/illustrated-bert/

BERT – Fine-tuning for Classification



Single sentence classification

Sentiment analysis, spam detection, etc.

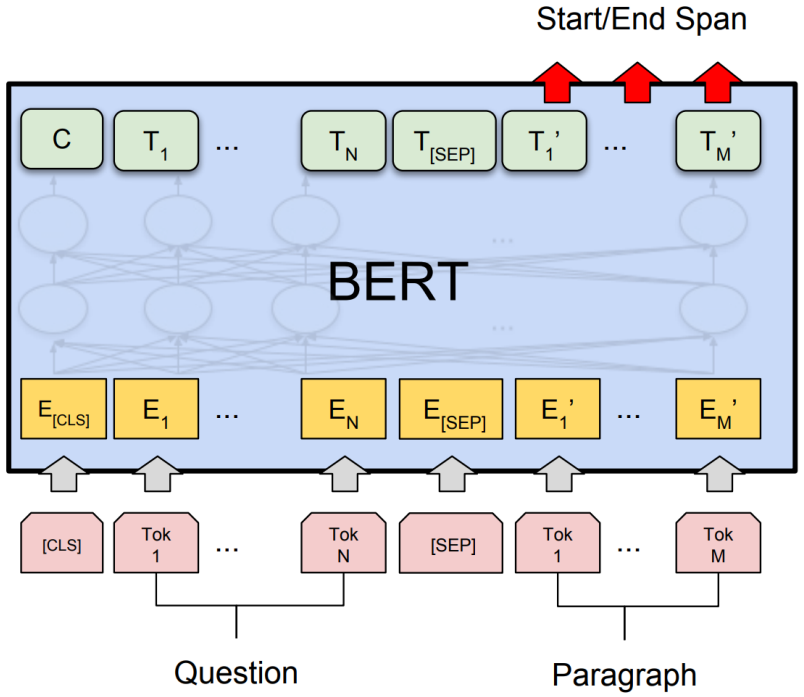


Pair of sentences classification

Entailment, paraphrase detection, etc.

Figures from Devlin et al. 18'

BERT – Fine-tuning for Machine Reading



(c) Question Answering Tasks: SQuAD v1.1

System	Dev EM	Dev F1	Test EM	Test F1
Leaderboard (Oct 8th, 2018)				
Human	-	-	82.3	91.2
#1 Ensemble - nlnet	-	-	86.0	91.7
#2 Ensemble - QANet	-	-	84.5	90.5
#1 Single - nlnet	-	-	83.5	90.1
#2 Single - QANet	-	-	82.5	89.3
Published				
BiDAF+ELMo (Single)	-	85.8	-	-
R.M. Reader (Single)	78.9	86.3	79.5	86.6
R.M. Reader (Ensemble)	81.2	87.9	82.3	88.5
Ours				
$\text{BERT}_{\text{BASE}}$ (Single)	80.8	88.5	-	-
$\text{BERT}_{\text{LARGE}}$ (Single)	84.1	90.9	-	-
$\text{BERT}_{\text{LARGE}}$ (Ensemble)	85.8	91.8	-	-
$\text{BERT}_{\text{LARGE}}$ (Sgl.+TriviaQA)	84.2	91.1	85.1	91.8
$\text{BERT}_{\text{LARGE}}$ (Ens.+TriviaQA)	**86.2**	**92.2**	**87.4**	**93.2**

Figures from Devlin et al. 18'

RoBERTa / ALBE

Derivatives from BERT

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Mar 12, 2020	ALBERT + DAAF + Verifier (ensemble) *PINGAN Omni-Sinitic*	**90.386**	**92.777**
2 Jan 10, 2020	Retro-Reader on ALBERT (ensemble) *Shanghai Jiao Tong University* http://arxiv.org/abs/2001.09694	90.115	92.580
3 Nov 06, 2019	ALBERT + DAAF + Verifier (ensemble) *PINGAN Omni-Sinitic*	90.002	92.425
4 Sep 18, 2019	ALBERT (ensemble model) *Google Research & TTIC* https://arxiv.org/abs/1909.11942	89.731	92.215
4 Feb 25, 2020	Albert_Verifier_AA_Net (ensemble) *QIANXIN*	89.743	92.180
5 Jan 23, 2020	albert+transform+verify (ensemble) *qianxin*	89.528	92.059
6 Mar 06, 2020	ALBert-LSTM (ensemble) *oppo.tensorlab*	89.269	91.777
7 Dec 08, 2019	ALBERT+Entailment DA (ensemble) *CloudWalk*	88.761	91.745

RTa

Models	e/High)
Single model (from lead	
BERT-large	1)
XLNet	2)
RoBERTa	3)
UPM	
XLNet + SG-Net Verifi	
ALBERT (1M)	1)
ALBERT (1.5M)	**5)**
Ensembles (from leader	
BERT-large	
XLNet + SG-Net Verifi	
UPM	
XLNet + DAAF + Verif	
DCMN+	3)
ALBERT	**6)**

+5%