Machine Reading, Question Answering & Dialog

AMMI – Deep NLP – Part III

Angela Fan, Louis Martin, Antoine Bordes

Facebook AI Research (Paris)

March 18-22, 2019

Who are we?

Angela
angelafan@fb.com
PhD Student FAIR Paris

Louis
louismartin@fb.com
PhD Student FAIR Paris

Antoine
abordes@fb.com
Director FAIR Paris

+ Help from **Sebastian Riedel** (Research Scientist FAIR London)

This Class:

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

Quick schedule

- Monday: 2pm lecture “Machine Reading” [Antoine]
- Tuesday: 9am labs “Machine Reading” [Angela/Louis] + office hours 11m [All]
- Wednesday: 2pm lecture “Question Answering” [Antoine] + office hours 4pm [All]
- Thursday: 9am lecture “Dialog” [Antoine] + 11:15am exam [All]
- Friday: 2pm labs “Dialog” [Angela/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

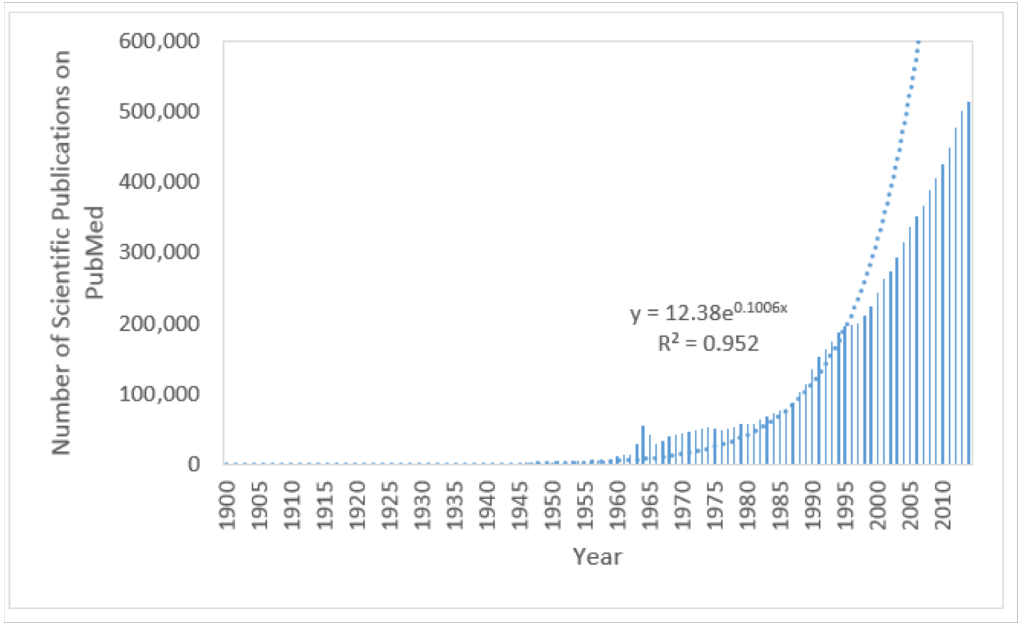
BUT THERE HAS BEEN A LOT OF PROGRESS AND MACHINE READING RESEARCH ACTIVITY HAS SKYROCKETED

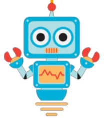
BY ANTHONY CUTHBERTSON ON 1/15/18 AT 8:00 AM

Main (big) motivation

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

Motivation 1: Information Overload



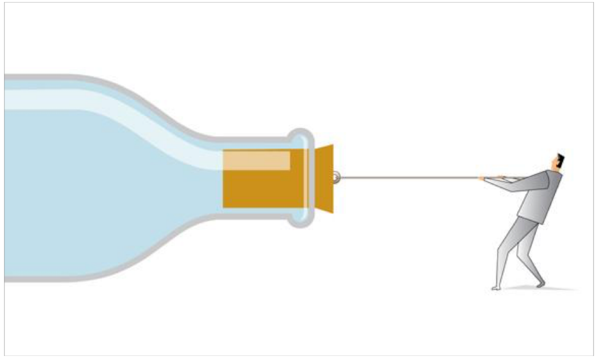


uses for

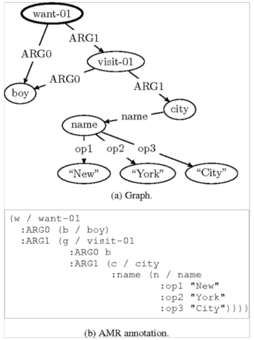
[Information Need]

Motivation 2: The Knowledge Acquisition Bottleneck

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

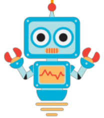


[Knowledge]



(a) Graph.
(b) AMR annotation.

[Meaning]



uses for

[Information Need]

Applications: Question Answering

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.

?

What city did Tesla move to in 1880?

Prague

[Text]

[Meaning]

[Information Need]

Applications: Helping Agents to learn Faster

Branavan et al., JMLR'12

The natural resources available where a population settles affects its ability to produce food and goods. Build your city on a plains or grassland square with a river running through it if possible.

?

[Text]

[Meaning]

[Information Need]

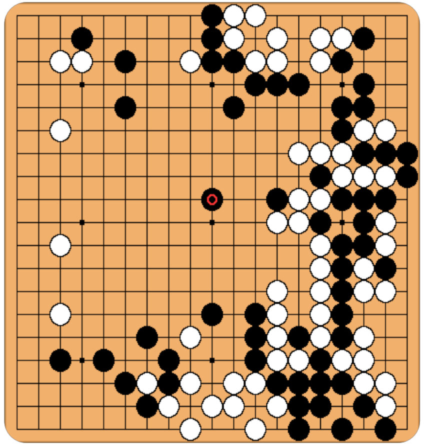
Applications: Helping Agents to learn Faster

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]

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]

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

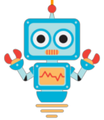
Machine Reading

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

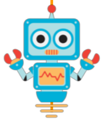
[Information Need]

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.

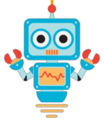
?

[Passage of Text]



converts into

[Meaning]

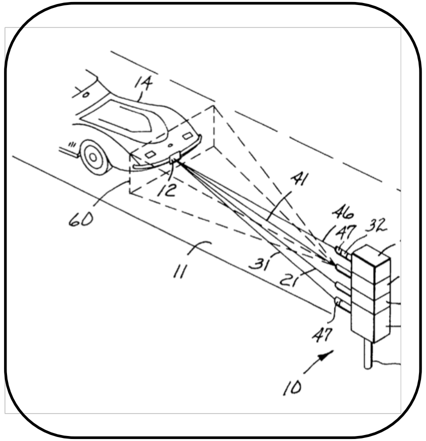


uses for

[Information Need]

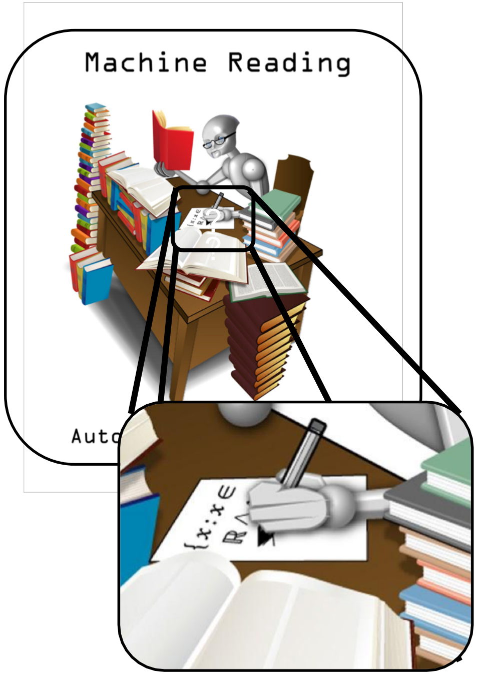
Timeline of Machine Reading

Something else
entirely!

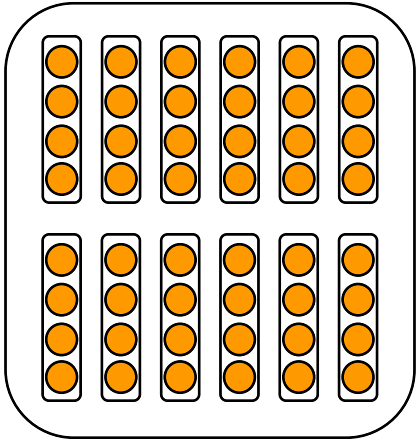


before 2006

Text to Meaning
Representations

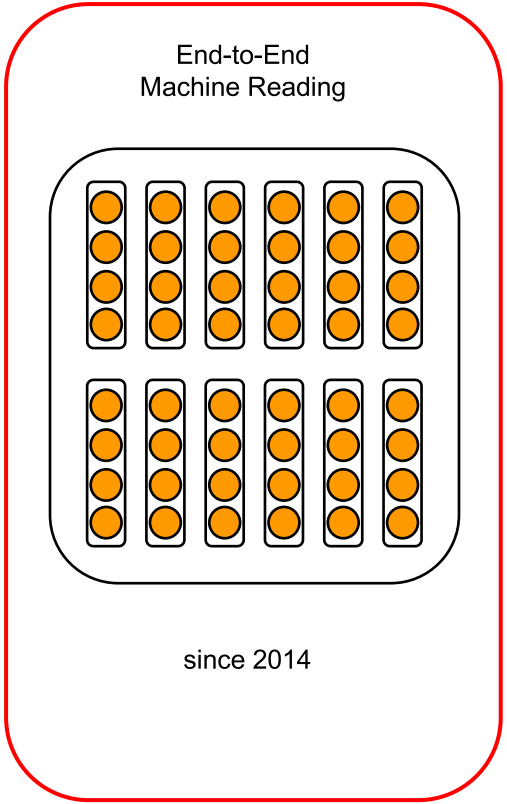


End-to-End
Machine Reading



since 2014

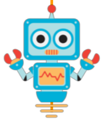
Today we cover:



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.

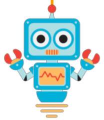
[Passage of Text]



converts into

?

[Meaning]



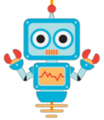
uses for

[Information Need]

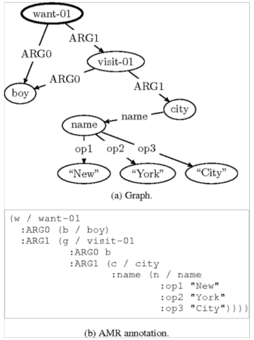
Symbolic Approaches (until 2014 or so)

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.

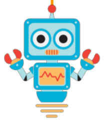
[Passage of Text]



converts into



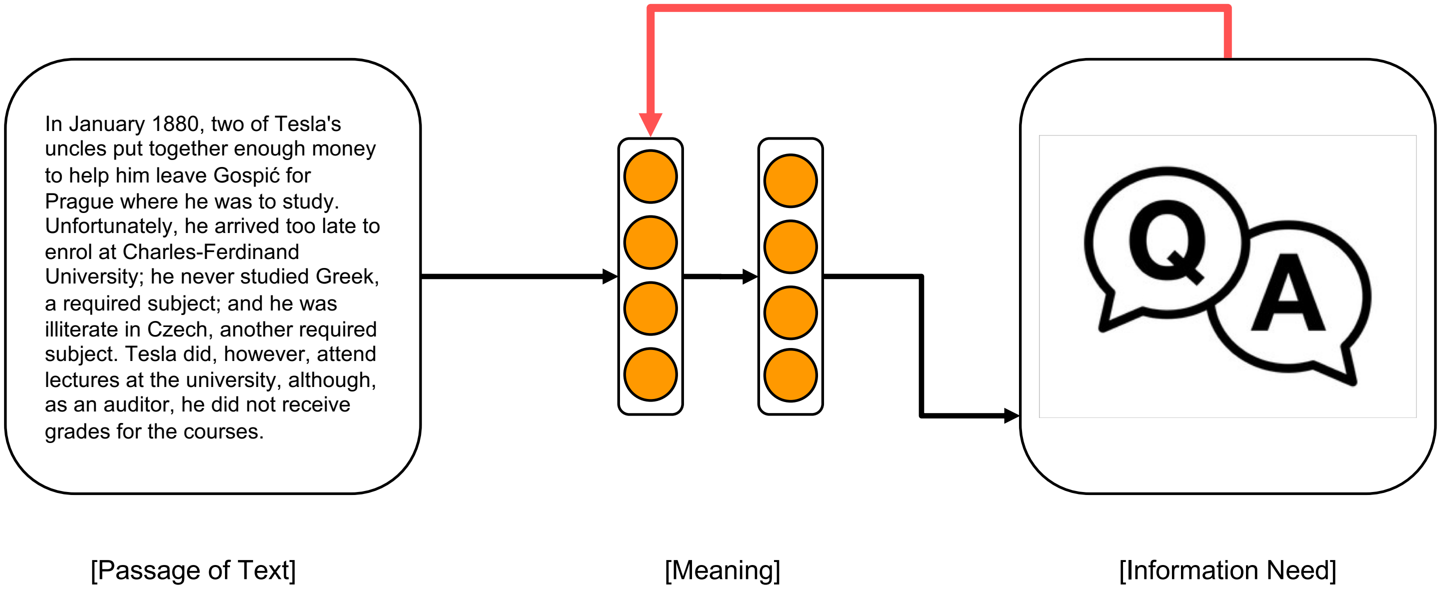
[Meaning]



uses for

[Information Need]

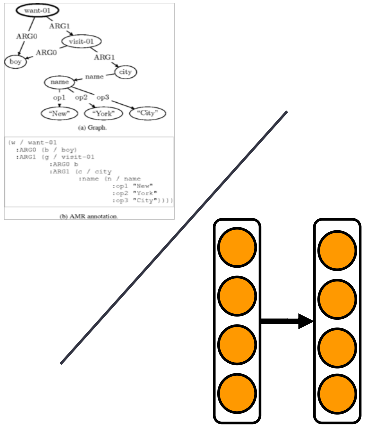
End-to-End Approaches (since 2014 or so)



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]



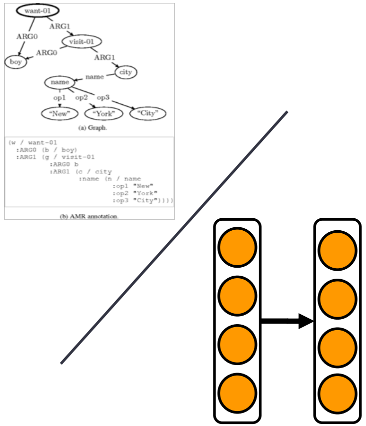
[Meaning]

- 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.

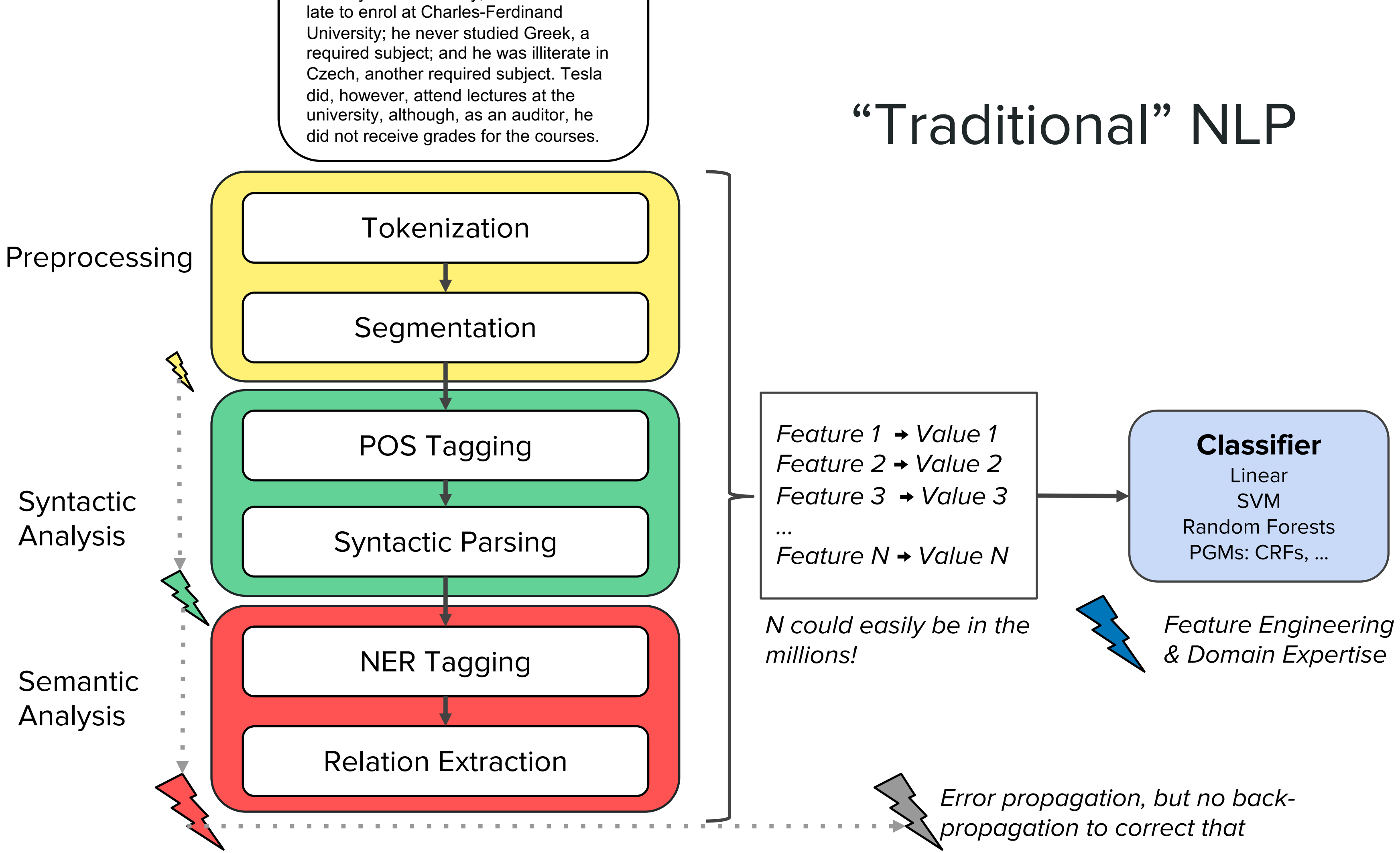
[Passage of Text]



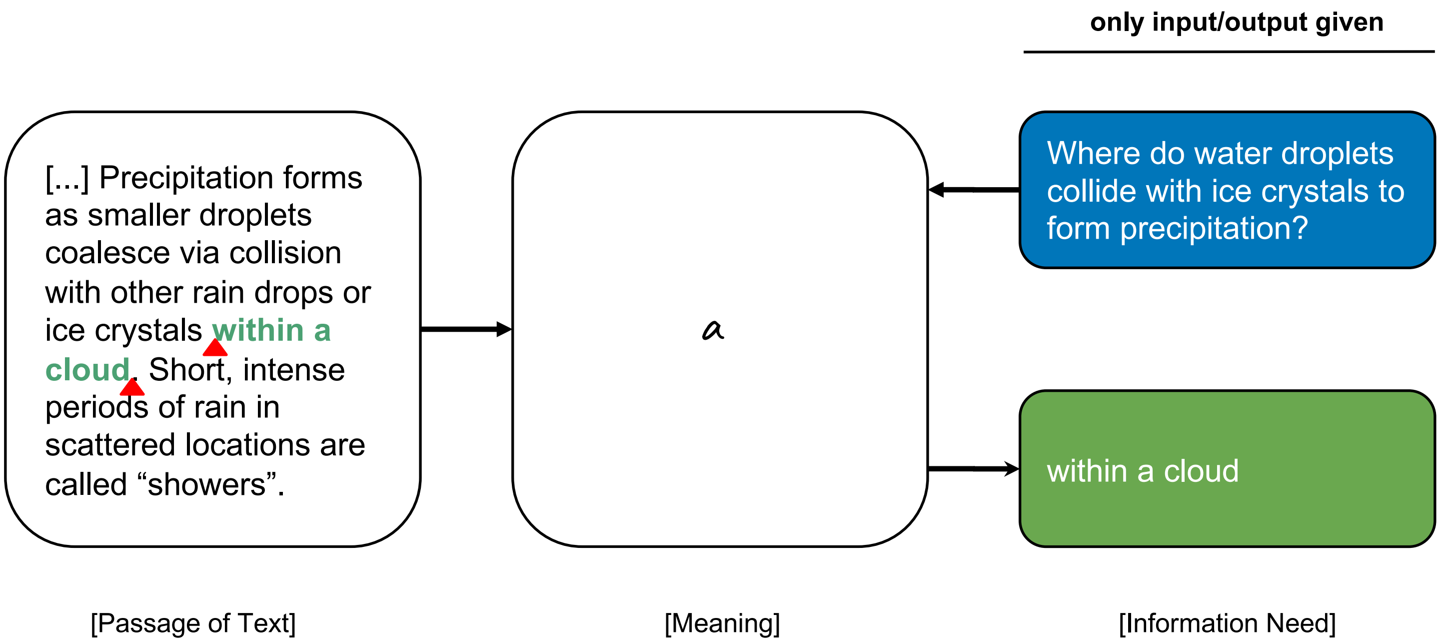
[Meaning]

- Ambiguity
- Variation
- Coreference
- Common Sense
- Scale
- ...

“Traditional” NLP



End-to-end System



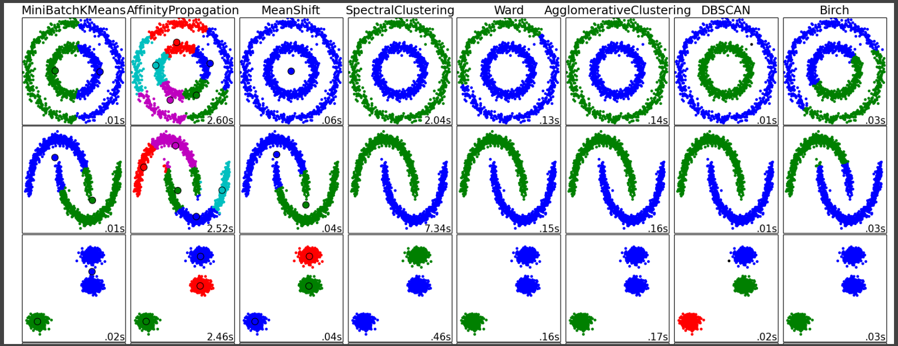
Machine Reading / Data

Limits of Big Model + Big Data

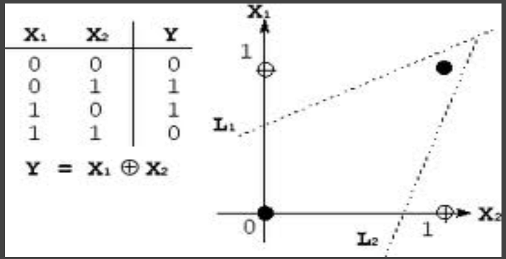
- End-to-end systems need a lot of cleaned data!
- Getting real conditions large-scale data is difficult:
  - Real large-scale data is complex, noisy, unlabeled...
  - Interpretation of success or failure is complex
- Complicates the design of innovative learning systems
- Some recent process with large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

A long history of artificial problems in ML

Two moons and friends (clustering)



XOR (neural networks)



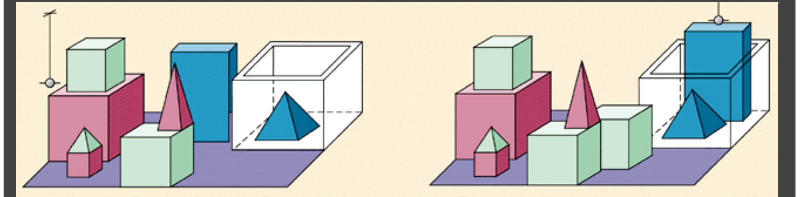
Many in the UCI repository (regression, classification)

UCI
Machine Learning Repository

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

And in early AI too

Block Worlds (Winograd 1971)

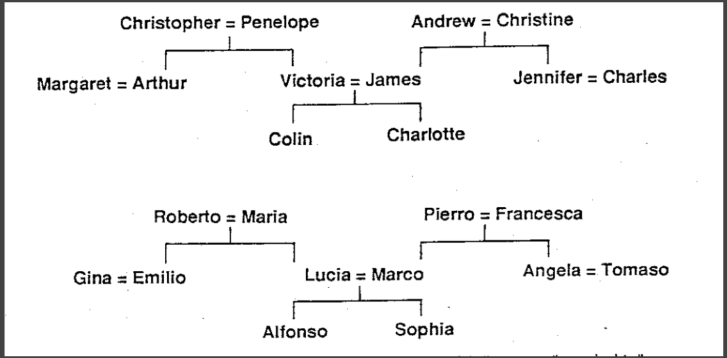


(a) "Pick up a big red block."

(b) "Find a block which is taller than the one you are holding and put it into the box."

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

Simple grammar

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

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

(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 Schemas

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
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

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

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

Dataset Name	Task Format	Supervision type	Total Size	Authors / Reference
TREC-QA	Query log, IR + free form	Human verification	1,479	Voorhees and Tice (2000)
QuizBowl	Trivia Question Answering	Expert Creation	37,225	Boyd-Graber et al (2012)
WebQuestions	NL question + KB	Google Search API & Human verification	5,810	Berant et al. (2013)
MCTest	Multiple Choice QA	crowdsourced	2640	Richardson et al. (2013)
CNN & Daily Mail	Cloze, Multiple Choice QA	Distant Supervision	387,420 + 997,467	Hermann et al. (2015)
WikiQA	Extractive QA/ sentence selection Â with Bing queries	crowdsourced	3,047	Yang et al. (2015)
bAbI	**20 complex reasoning tasks with controlled language**	**Automatically Generated**	**20,000**	**Weston et al. (2015)**
SimpleQuestions	NL question + KB	KB + crowdsourced questions	108,442	Bordes et al (2015)
Children Book Test	Multiple Choice Cloze QA	Automatic (fill-the-blank)	687,343	Hill et al. (2016)
SQuAD (1.0 + 2.0)	**Extractive QA**	**Crowdsourced**	**107,702**	**Rajpurkar et al (2016), Rajpurkar and Jia et al (2018)**
ComplexQuestions	NL question + KB	Search API & Human verification	2,100	Bao et al. (2016)
MovieQA	Multiple choice QA, text & video.	crowdsourced	14,944	Tapawasi et al. (2016)
WhoDidWhat	Cloze, Multiple Choice QA	Distant Supervision	205,978	Onishi et al. (2016)
MS MARCO	Bing queries and NL answers	crowdsourced	100,000	Nguyen et al (2016)
Lambada	Cloze QA	Automatic (human verification)	10,022	Paperno et al. (2016)
WikiReading	KB query, NL text	Distant Supervision	18.58M	Hewlett et al. (2016)
TriviaQA	Trivia Question Answering	Expert Creation + Distant Supervision	662,659	Joshi et al. (2017)
SciQ	Multiple choice QA	crowdsourced	13,679	Welbl et al. (2017)
RACE	Multiple choice Exam questions	Expert Creation	97,687	Lai et al. (2017)
NewsQA	Extractive QA	crowdsourced	119,633	Trischler et al. (2017)
AI2 Science Questions	Multiple Choice Science Exam QA	Expert Creation	5,059	Allen Institute for AI (2017 release)
SearchQA	Trivia questions + Search Engine Results	Expert Creation + distant supervision	140,461	Dunn et al. (2017)
QUASAR-S & QUASAR-T	Cloze & free-form trivia questions	Distant supervision	37,362 + 43,013	Dhingra et al. (2017)
Wikihop & Medhop	KB query, NL text, multiple Choice	Distant Supervision	51,318+2,508	Welbl et al. (2018)
NarrativeQA	free-form answer generation	crowdsourced	46,765	Kocisky□ et al. (2018)

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 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”.

Question + Answer

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

within a cloud

Task: Given a paragraph and a question about it, predict the text span that states the correct answer.

Stanford Question Answering Dataset (SQuAD)

Text Pass

[...] Precipit
as smaller
coalesce vi
with other r
ice crystals
cloud. Sho
periods of r
scattered lo
called "sho

on + Answer

o water droplets
th ice crystals to
cipitation?

cloud

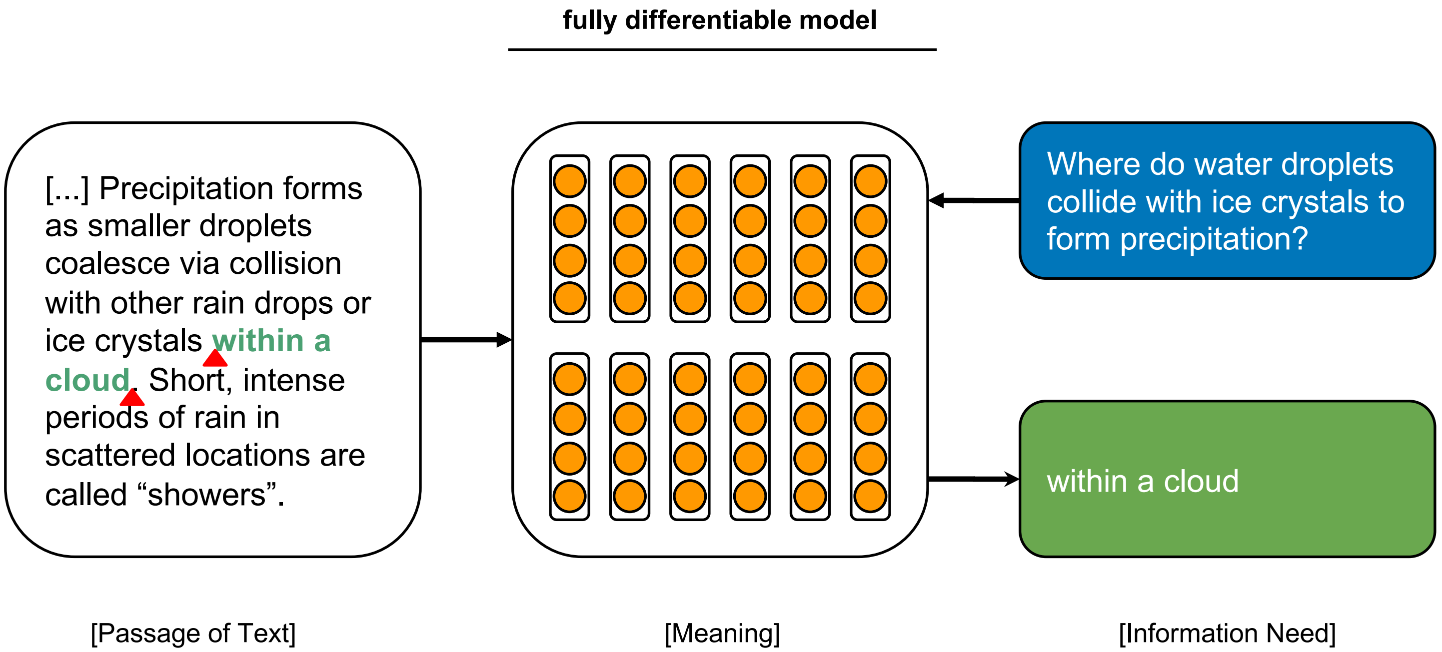
Task: Giv
predict th

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, 20	BERT + Synthetic Self-Training (ensemble)	84.292	86.967
5 Dec 16, 20			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

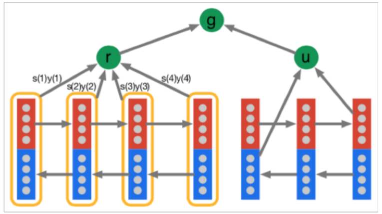
End-to-end Machine Reading for Question Answering



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]

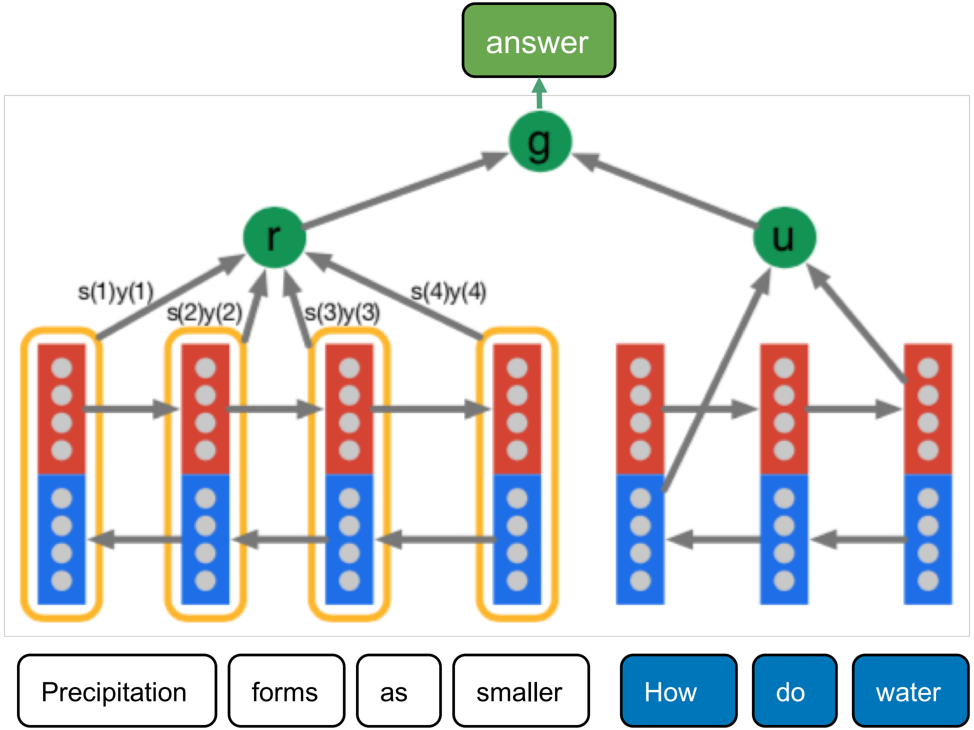
[Meaning]

[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

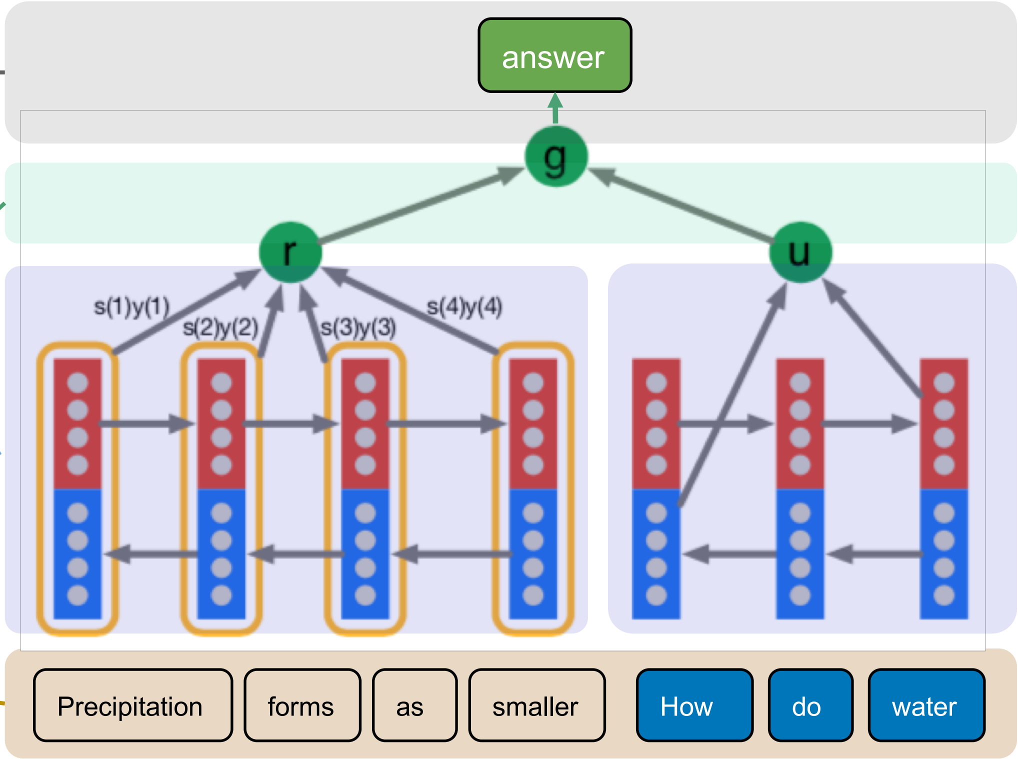
The Attentive Reader Model: Overview

Answer Selection: answer prediction

Sequence Interaction: Matching text with question

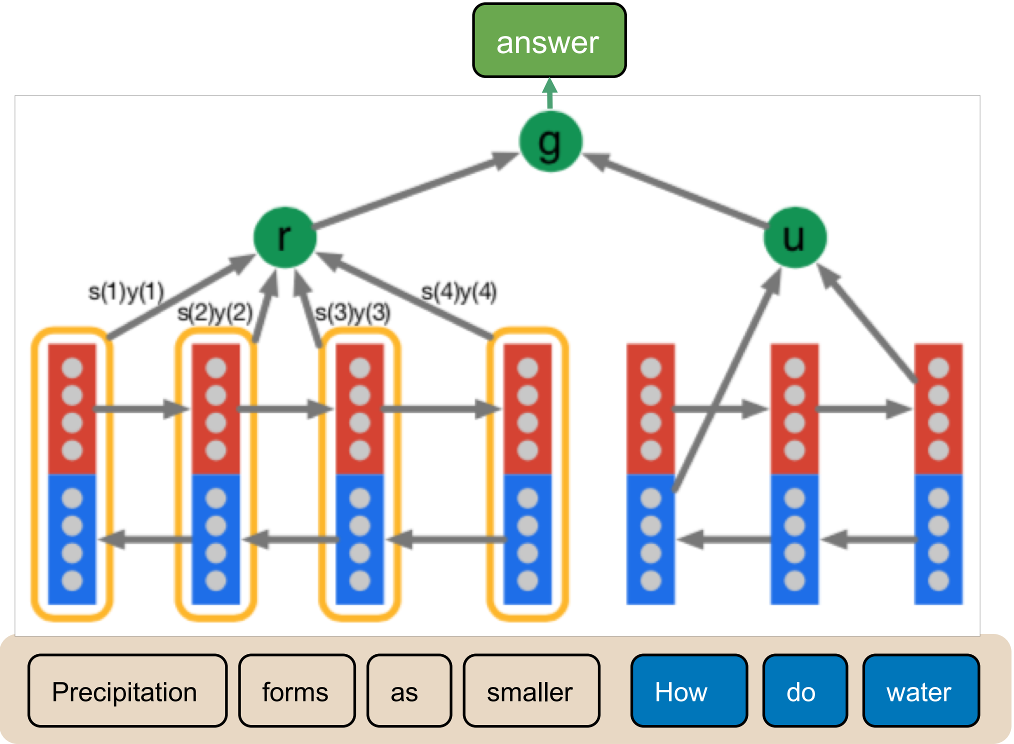
Composition: incorporating context around words

Input: Representing symbols as vectors



Modified visualization from Hermann et. al. NIPS'15

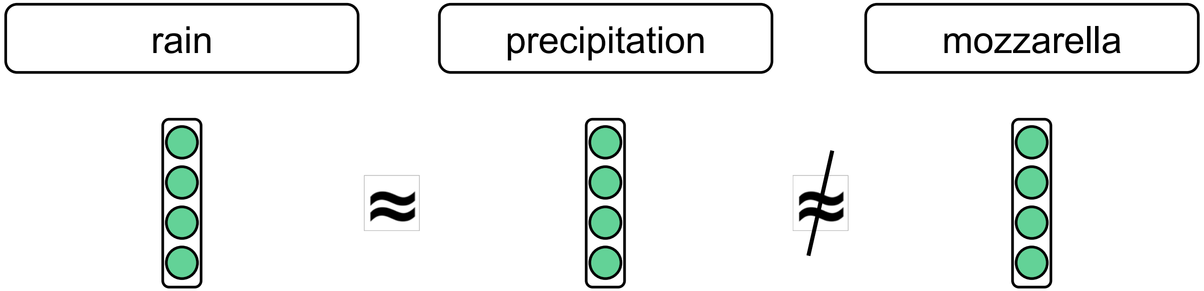
The Attentive Reader Model: Overview



Input: Representing symbols as vectors

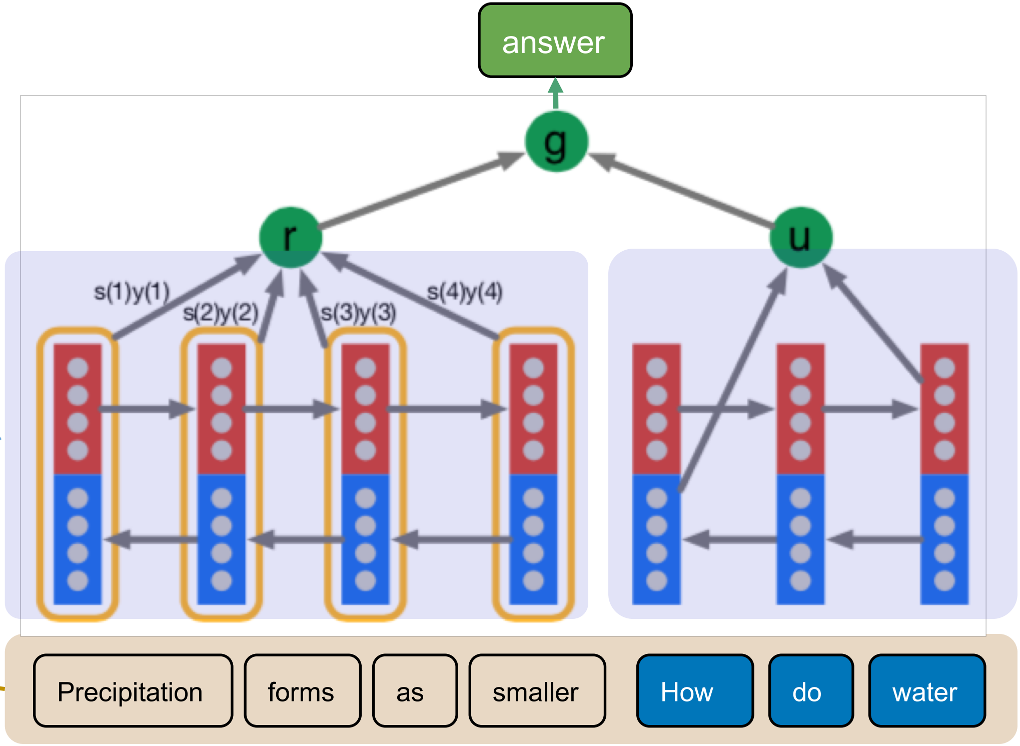
Modified visualization from Hermann et. al. NIPS'15

Representations for words: Embeddings



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

The Attentive Reader Model: Overview

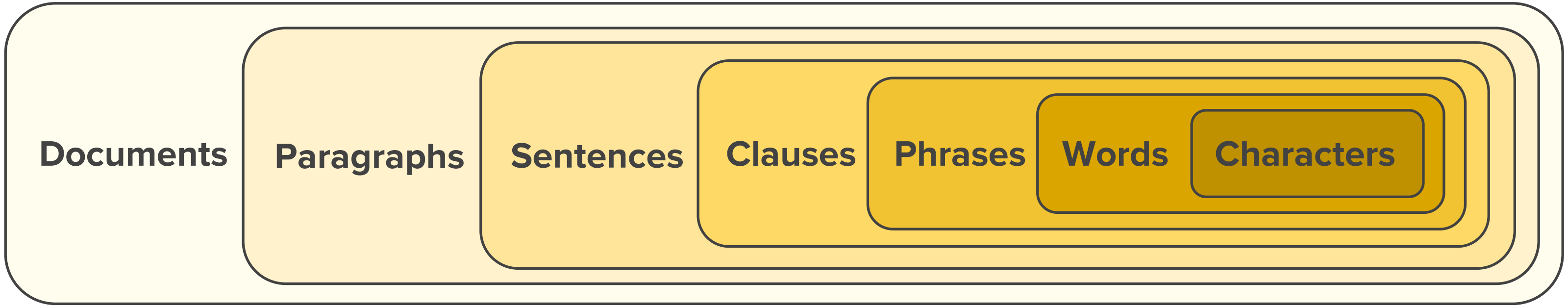


Composition: incorporating context around words

Input: Representing symbols as vectors

Modified visualization from Hermann et. al. NIPS'15

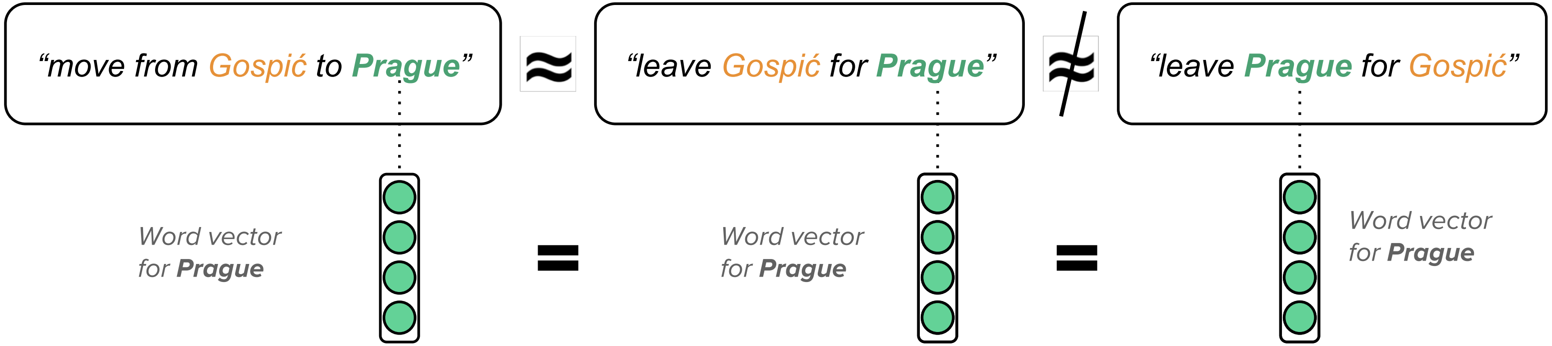
Language is compositional



Challenges

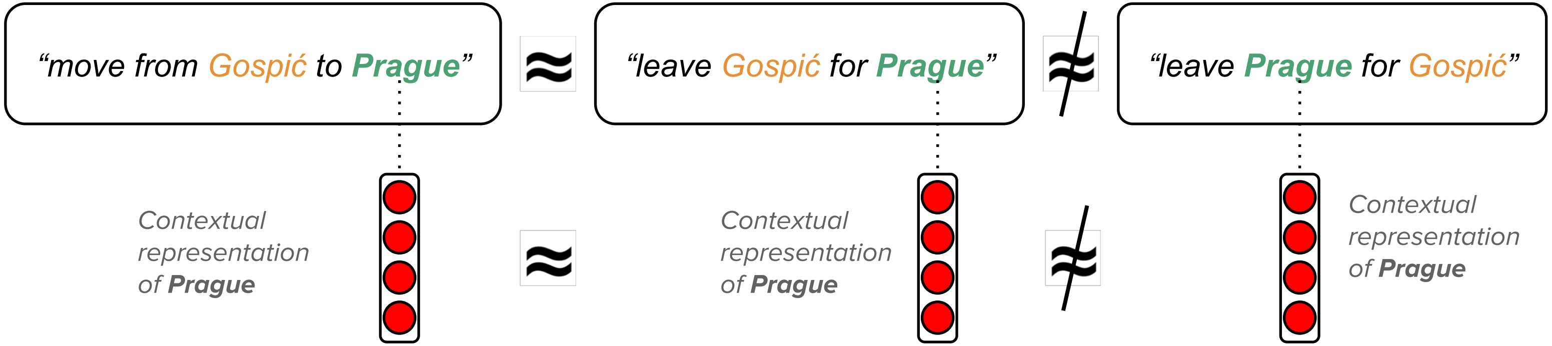
- Inductive bias: which composition function to use?
  - sequence, tree or more general graph structures?
  - varies for different levels
- Capturing long-range dependencies
  - co-reference (tracking entities)
  - effective information flow: ease of learning

Representing Words in Context



- Word representations should vary depending on context

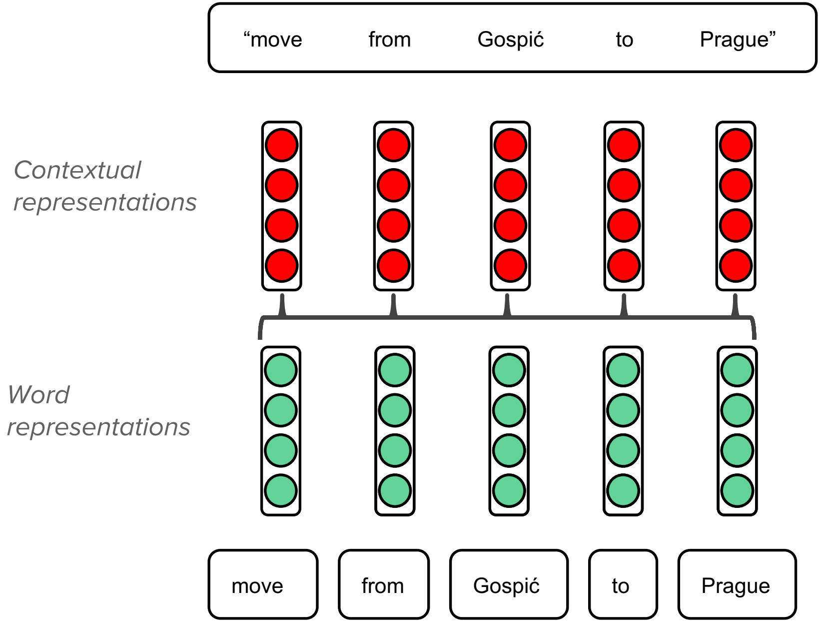
Representing Words in Context



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

Representing Words in Context

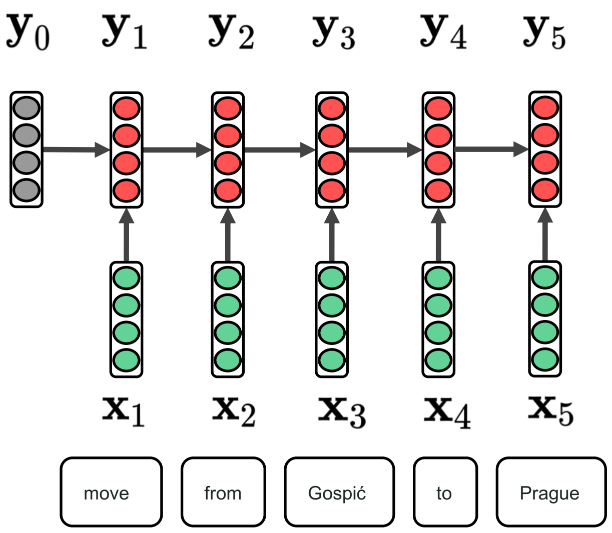
- composition of word vectors into contextualized word representations
- use vector composition function



Recurrent Neural Network Layers

- **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



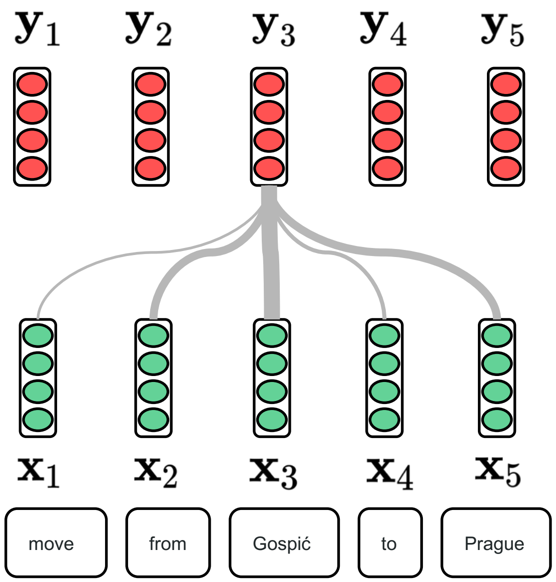
$$\mathbf{y}_t = f(\mathbf{x}_t, \mathbf{y}_{t-1})$$

Tree-variants:

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

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)*
  - *Good for:* phrases, sentences, paragraphs



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

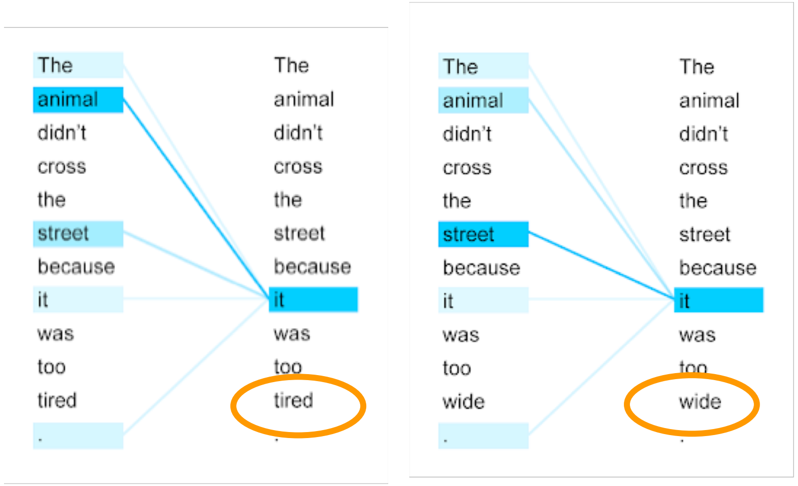
$$\tilde{\mathbf{x}}_t^k = \sum_{j=1}^{T} \alpha_{j,t}^k \mathbf{x}_j \qquad k = 1, \ldots, K$$

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

α_t^k : k^{th} self-attention weights for token

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

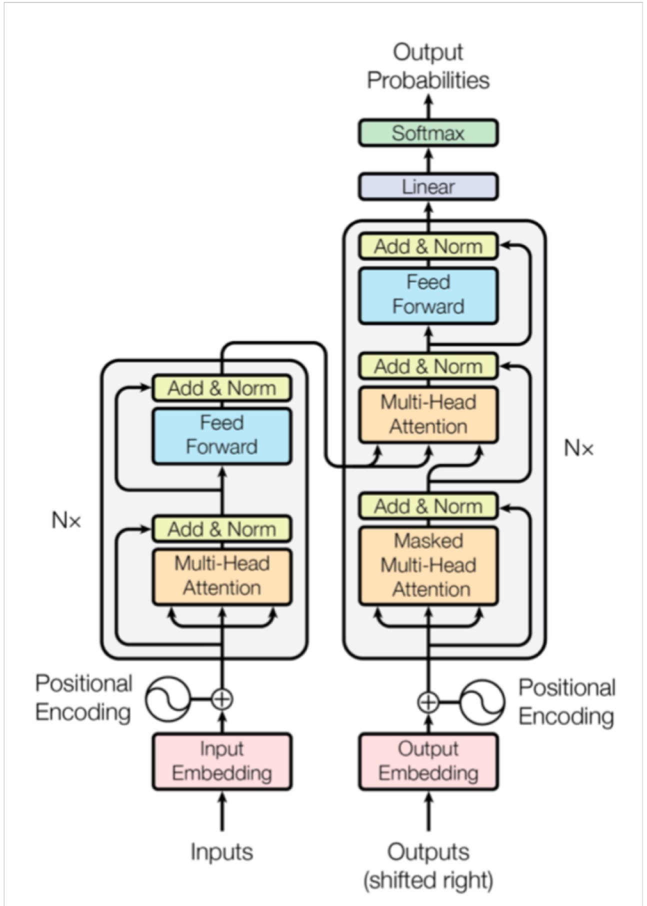
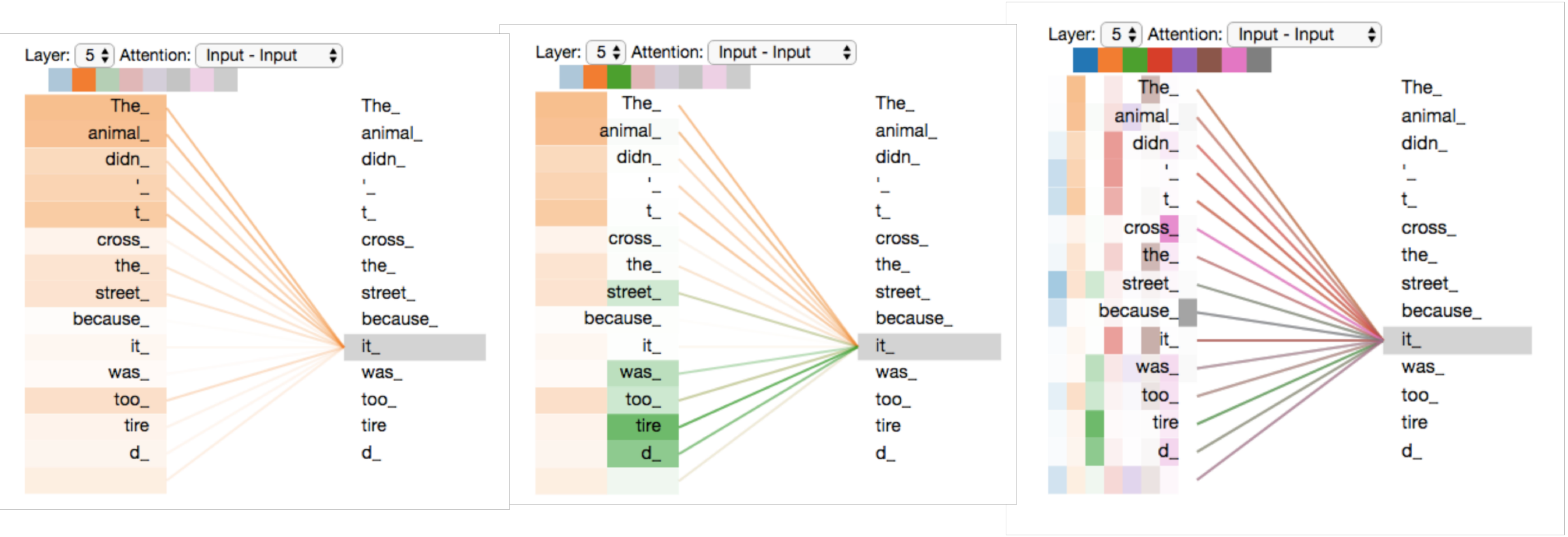


Figure from Vaswani et al., NIPS'17

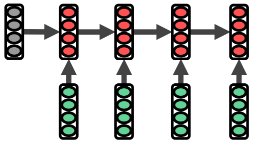
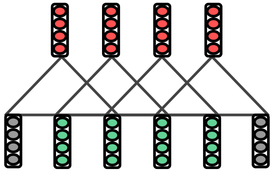
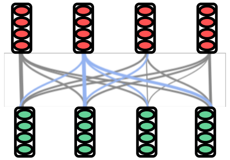
Multi-head attention



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

Compositional Sequence Encoders - Overview

- Language is compositional!
  - Characters → Words → Phrases → Clauses → Sentences → Paragraphs → Documents

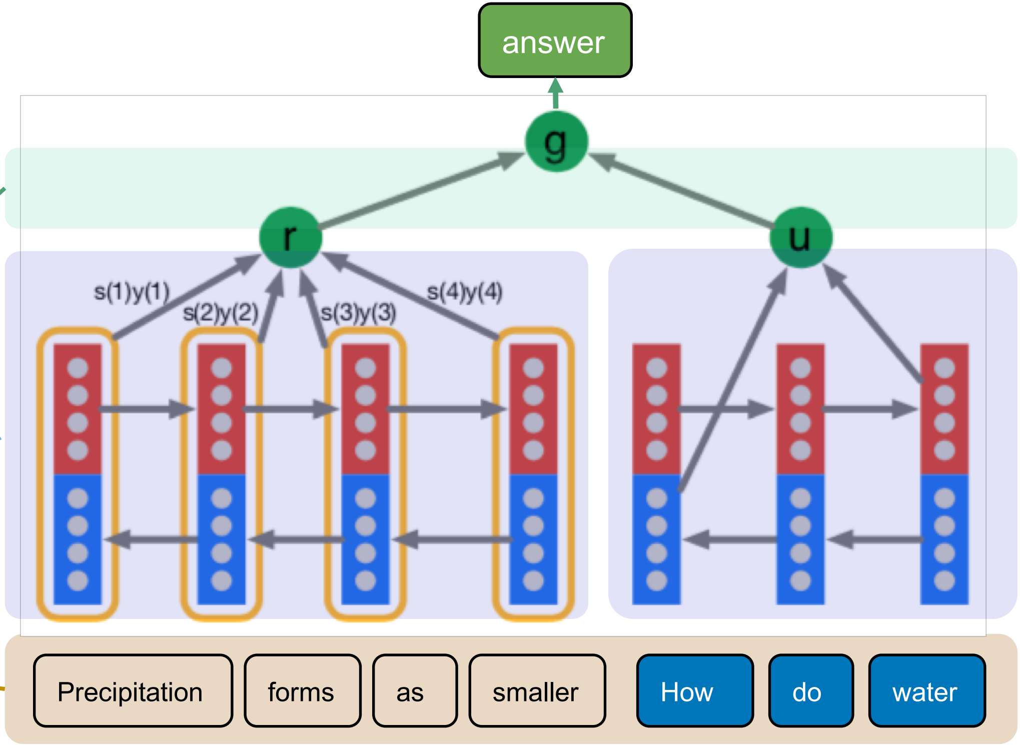
Architecture	**RNN** (LSTM, GRU)	**CNN**	**Self-Attention**
Illustration			
Function $\mathbf{y}_t =$	$f(\mathbf{x}_t, \mathbf{y}_{t-1})$	$f(\mathbf{x}_{t-k}, \ldots, \mathbf{x}_{t+k})$	$f(\mathbf{x}_1, \ldots, \mathbf{x}_T)$
Advantages	- unlimited context - recency bias	- parallelizable → fast - local n-gram patterns	- parallelizable → fast - long-range dep
Disadvantages	- slower - strong recency bias - long-range dep	- limited context - strong locality bias - long-range dep	- harder to train - careful setup of hyper-parameters

The Attentive Reader Model: Overview

Sequence Interaction: Matching text with question

Composition: incorporating context around words

Input: Representing symbols as vectors



Modified visualization from Hermann et. al. NIPS'15

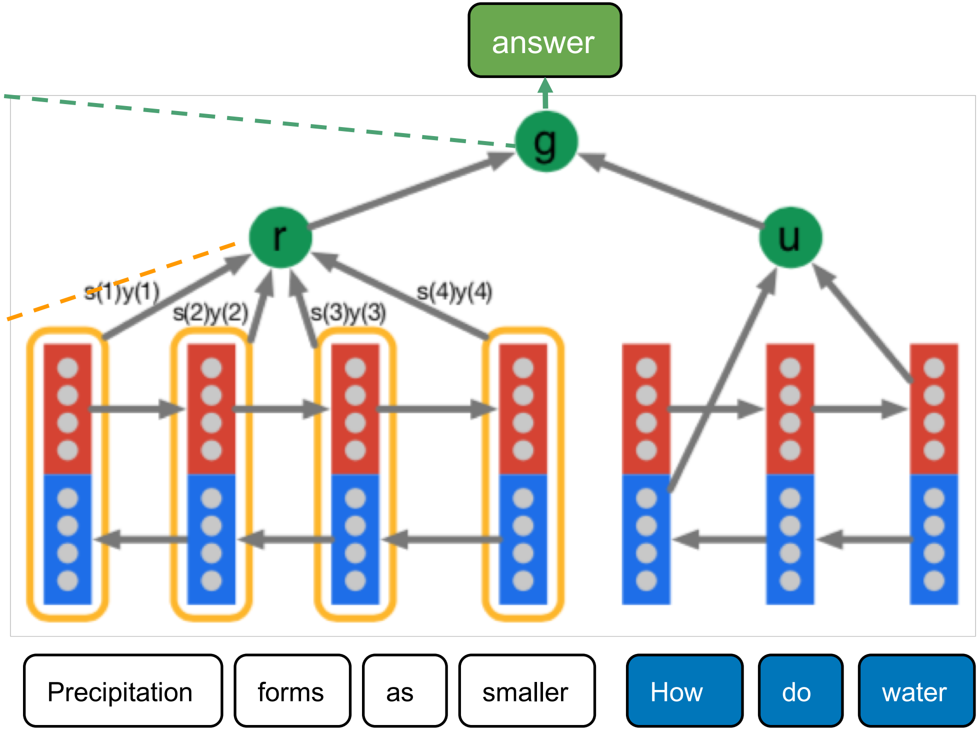
Modelling sequence interactions

- **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

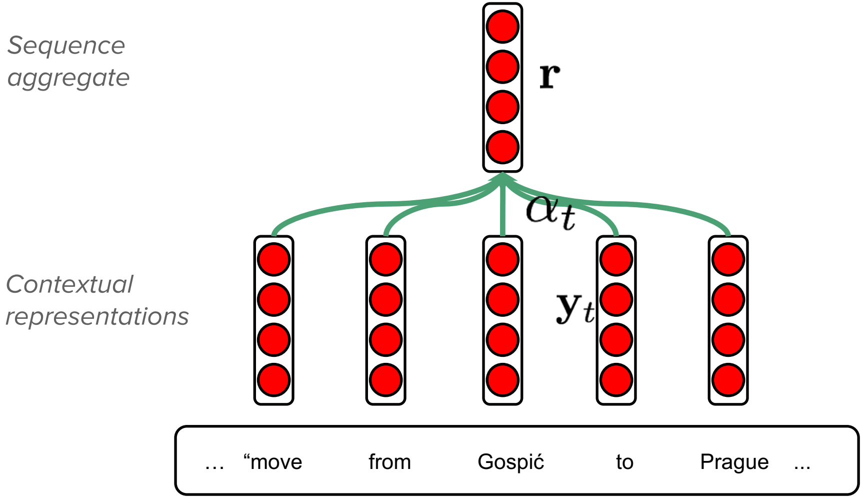
attention-weighted sum of contextualised word representations



Modified visualization from Hermann et. al. NIPS'15

Modelling sequence interactions: Attention

- **Attention:** relevance-weighted pooling of vectors across sequence
  - Attention mask computed can be conditional on question and text
  - Determines relevance of tokens for answering the question

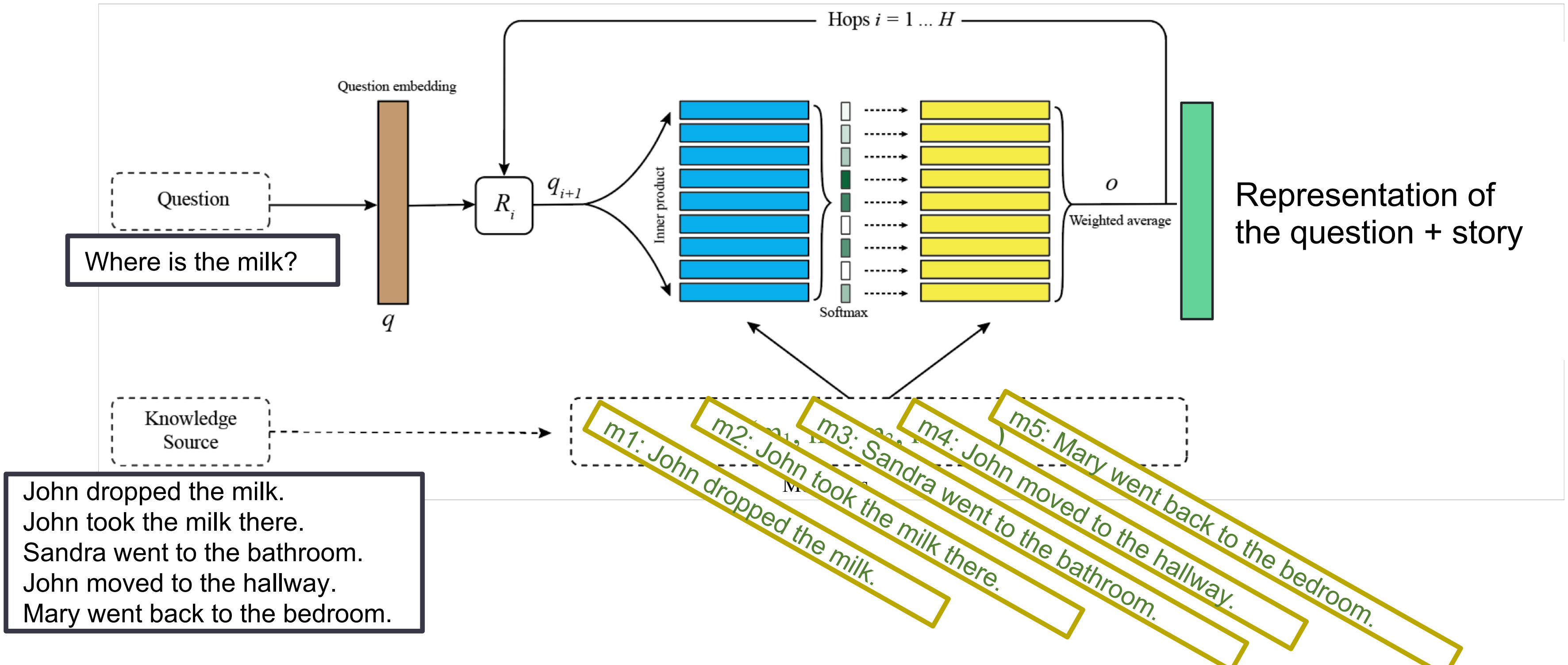


$$\mathbf{r} = \sum_{t=1}^{T} \alpha_t \mathbf{y}_t$$

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

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

S: 1 So they had to fall a long way .
2 So they got their tails fast in their mouths .
3 So they could n't get them out again .
4 That 's all . '
5 ` Thank you , ' said Alice , ` it 's very interesting .
6 I never knew so much about a whiting before . ' '
7 I can tell you more than that , if you like , ' said the Gryphon .
8 ` Do you know why it 's called a whiting ? ' '
9 I never thought about it , ' said Alice .
10 ` Why ? '
11 ` IT DOES THE BOOTS AND SHOES . '
12 the Gryphon replied very solemnly .
13 Alice was thoroughly puzzled .
14 ` Does the boots and shoes ! '
15 she repeated in a wondering tone .
16 ` Why , what are YOUR shoes done with ? '
17 said the Gryphon . '
18 I mean , what makes them so shiny ? '
19 Alice looked down at them , and considered a little before she gave her answer .
20 ` They 're done with blacking , I believe .

q: `Boots and shoes under the sea , ' the _______ went on in a deep voice , are done with a whiting .

C: Alice, BOOTS, Gryphon, SHOES, answer, fall, mouths, tone, way, whiting.

MemNNs (window + self-sup.): **Gryphon**

S: 1 He thought that Old Mr. Toad was trying to fool him .
2 Presently Peter Rabbit came along .
3 He found Jimmy Skunk sitting in a brown study .
4 He had quite forgotten to look for fat beetles , and when he forgets to do that you may make up your mind that Jimmy is doing some hard thinking .
5 `` Hello , old Striped-coat , what have you got on your mind this fine morning ? ''
6 cried Peter Rabbit .
7 `` Him , '' said Jimmy simply , pointing down the Lone Little Path .
8 Peter looked .
9 `` Do you mean Old Mr. Toad ! ''
10 he asked .
11 Jimmy nodded .
12 `` Do you see anything queer about him ? ''
13 he asked in his turn .
14 `` Do you see anything queer about him ? ''
15 he asked .
16 Peter stared down the Lone Little Path .
17 `` No , '' he replied , `` except that he seems in a great hurry . ''
18 `` That 's just it , '' Jimmy returned promptly .
19 `` Did you ever see him hurry unless he was frightened ? ''
20 Peter confessed that he never had

q: `` Well , he is n't _______ now , yet just look at him go '' retorted Jimmy .

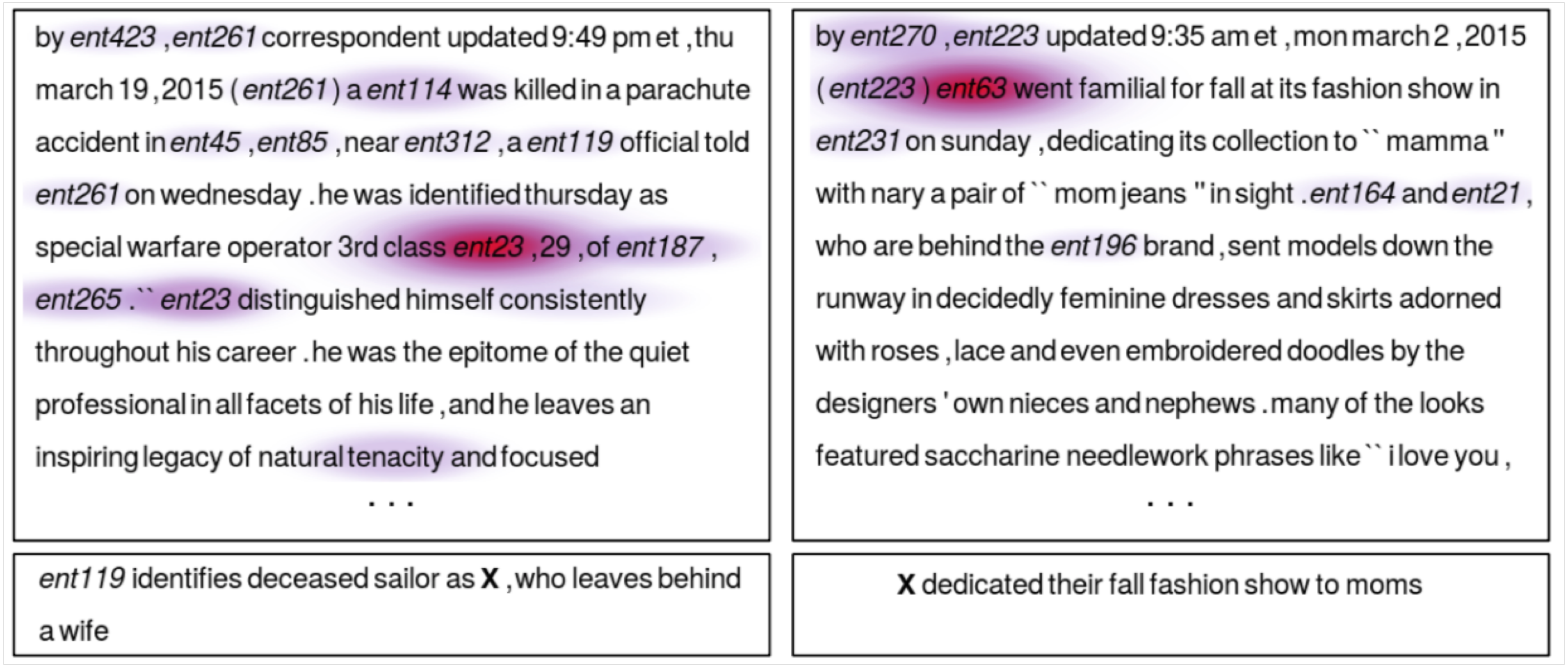
C: Do, came, confessed, frightened, mean, replied, returned, said, see, thought.

MemNNs (window +self-sup.): **frightened**

Memory Networks on Children Book Test dataset

Visualization from Hill et. al., ICLR'16

Example: Learned attention patterns



Attentive Reader on QACNN/DailyMail dataset

Visualization from Hermann et. al. NIPS'15

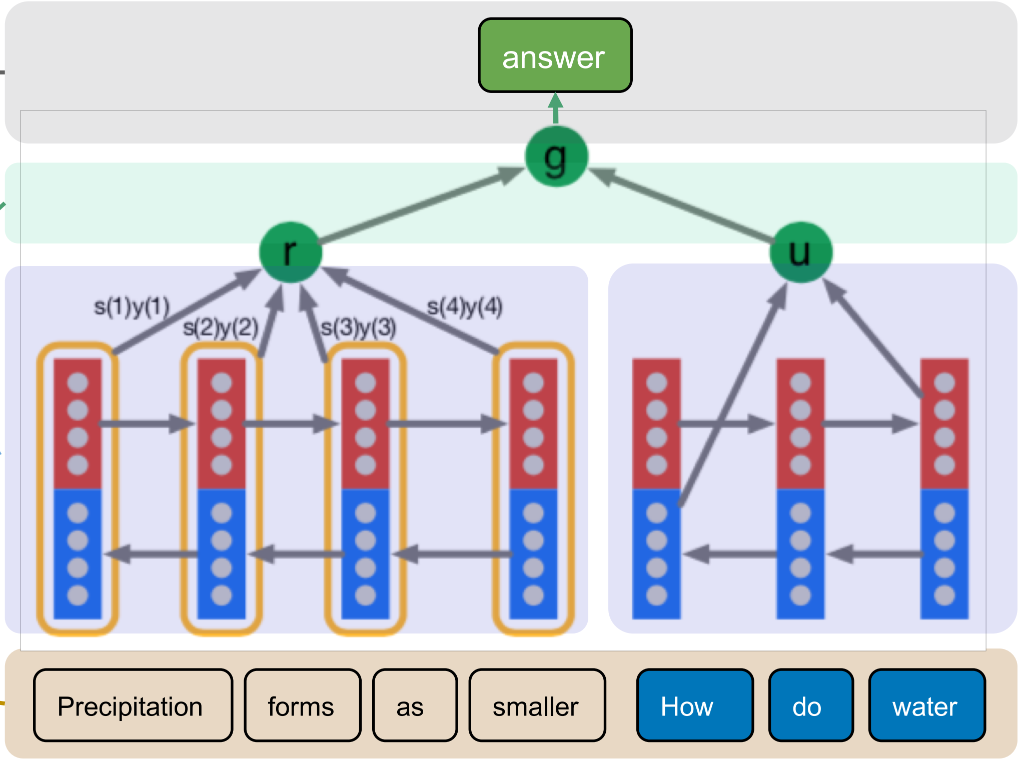
The Attentive Reader Model: Overview

Answer Selection:
answer prediction

Sequence Interaction:
Matching text with question

Composition: incorporating context around words

Input: Representing symbols as vectors



Modified visualization from Hermann et. al. NIPS'15

Answer prediction

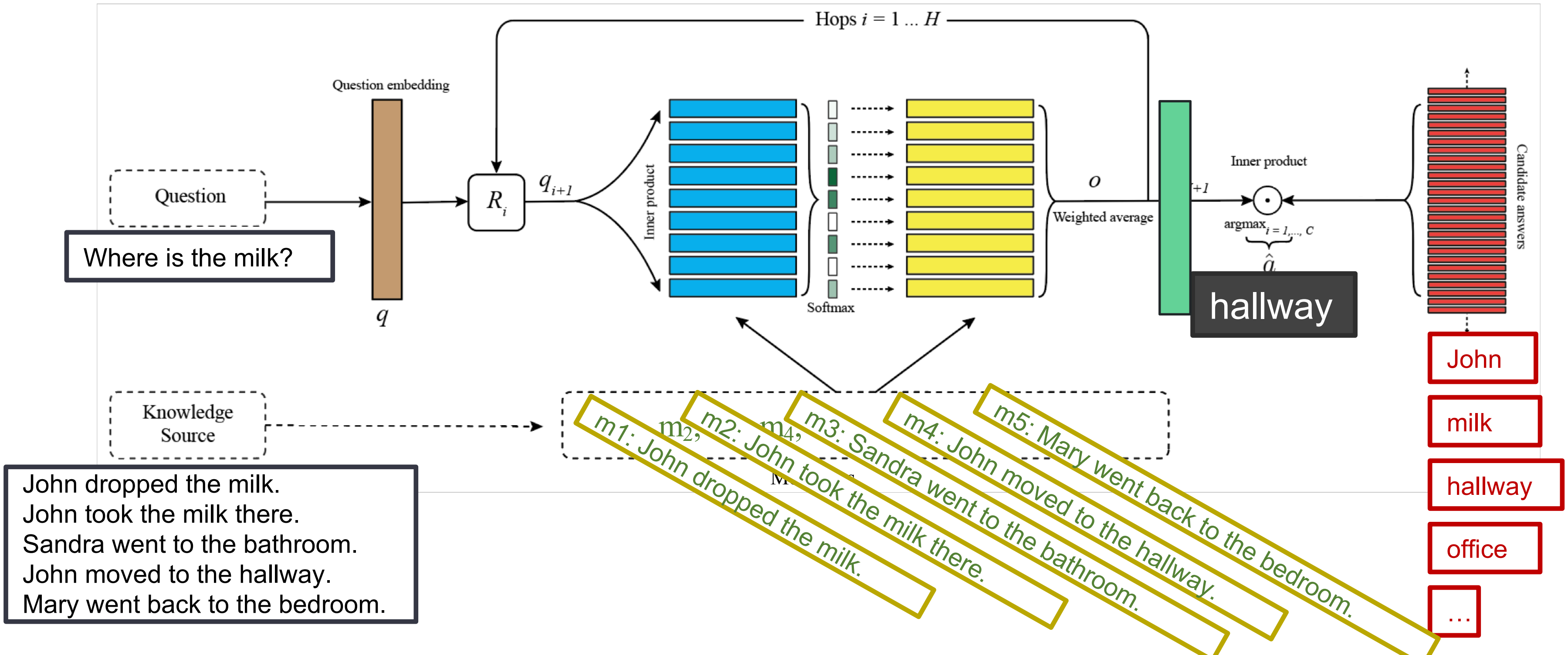
- Usually linear projection

- **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



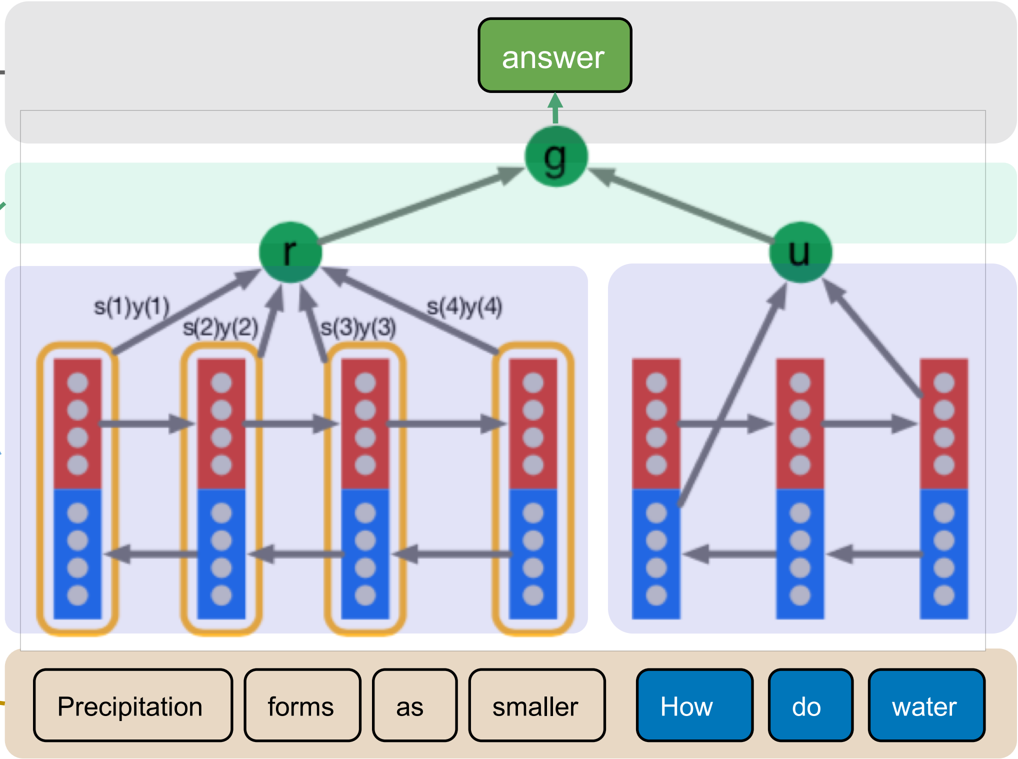
The Attentive Reader Model: Overview

Answer Selection: answer prediction

Sequence Interaction: Matching text with question

Composition: incorporating context around words

Input: Representing symbols as vectors



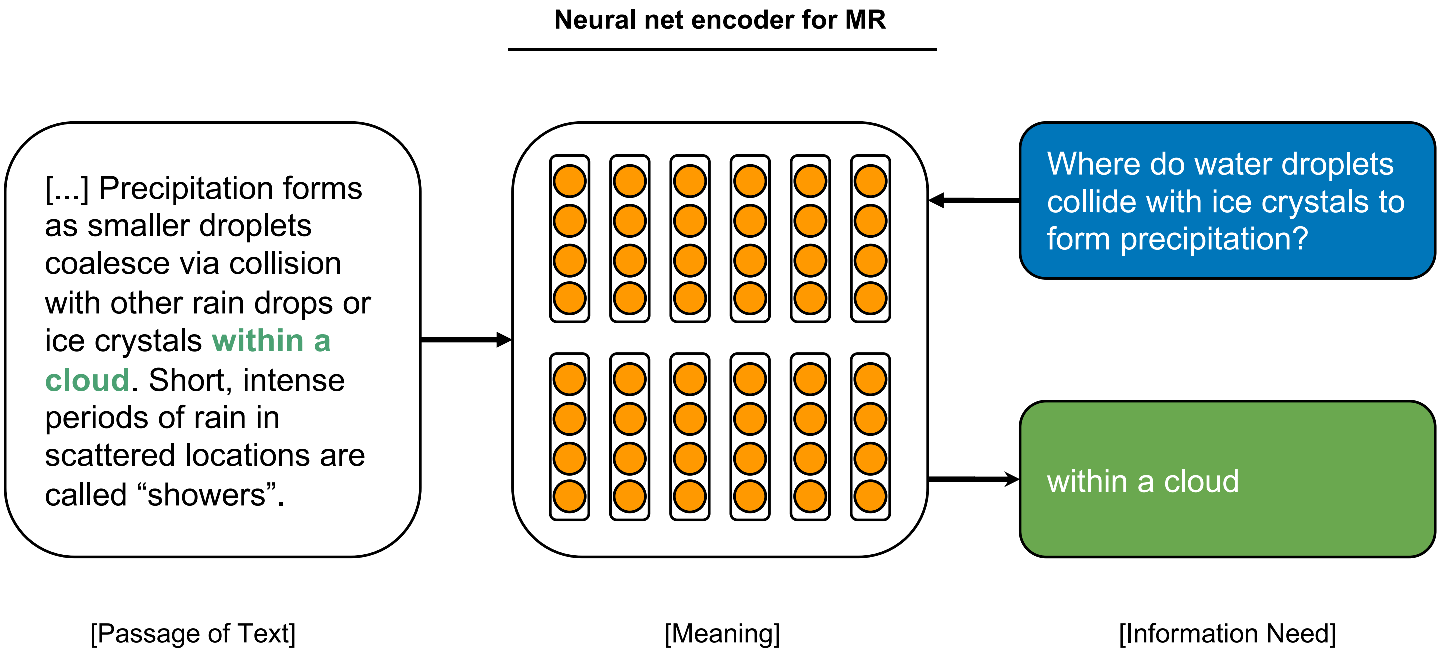
Modified visualization from Hermann et. al. NIPS'15

Conclusion

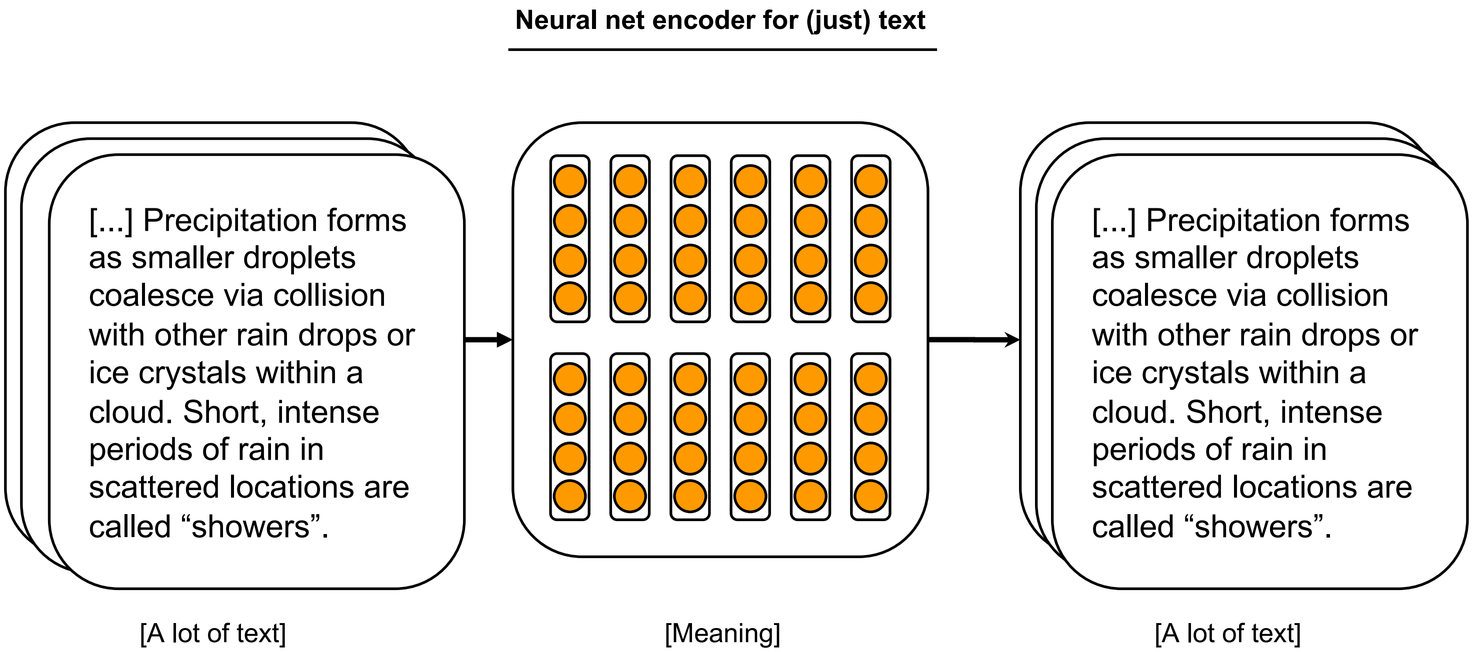
- We gathered all ingredients to build state-of-the-art supervised Machine Reading systems!
- architectures work well in practice
  … as long as we stay in-domain and questions are simple
- We covered only extractive and multiple choice questions settings but there is also generative machine reading
- Practice in Labs tomorrow on bAbI!

Machine Reading / Current Trend

Supervised training



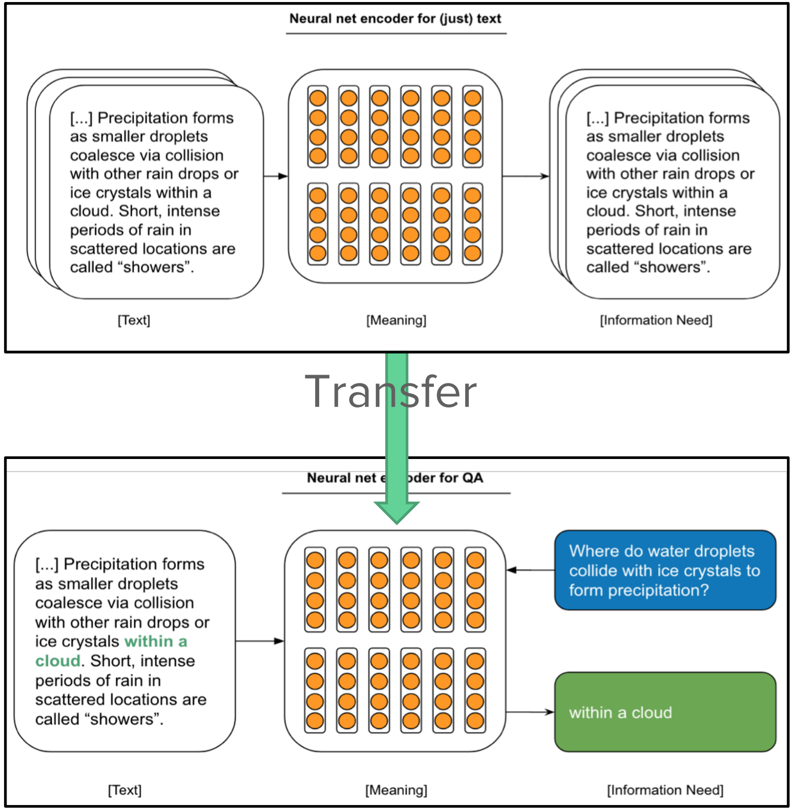
Unsupervised pretrained representations



Lifting over pretrained representations

Pretrained Language Model

Machine Reading



How is this different from pretrained word embeddings?

Pretrained Word Embeddings (word2vec)

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

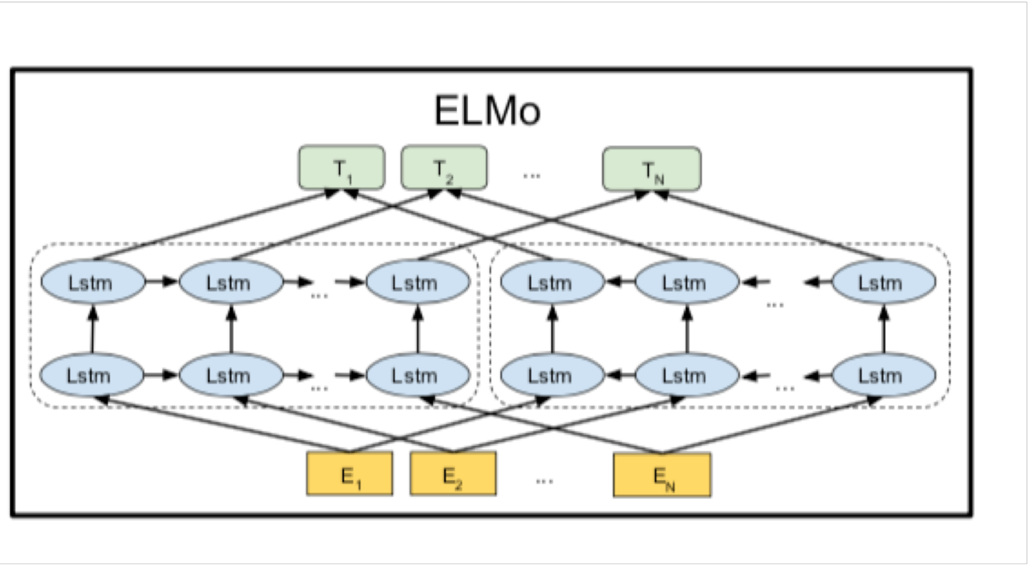
Pretrained Contextualized Embeddings (e.g. ELMo, BERT)

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

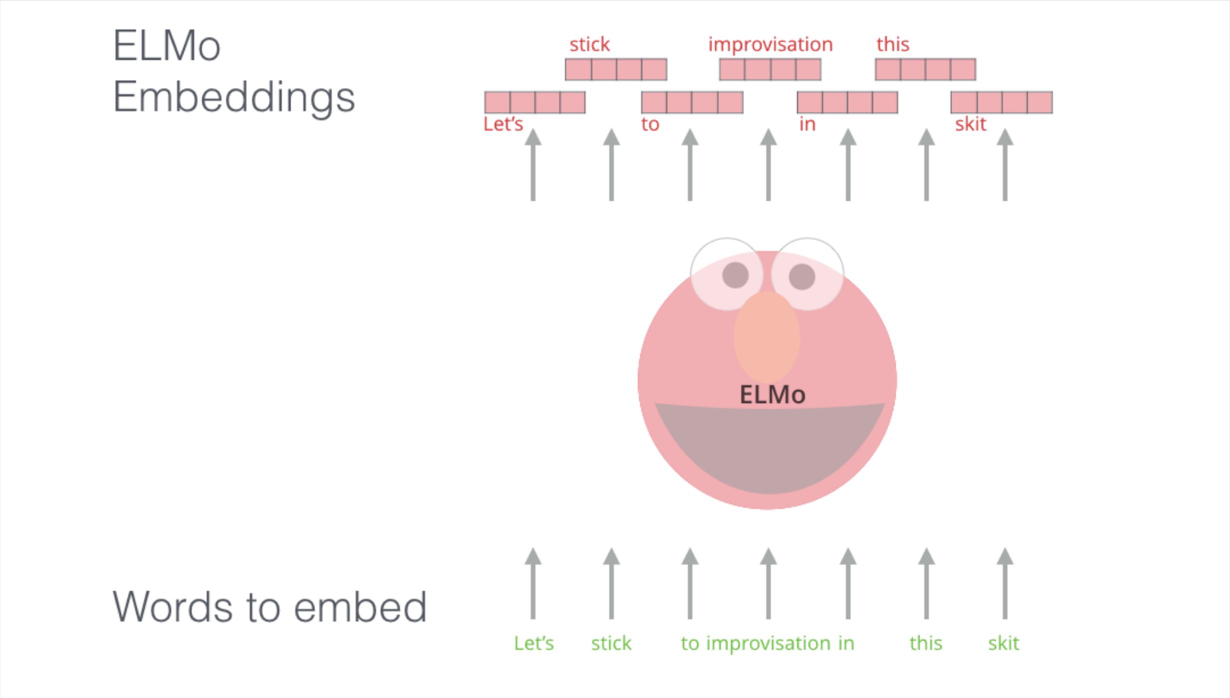
ELMo: **E**mbeddings from **L**anguage **Mo**dels

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 forward and backward representations into final contextual embeddings

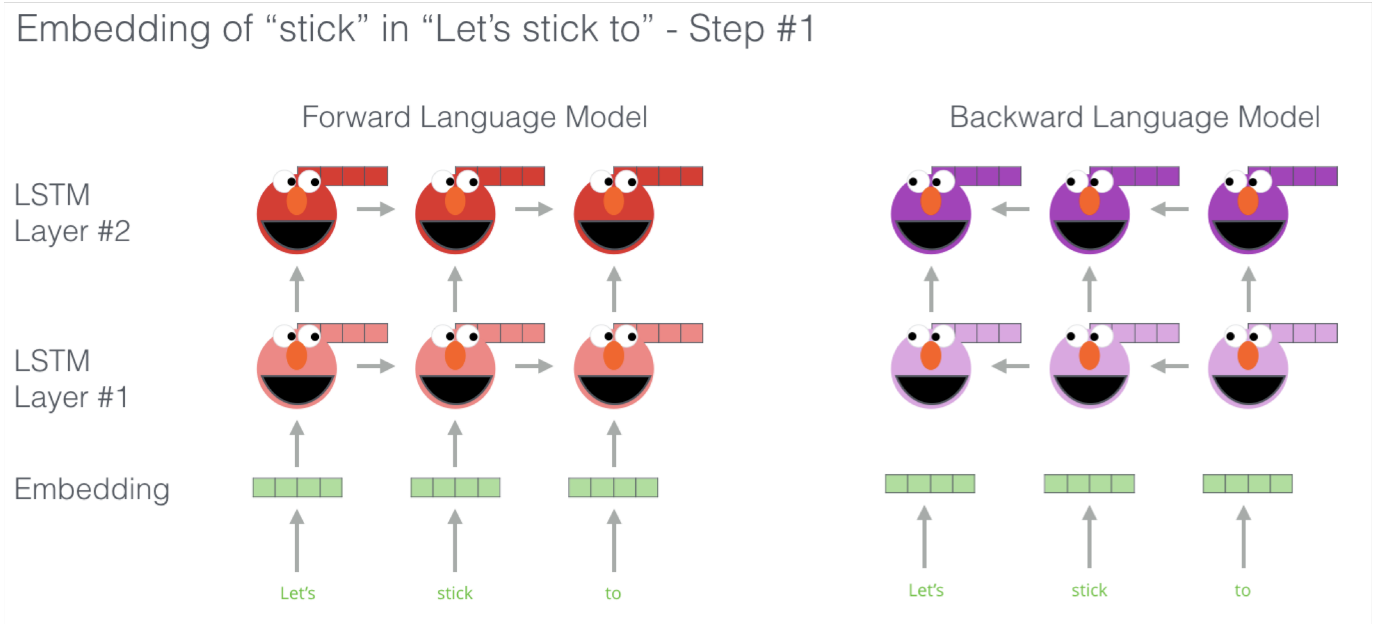


ELMo: **E**mbeddings from **L**anguage **Mo**dels



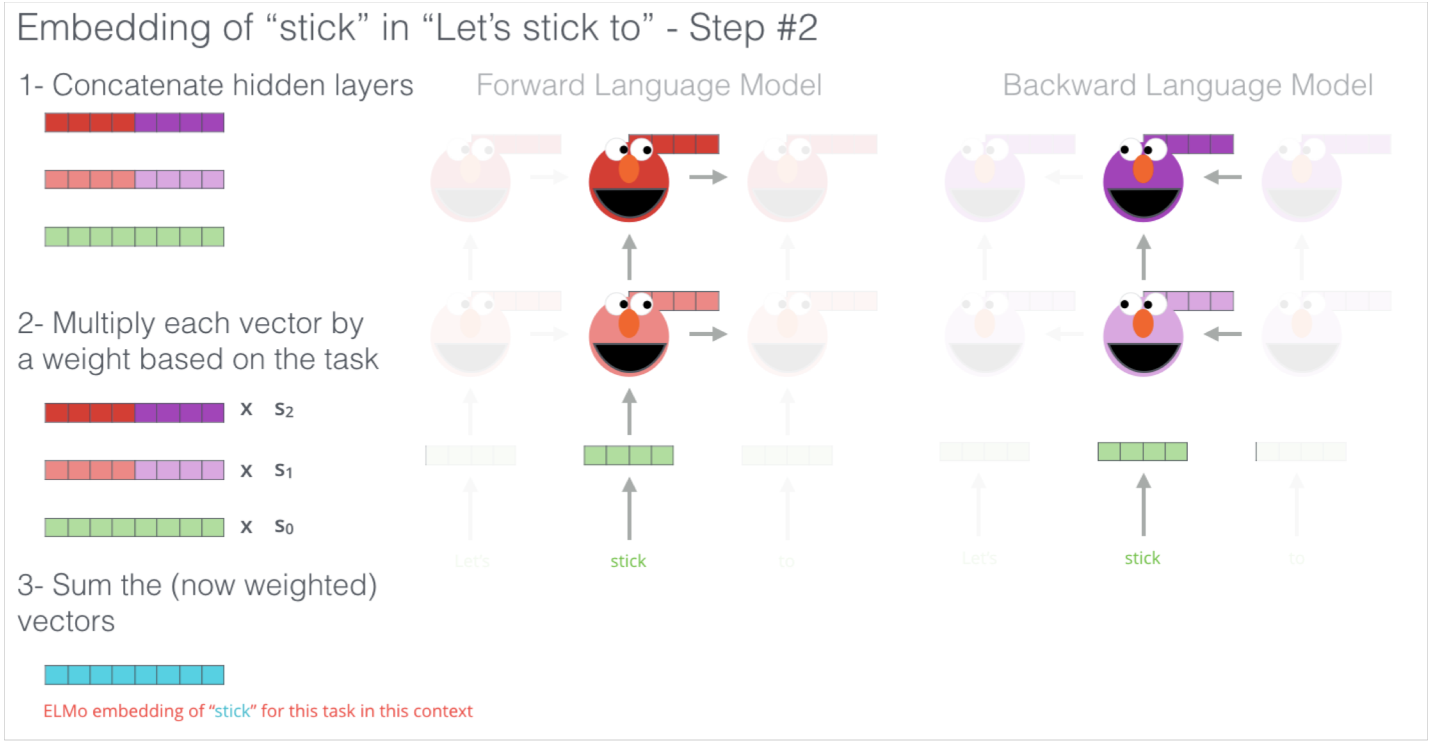
Figures from http://jalammar.github.io/illustrated-bert/

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



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

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



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 .

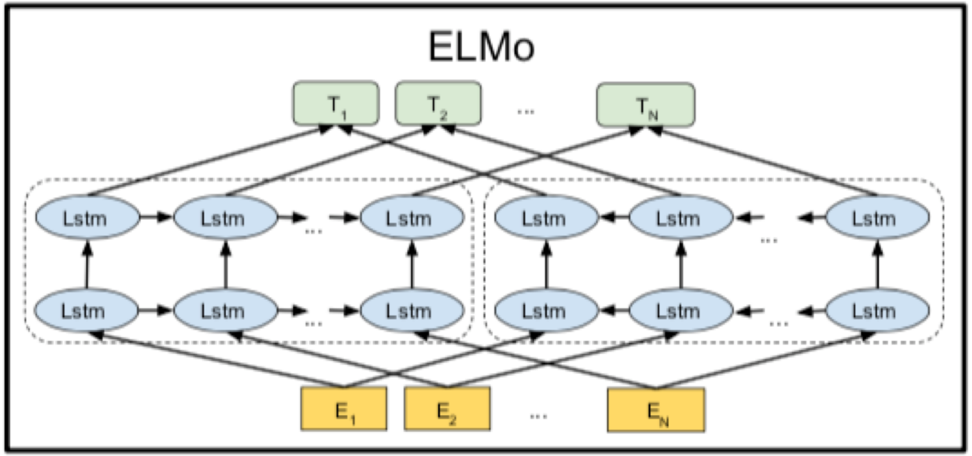
Problems with ELMo

- 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!

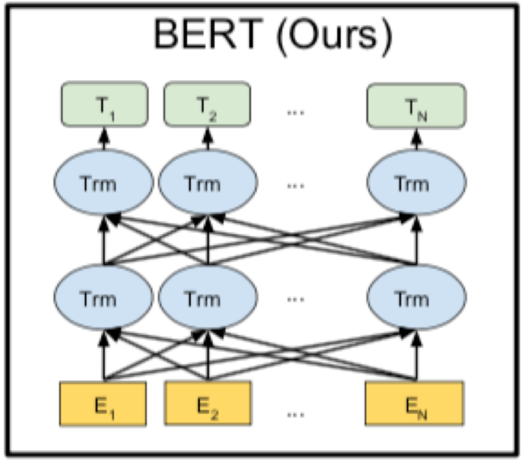
BERT - **B**idirectional **E**ncoder **R**epresentations from **T**ransformers

Devlin et al., NAACL'19

Solutions: use Transformer + encoder layers instead of decoder layers



(OpenAI GPT)

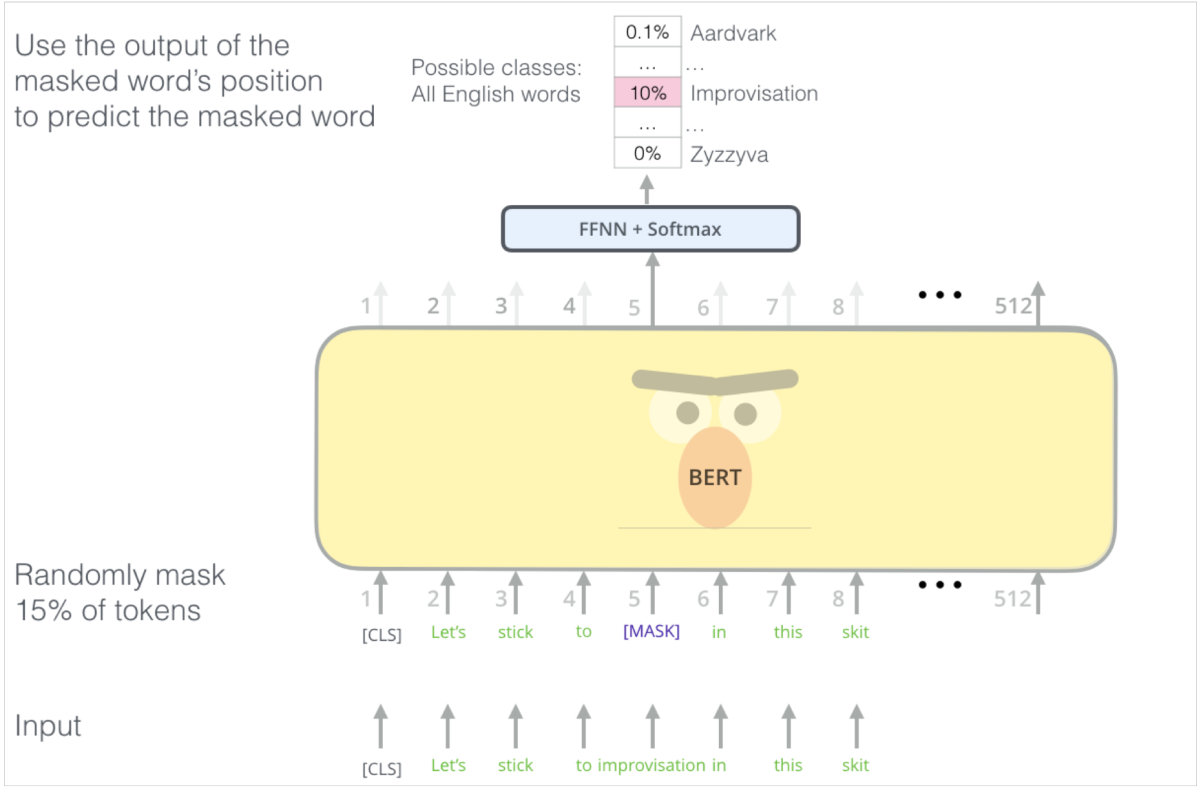


Innovation with multiple pretraining tasks

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

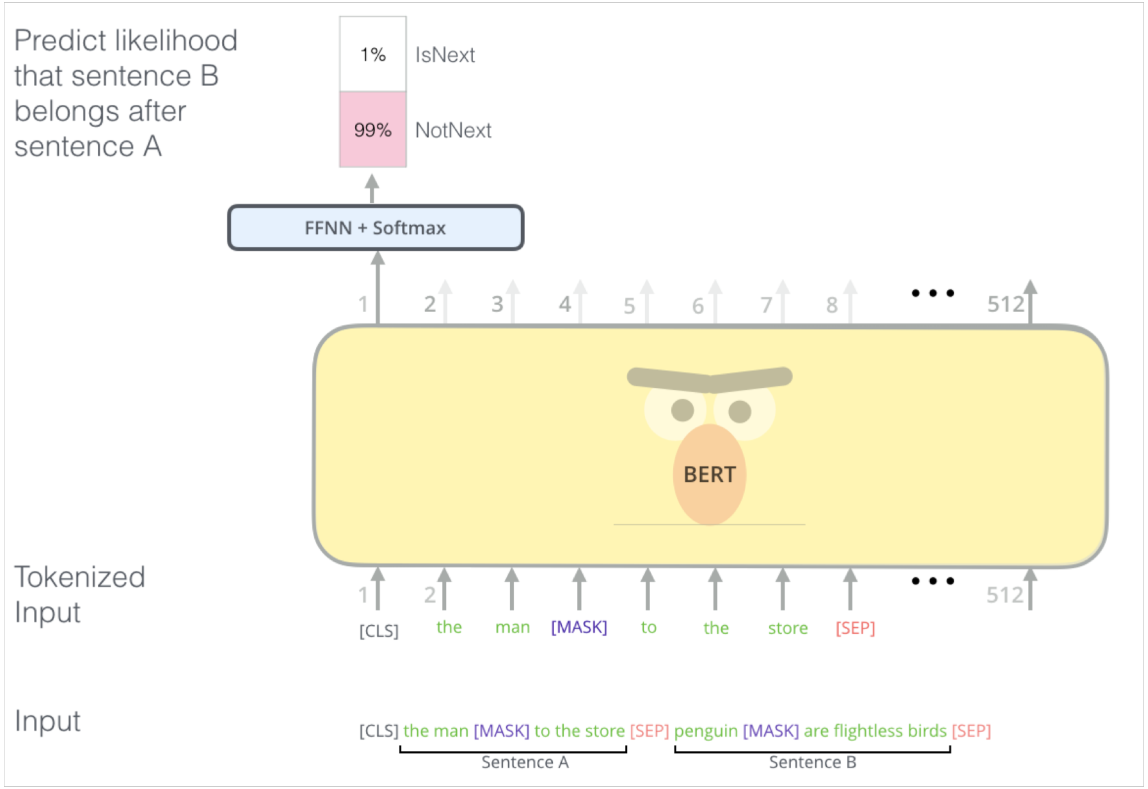


Figures from http://jalammar.github.io/illustrated-bert/

BERT – Pretraining 2: next sentence prediction

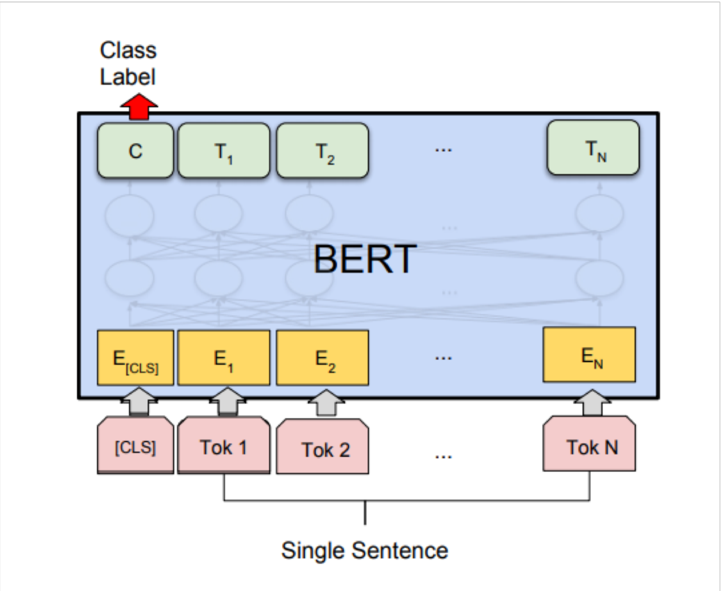
- Given two sentences, does the first follow the second?
- Teaches BERT about relationship between two sentences
- 50% of the time the actual next sentence, 50% random

BERT – Pretraining 2: next sentence prediction



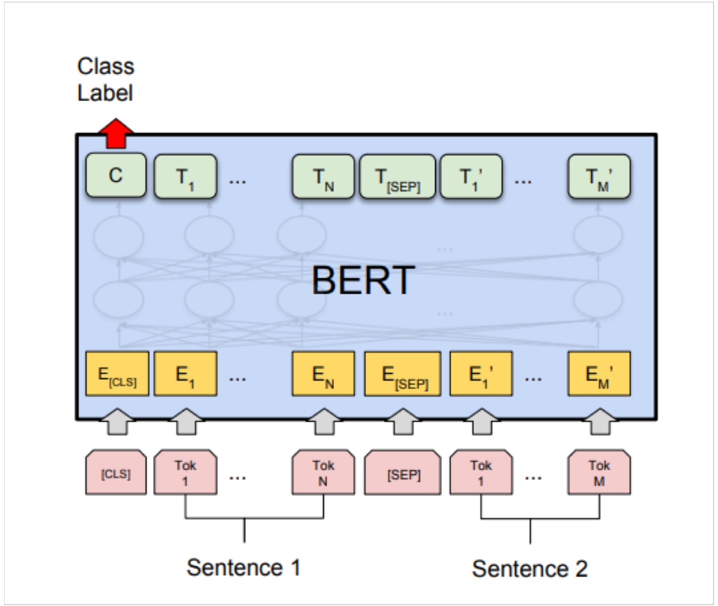
Figures from http://jalammar.github.io/illustrated-bert/

BERT – Fine-tuning for Classification



Single sentence classification

Sentiment analysis, spam detection, etc.

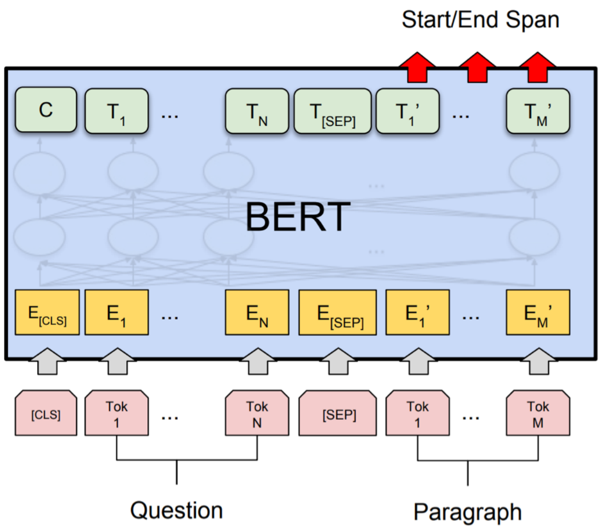


Pair of sentences classification

Entailment, paraphrase detection, etc.

Figures from Devlin et al. 18'

BERT – Fine-tuning for Machine Reading



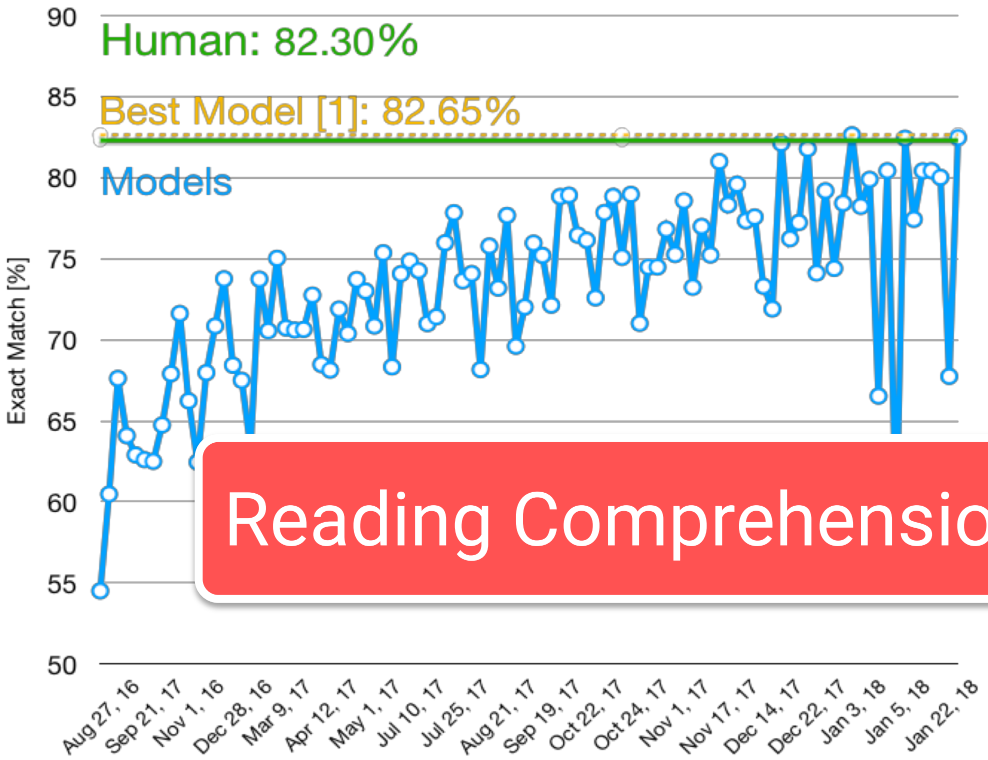
(c) Question Answering Tasks:
SQuAD v1.1

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'

Machine Reading / Open Problems

Progression of SQuAD Model Performance



Reading Comprehension Solved?

Challenge 1: Robustness

What is the name of the quarterback who was 38 in Super Bowl XXXIII?

The past record was held by quarterback John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38.

Challenge 1: Robustness

What is the name of the quarterback who was 38 in Super Bowl XXXIII?

John Elway

The past record was held by quarterback **John Elway**, who led the Broncos to victory in Super Bowl XXXIII at age 38.

Challenge 1: Robustness

What is the name of the quarterback who was 38 in Super Bowl XXXIII?

The past record was held by quarterback John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38. Quarterback Jeff Dean had a jersey number 37 in Champ Bowl XXXIV.

Challenge 1: Robustness

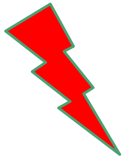
What is the name of the quarterback who was 38 in Super Bowl XXXIII?

Jeff Dean

The past record was held by quarterback John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38. Quarterback **Jeff Dean** had a jersey number 37 in Champ Bowl XXXIV.

Challenge 1: Robustness

What is the name of the quarterback who was 38 in Super Bowl XXXIII?



Jeff Dean

The past record was held by quarterback John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38. Quarterback **Jeff Dean** had a jersey number 37 in Champ Bowl XXXIV.

- Reading Comprehension models can easily be fooled by adding adversarial sentences (Jia et al., ACL’17)

Adversarial Examples for Training / Regularization

- Make models adhere to higher-level rules
- What are these rules, how can we formulate / integrate them?
  - Appending Sentences + KB rules (Jia et al. 2017)
  - Erasing words (Li et al. 2017)
  - Character flips (Ebrahimi et al. 2018)
  - Paraphrases (Iyyer et al. 2018)
  - Semantic equivalence (Ribeiro et al. 2018)
  - KB rules (Minervini et al. 2018)

Data augmentation

Adversarial regularisation

Challenge 2: Solvability

Can the question actually be answered? (Rajpurkar et al. 2018)

What was the name of the 1937 treaty?

[UNANSWERABLE]

… Other legislation followed, including the Migratory Bird Conservation Act of 1929, a 1937 treaty prohibiting the hunting of right and gray whales, and the Bald Eagle Protection Act of 1940.

From: Rajpurkar et al. 2018

Challenge 2: Solvability

Can the question actually be answered? (Rajpurkar et al. 2018)

What was the name of the 1937 treaty?

[UNANSWERABLE]

… Other legislation followed, including the Migratory Bird Conservation Act of 1929, a 1937 treaty prohibiting the hunting of right and gray whales, and the Bald Eagle Protection Act of 1940.

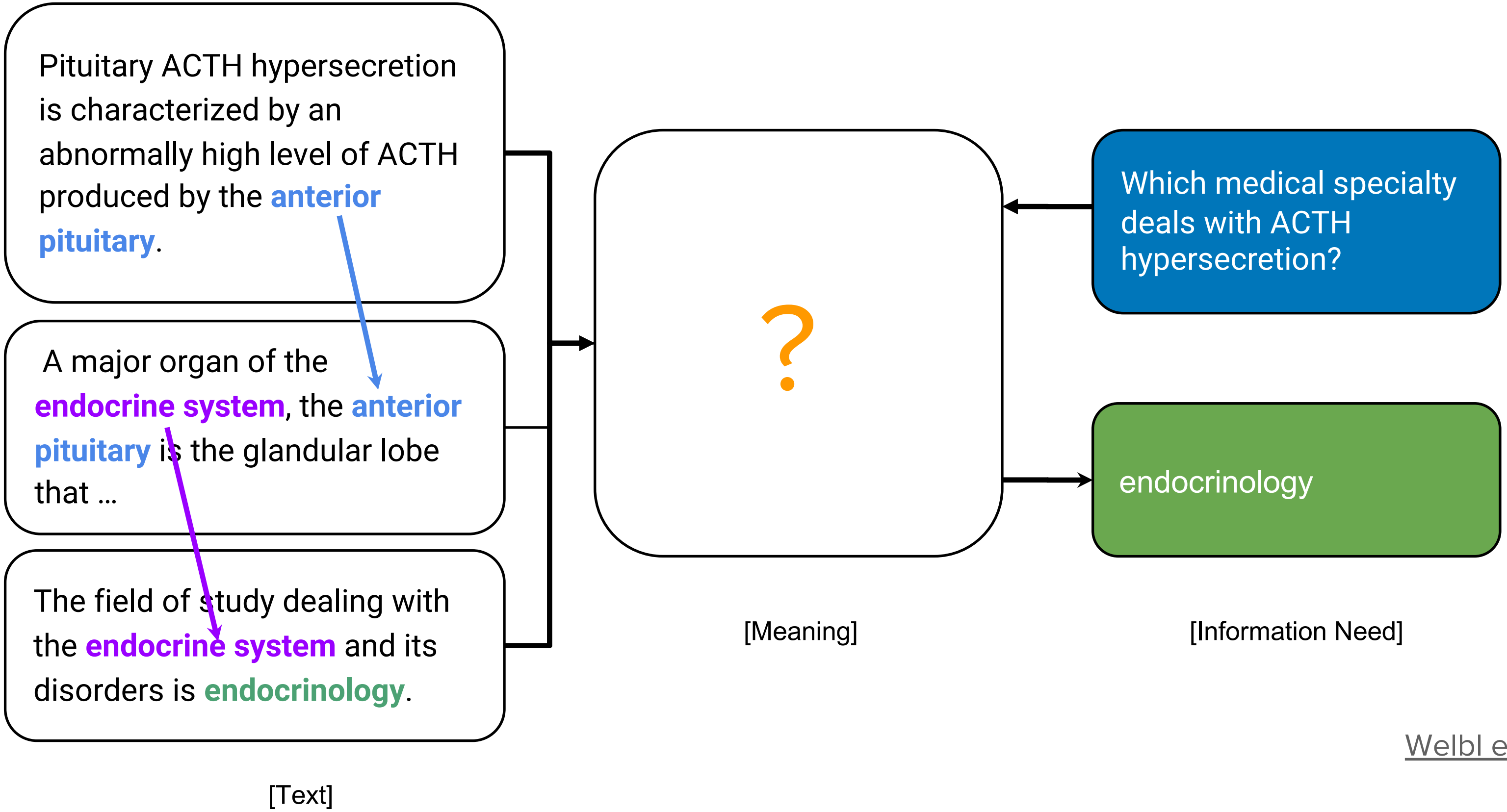
System	SQuAD 1.1 test EM	SQuAD 1.1 test F1	SQuAD 2.0 dev EM	SQuAD 2.0 dev F1	SQuAD 2.0 test EM	SQuAD 2.0 test F1
BNA	68.0	77.3	59.8	62.6	59.2	62.1
DocQA	72.1	81.0	61.9	64.8	59.3	62.3
DocQA + ELMo	**78.6**	**85.8**	**65.1**	**67.6**	**63.4**	**66.3**
Human	82.3	91.2	86.3	89.0	86.9	89.5
Human–Machine Gap	3.7	5.4	**21.2**	**21.4**	**23.5**	**23.2**

From: Rajpurkar et al. 2018

Challenge 3: Limited Supervision

- Strong results with large annotated training sets
- How about smaller datasets?
  - Ideally: shift from 100K to 1K training points
  - less costly, large-scale annotation
- Approaches:
  - domain adaptation, e.g. Wiese et al. (2017)
  - Synthetic data generation, e.g. Dhingra et al. (2018)
  - transfer learning, e.g. Mihaylov et al. (2017)
  - **(un?-)supervised pretraining, e.g. ELMo, Peters et al. (2018)**

Challenge 4: Reasoning with Text



Welbl et al. (2018)

Challenge 4: Reasoning with Text

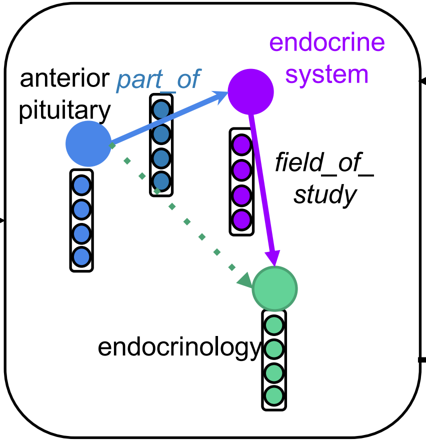
Pituitary ACTH hypersecretion is characterized by an abnormally high level of ACTH produced by the **anterior pituitary**.

A major organ of the **endocrine system**, the **anterior pituitary** is the glandular lobe that …

The field of study dealing with the **endocrine system** and its disorders is **endocrinology**.

[Text]

Reasoning with Structured Knowledge



[Meaning]

Which medical specialty deals with ACTH hypersecretion?

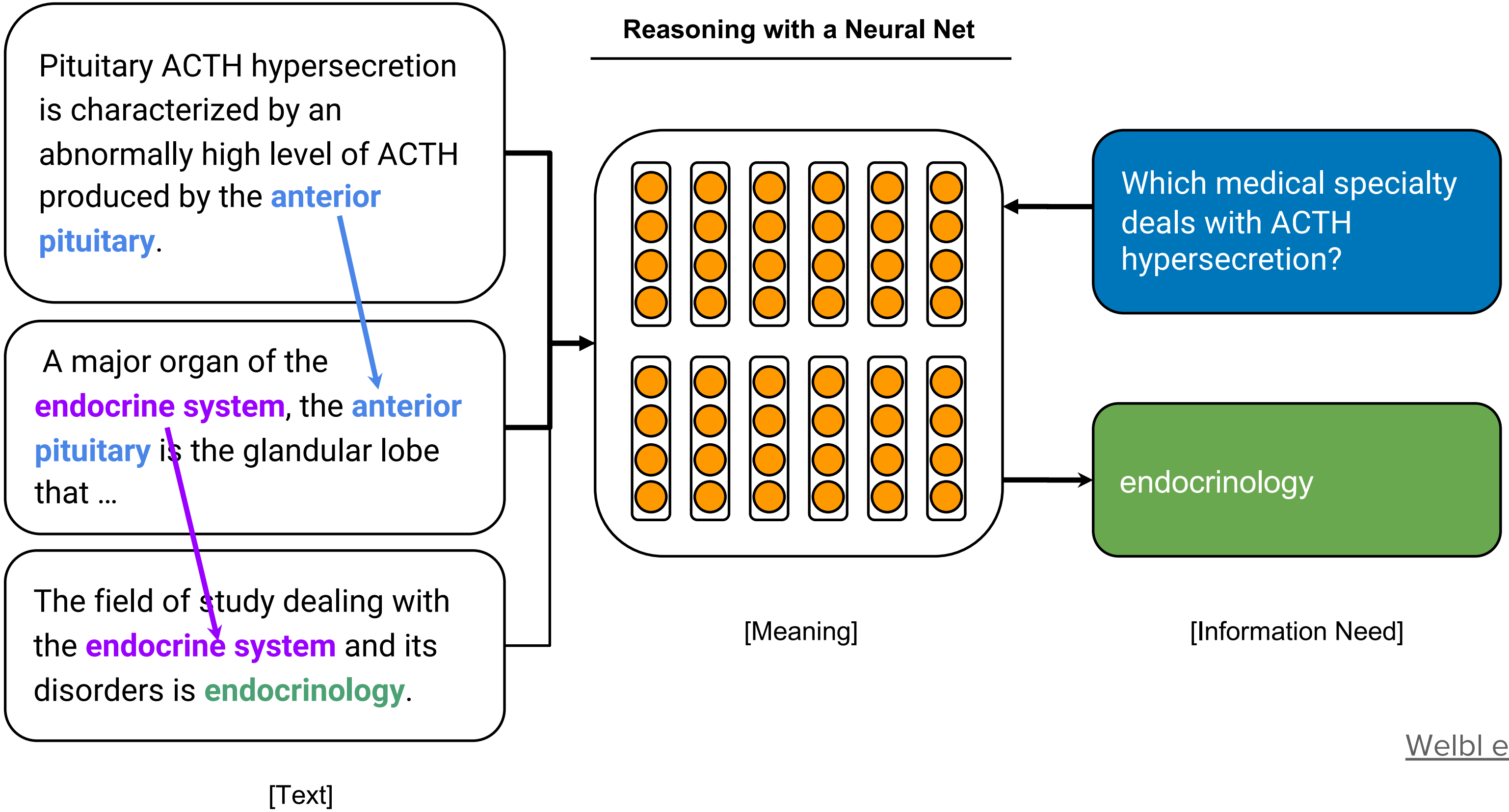
endocrinology

[Information Need]

part_of(A,B) ∧ *lfield_of_study*(B,C)
→ *medical_specialty*(A,C)

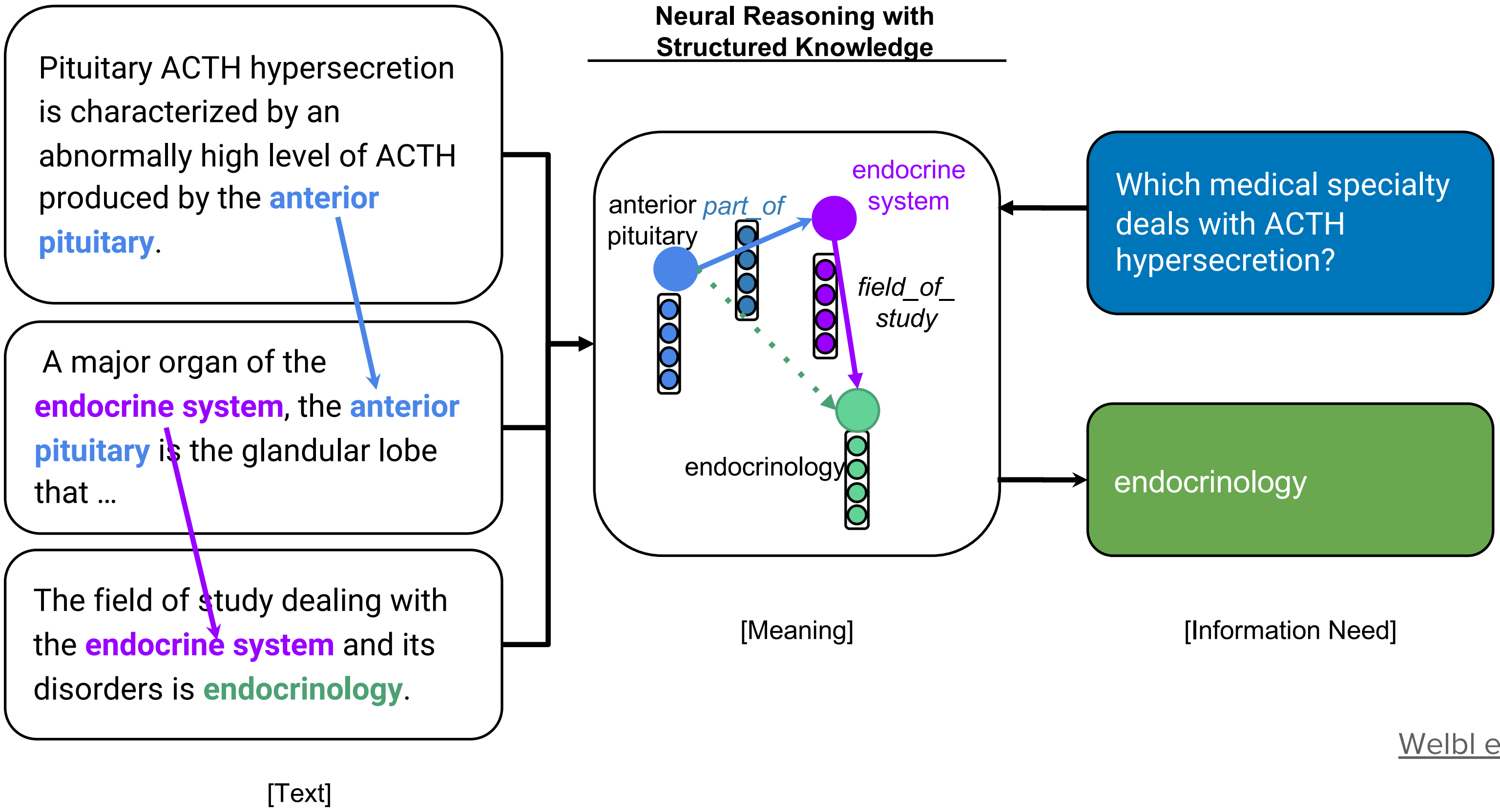
Welbl et al. (2018)

Challenge 4: Reasoning with Text



Welbl et al. (2018)

Challenge 4: Reasoning with Text



Welbl et al. (2018)

Summary: Where models work well today

- Question is answerable
- Relevant paragraph not too long
- Inferring answer is not too complex
- Pattern matching / soft text alignment between question and text
- Same domain during training and test time
- Relevant paragraph / text is given

Is all this model complexity necessary?

- BERT has 340M parameters
- 1-layer BiLSTM with a word-in-question feature works well on SQuAD (Weissenborn 17)

Should we rather:

- Build model architectures more carefully?
- Think more carefully about our training data?

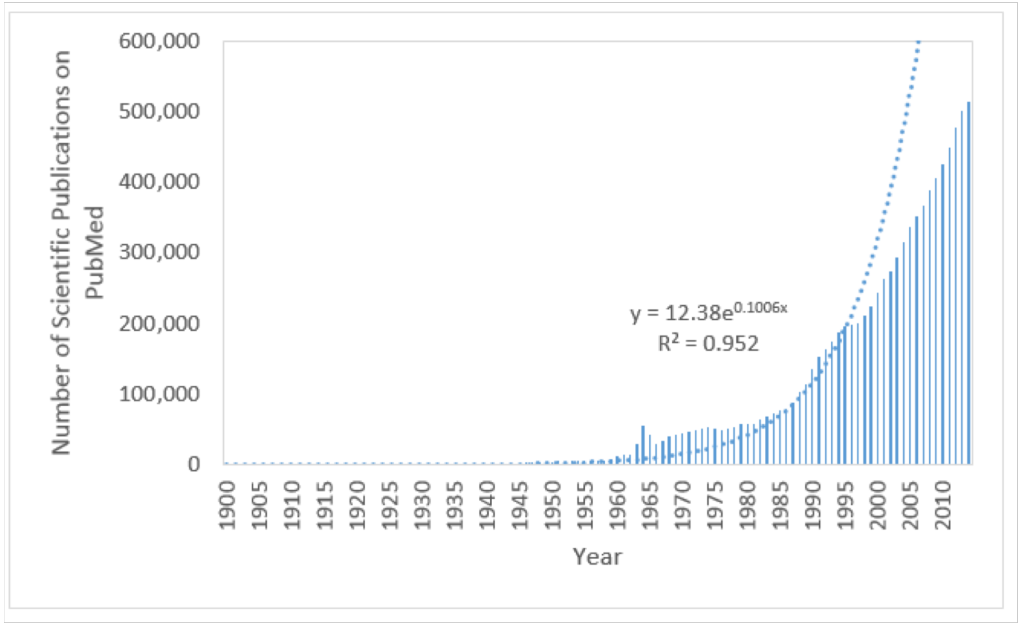
Take home:

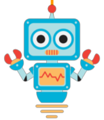
- **Don’t over-engineer** before establishing a decent baseline
- **Look at your datasets**! Are they challenging enough for the research you want to conduct?

Open Domain Question Answering

Is Machine Reading actually useful?

Motivation 1: Information Overload



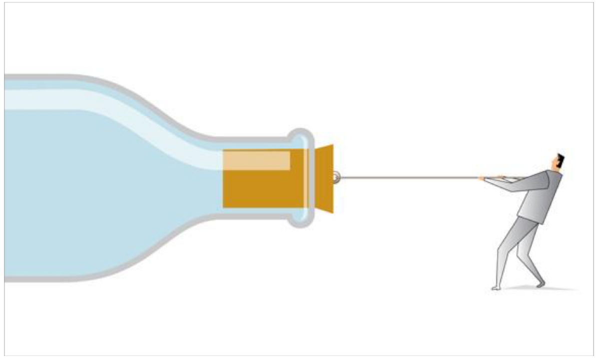


[Information Need]

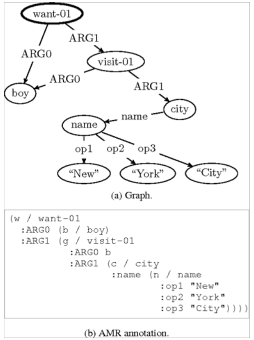
uses for

Motivation 2: The Knowledge Acquisition Bottleneck

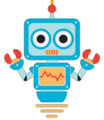
“The problem of knowledge acquisition is the critical bottleneck problem in artificial intelligence.”
E. A. Feigenbaum 1984



[Knowledge]



[Meaning]



uses for

[Information Need]

Open domain Question Answering

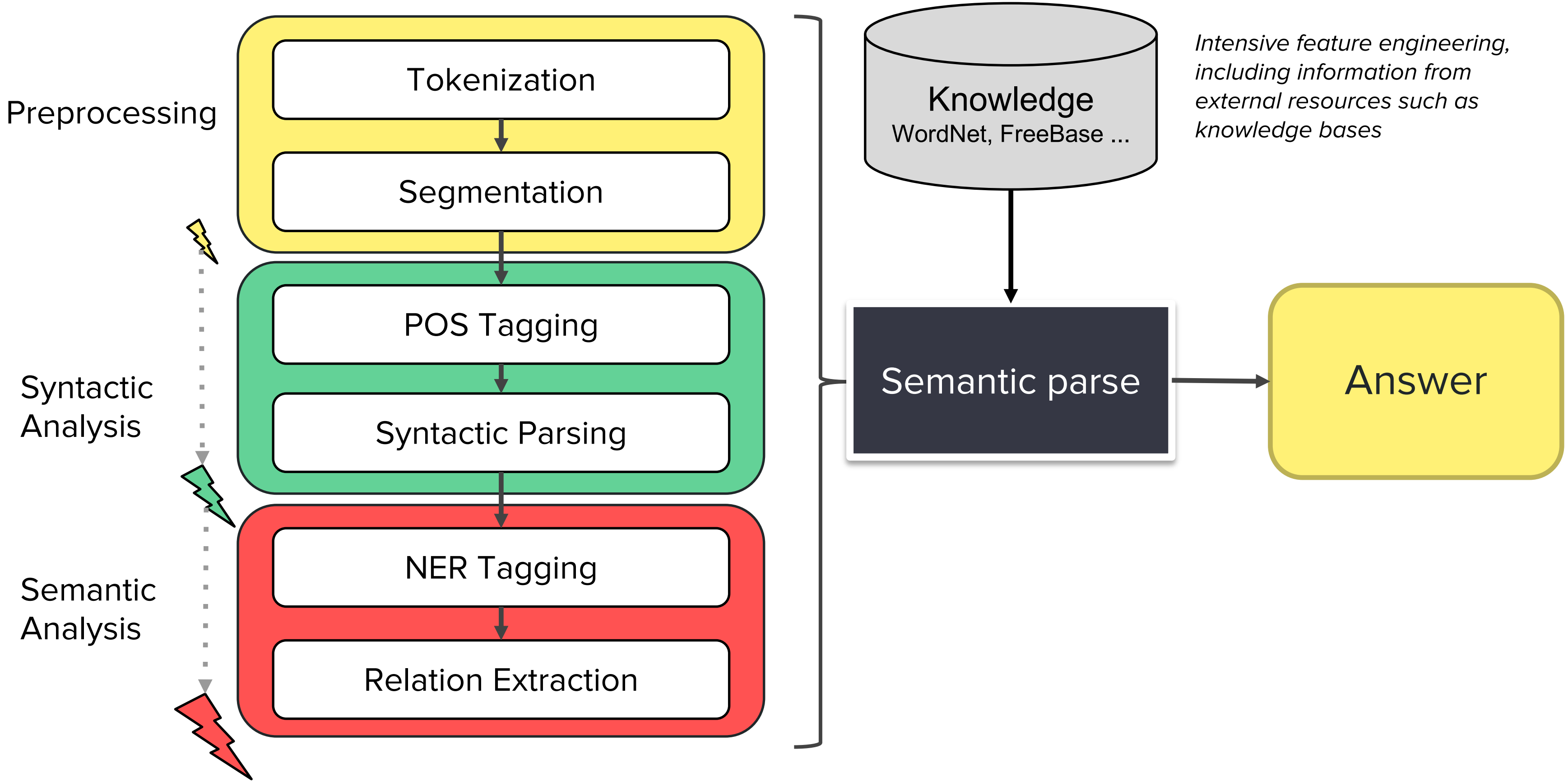
- Open domain QA: answer any question using very large knowledge sources
- Goes beyond Machines Reading that expects a paragraph to be given

- Open domain = question on any topic not a restricted subset

- In the following

1. Traditional approaches using Knowledge Bases
2. New approaches based on end-to-end Machine Reading

“Traditional” NLP for open domain QA



Semantic Parsing

Ewan forgot the mozarella in his car

$$\exists x0\ \mathrm{named}(x0, ewan, person) \wedge$$
$$\exists x1\ \mathrm{mozzarella}(x1) \wedge$$
$$\exists x2\ \mathrm{car}(x2) \wedge \mathrm{of}(x2,x0) \wedge \mathrm{in}(x1, x2) \wedge$$
$$\exists e\ \ \mathrm{event}(e) \wedge \mathrm{forget}(e) \wedge \mathrm{agent}(e, x0) \wedge \mathrm{patient}(e, x1)$$

[Text] [Meaning] [Information Need]

Semantic parses are logical forms in PROLOG, SQL, SPARQL, etc.

Knowledge Bases

- KB: structured repository of knowledge (usually relational DB)
- Goal: encode knowledge so that it can be queried by semantic parses **efficiently**
- Scale can be huge: billions of facts, millions of entities

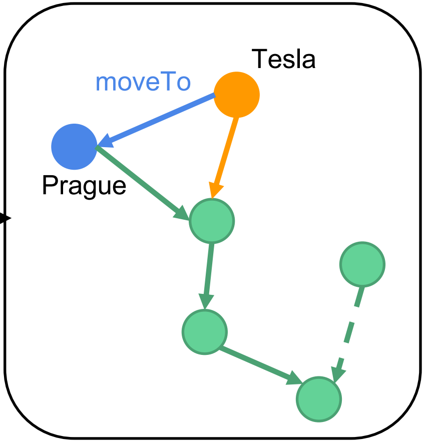
- KB can be generic or specific
- Examples: Cyc, WikiData, DBPedia, Google KG, GeneOntology, IMDB, etc.

- Key challenge is their construction!
  - Manually: Crowdsourcing, paid experts
  - Automatically: Information extraction or Automatic KB Construction

Automatic Knowledge Base Construction

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.

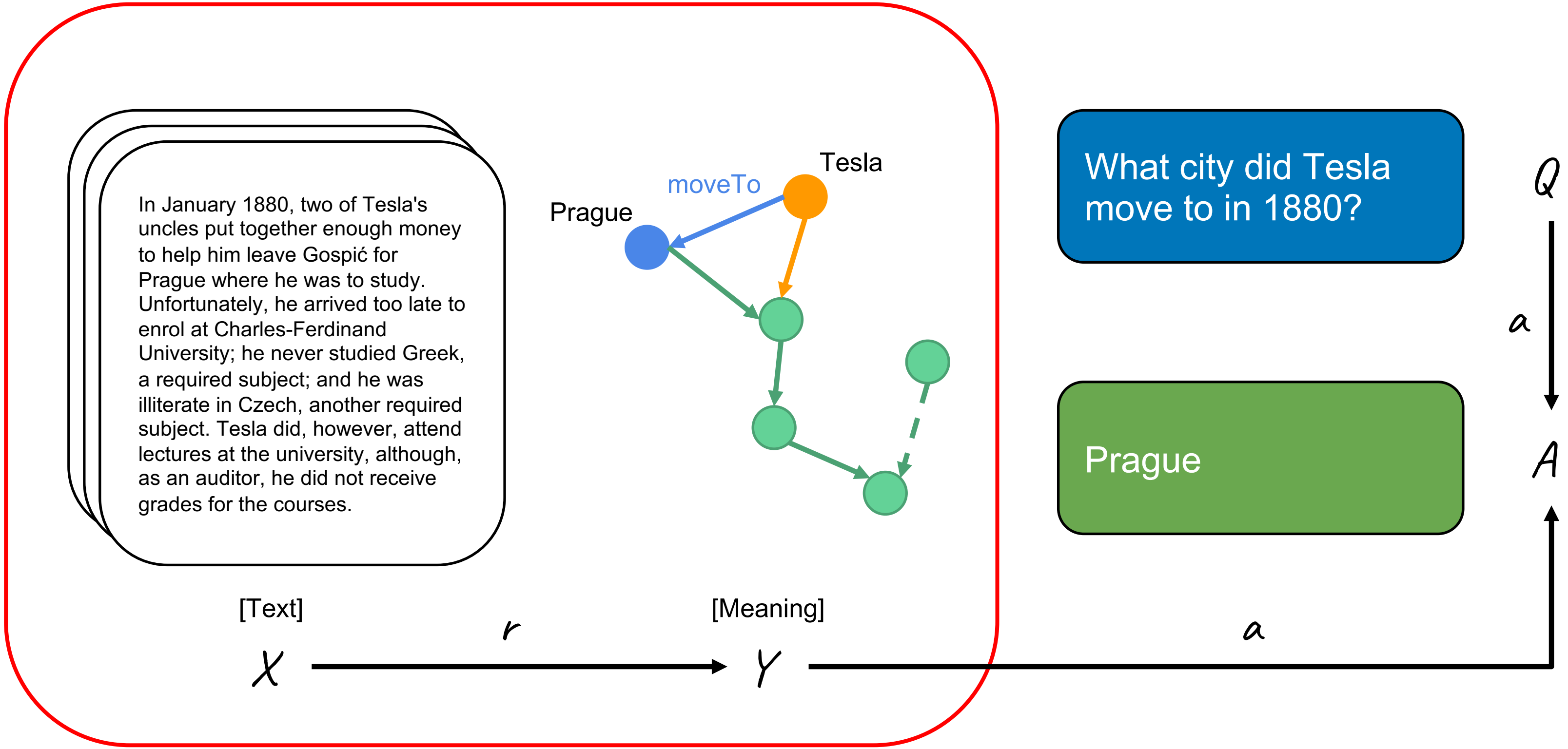
[Text]



[Meaning]

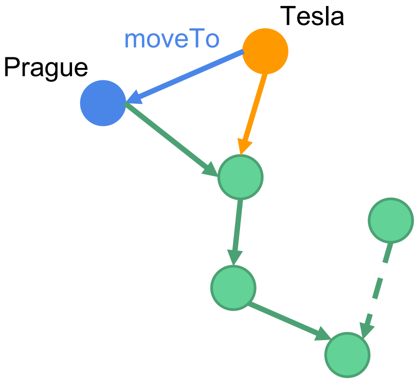
[Information Need]

Knowledge Graph Construction



Knowledge Graph Construction

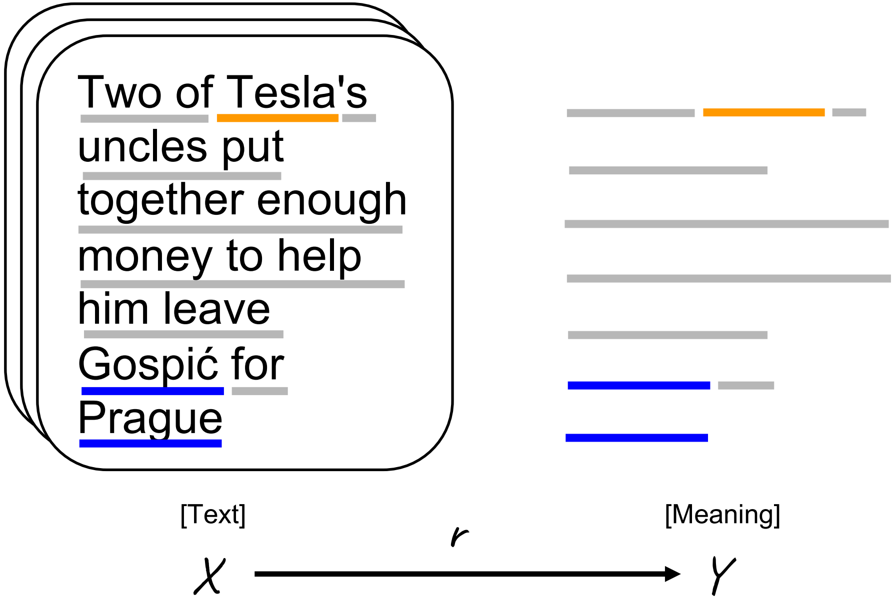
In January 1880, two of Tesla's uncles put together enough money to help him leave Gospić for Prague where he was to study. Unfortunately, he arrived too late to enrol at Charles-Ferdinand University; he never studied Greek, a required subject; and he was illiterate in Czech, another required subject.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



[Text] [Meaning]

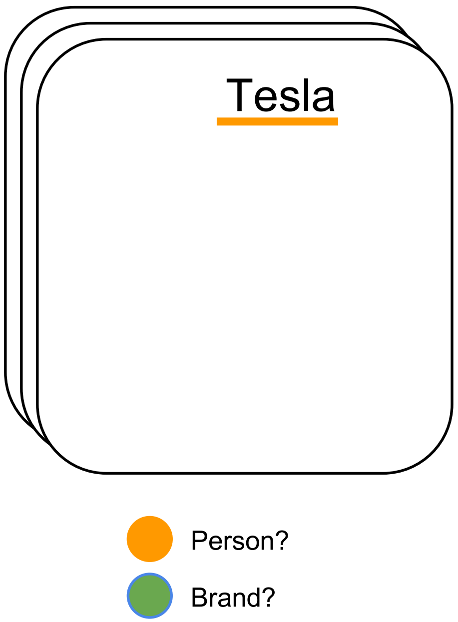
$$X \xrightarrow{r} Y$$

Entity Extraction



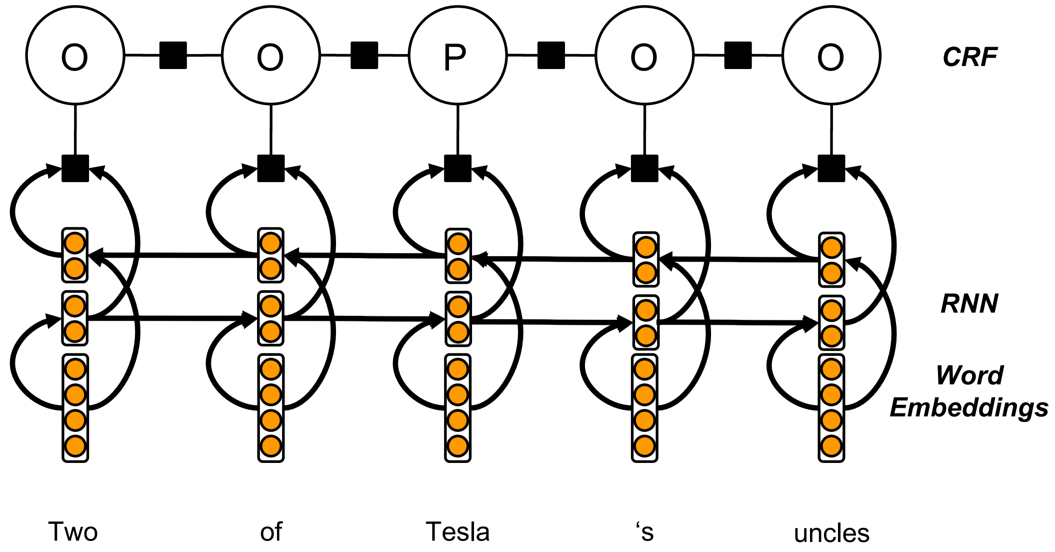
- Linear Chain CRF
- Bi-directional RNNs
- Hybrid RNN & CRFs

Challenge: Ambiguity

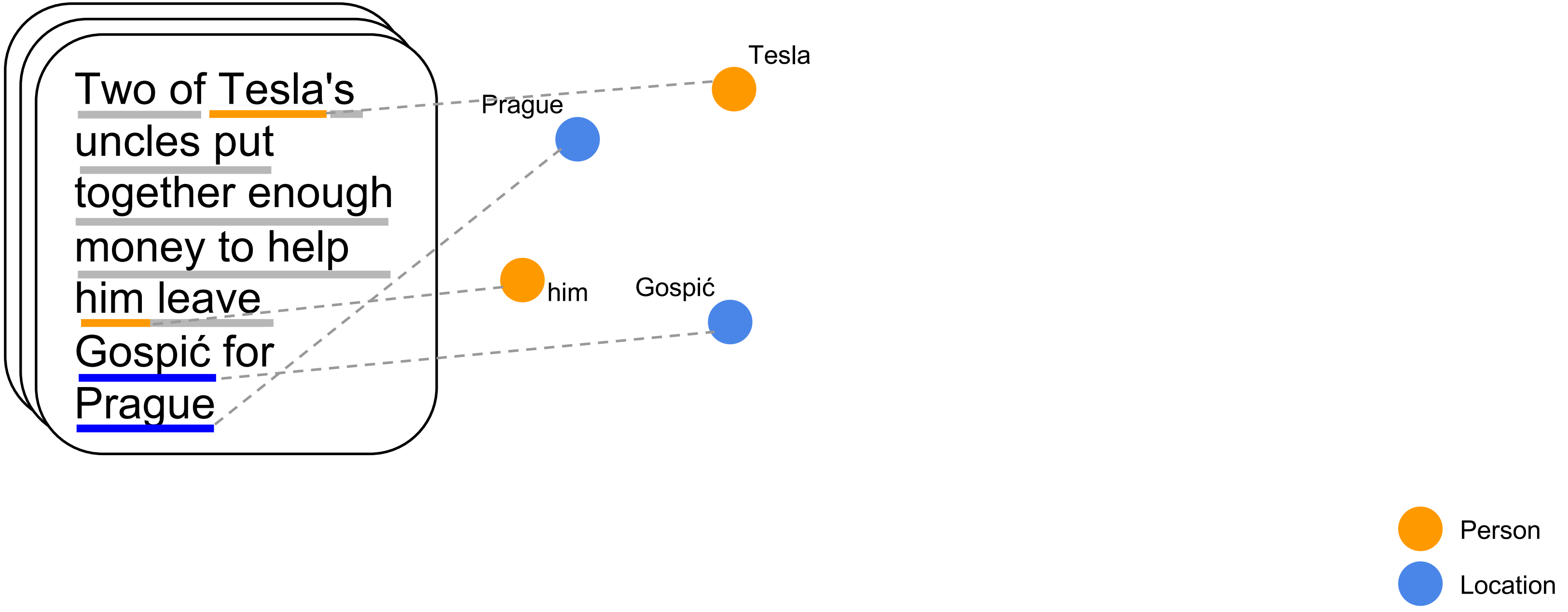


Conditional Random Fields with RNN Potentials

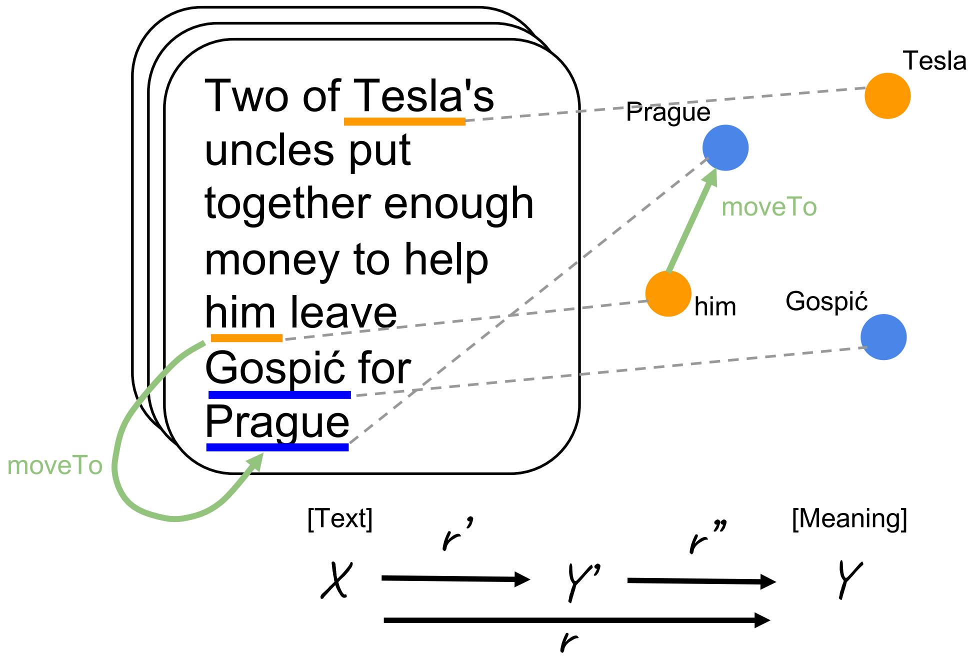
Huang et al., 2015



Instantiate Nodes



Relation Extraction



- Neural Classification
- Distant Supervision

Challenge: Variation

Two of Tesla's uncles put together enough money to help **him leave Gospić for Prague**

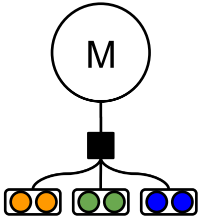
Two of Tesla's uncles put together enough money to help **him move to Prague**

Two of Tesla's uncles put together enough money to help **him settle in Prague**

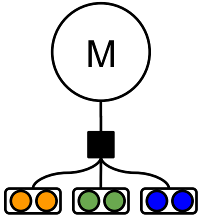
Relation Classification

[Current SOTA neural RE model]

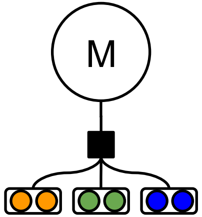
M (Tesla, moveTo, Prague)



him leave Gospić for Prague



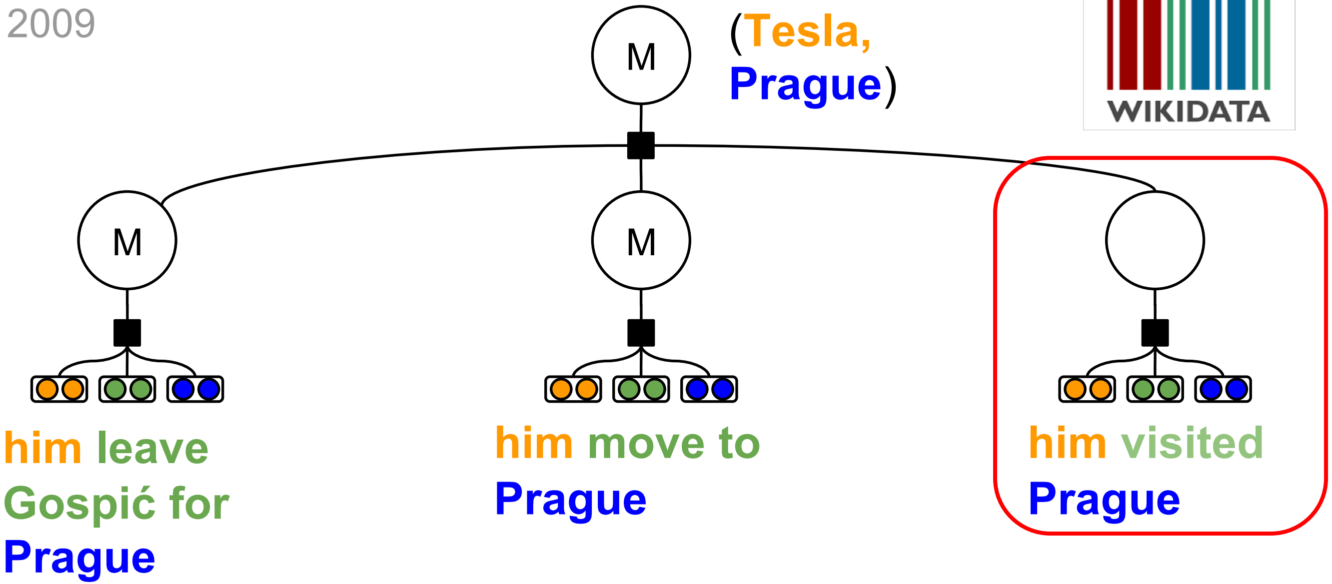
him move to Prague



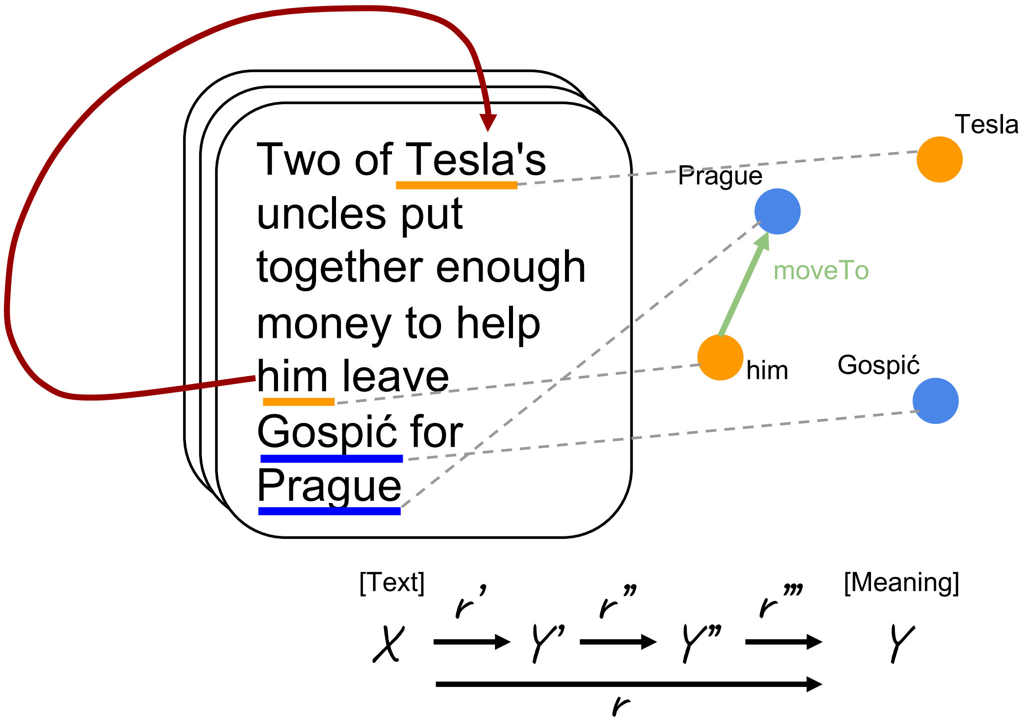
him settle in Prague

Distant Supervision & Multiple Instance Learning

Mintz et al., 2009

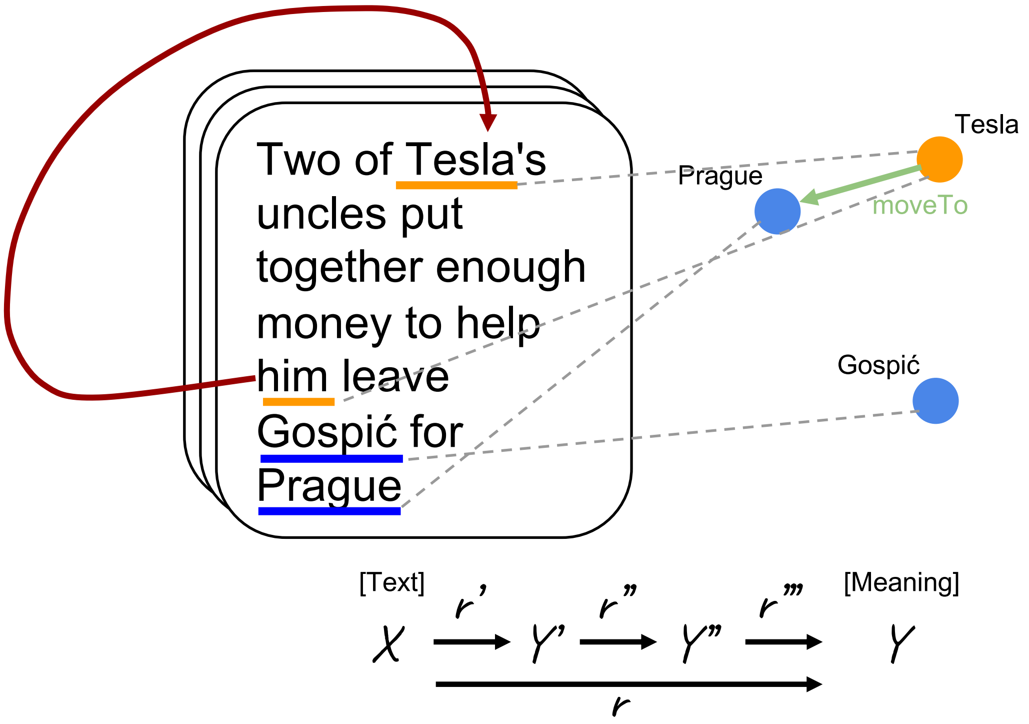


Coreference Resolution



- Neural Classification
- Latent Variables

Collapsing Nodes



Challenge: Common Sense

Two of Tesla's uncles put together enough money to help **him** leave Gospić for Prague

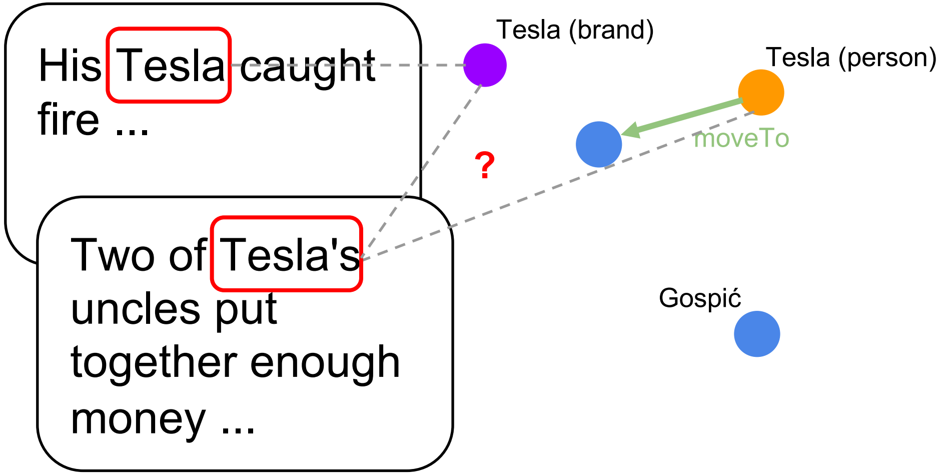
The **trophy** would not fit in the brown **suitcase** because **it** was too ***big***.

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

Surface

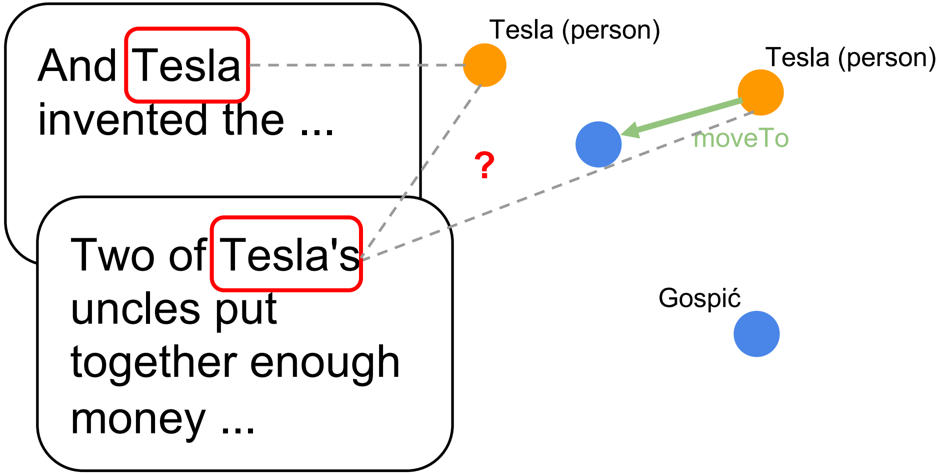
Common Sense

Entity Linking



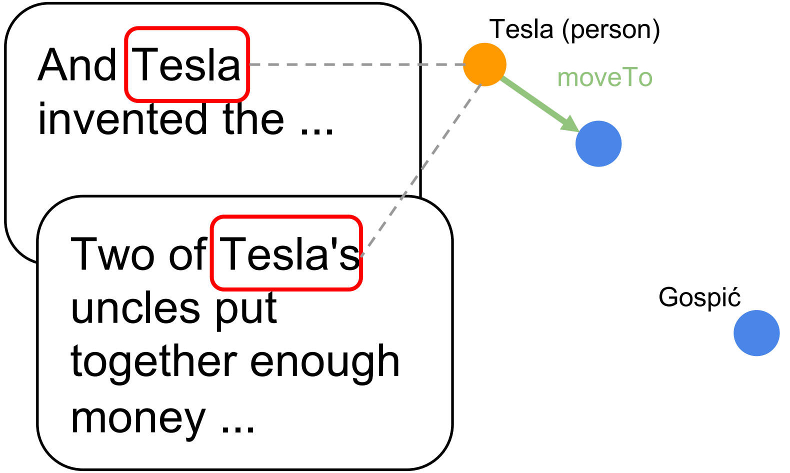
- Reranking ...
- Embeddings ...

Entity Linking



- Reranking ...
- Embeddings ...

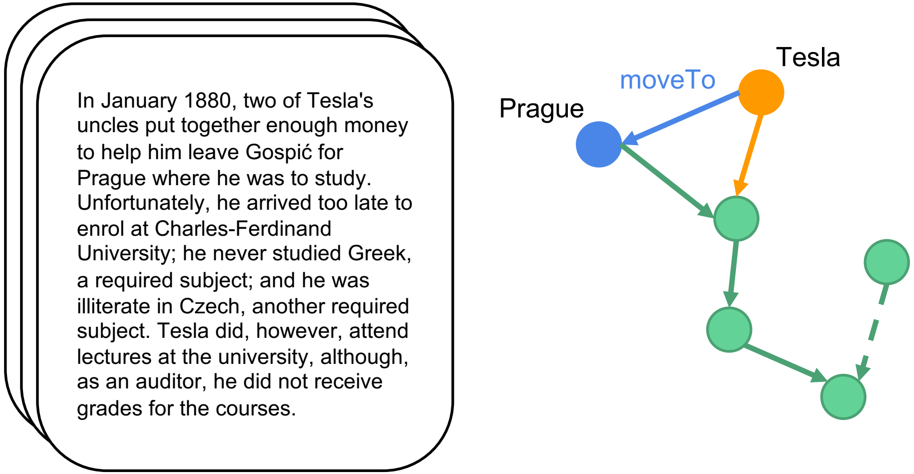
Collapsing



$$X \xrightarrow{r'} Y' \xrightarrow{r''} Y'' \xrightarrow{r'''} Y''' \xrightarrow{r''''} Y$$

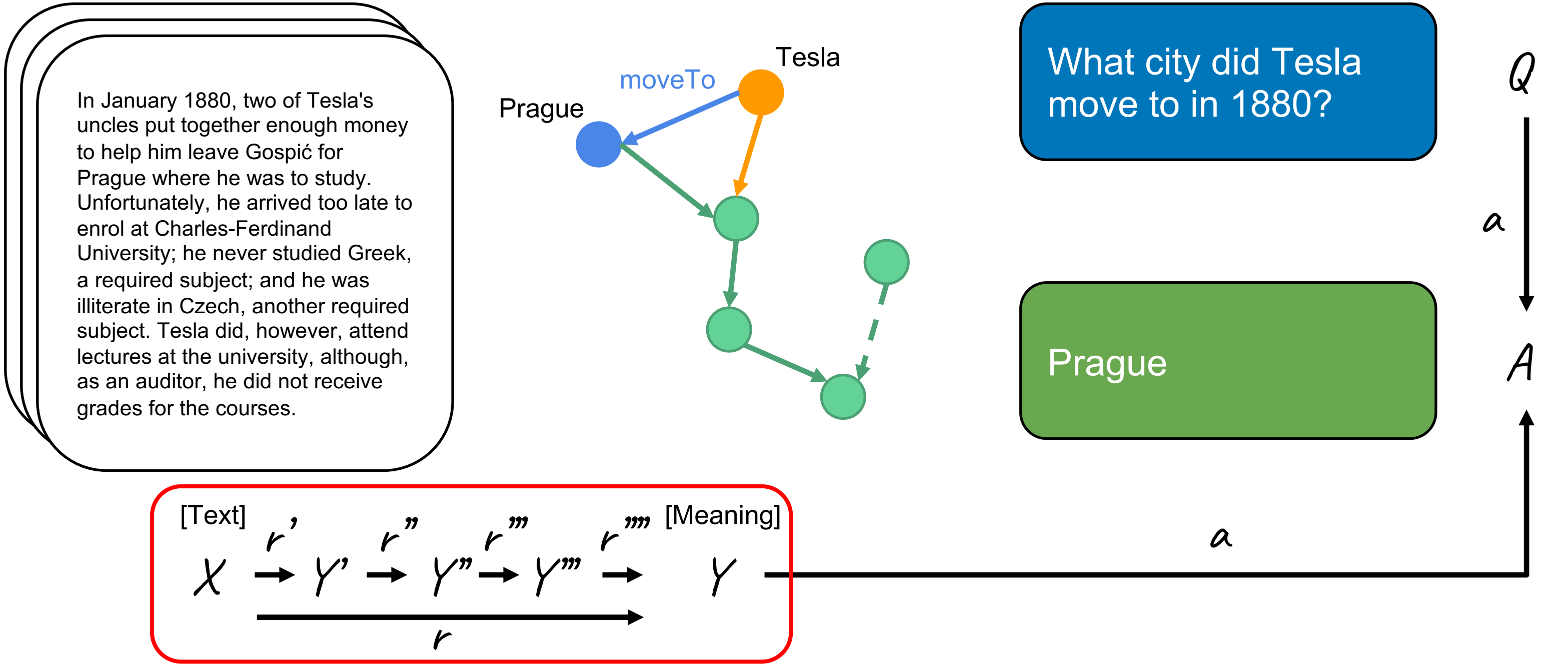
$$X \xrightarrow[r]{} Y$$

Strengths



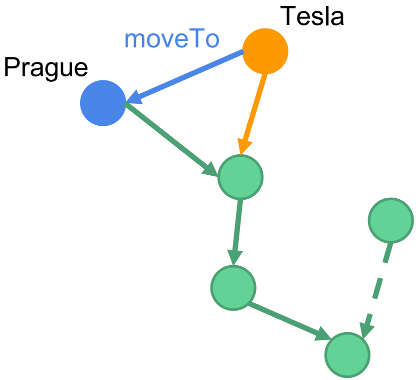
- Supports Reasoning
- Fast access
- Generalisation
- Interpretable
- Existing KBs can serve as supervision signal!

Weakness: Cascading errors



Weakness: Cascading errors

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.



What city did Tesla move to in 1880?

moveTo(Tesla,X)?

Prague

$$Q \xrightarrow{a'} Q' \xrightarrow{a''} A \qquad Q \xrightarrow{a} A$$

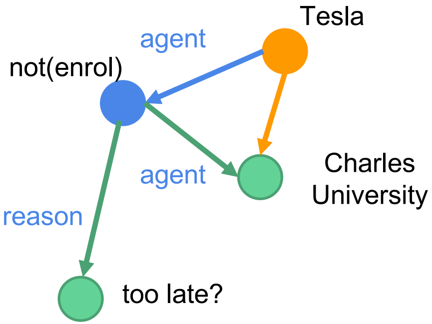
[Text] [Meaning]

$$X \xrightarrow{r'} Y' \xrightarrow{r''} Y'' \xrightarrow{r'''} Y''' \xrightarrow{r''''} Y \qquad X \xrightarrow{r} Y$$

$$Y \xrightarrow{a} A$$

Weakness: Engineering Schemas and Formalisms

Unfortunately, he arrived too late to enrol at Charles University



Why did he not enrol?

He arrived too late?

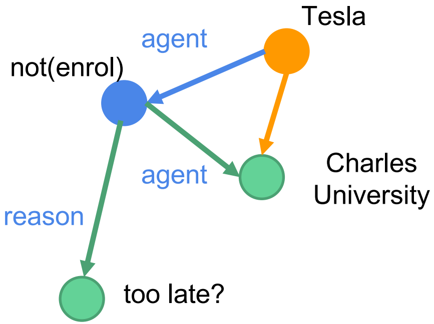
getting this right is hard

Weakness: Annotation

He et al., 2015

much easier

Unfortunately, he arrived too late to enrol at Charles University



Why did he not enrol?

He arrived too late?

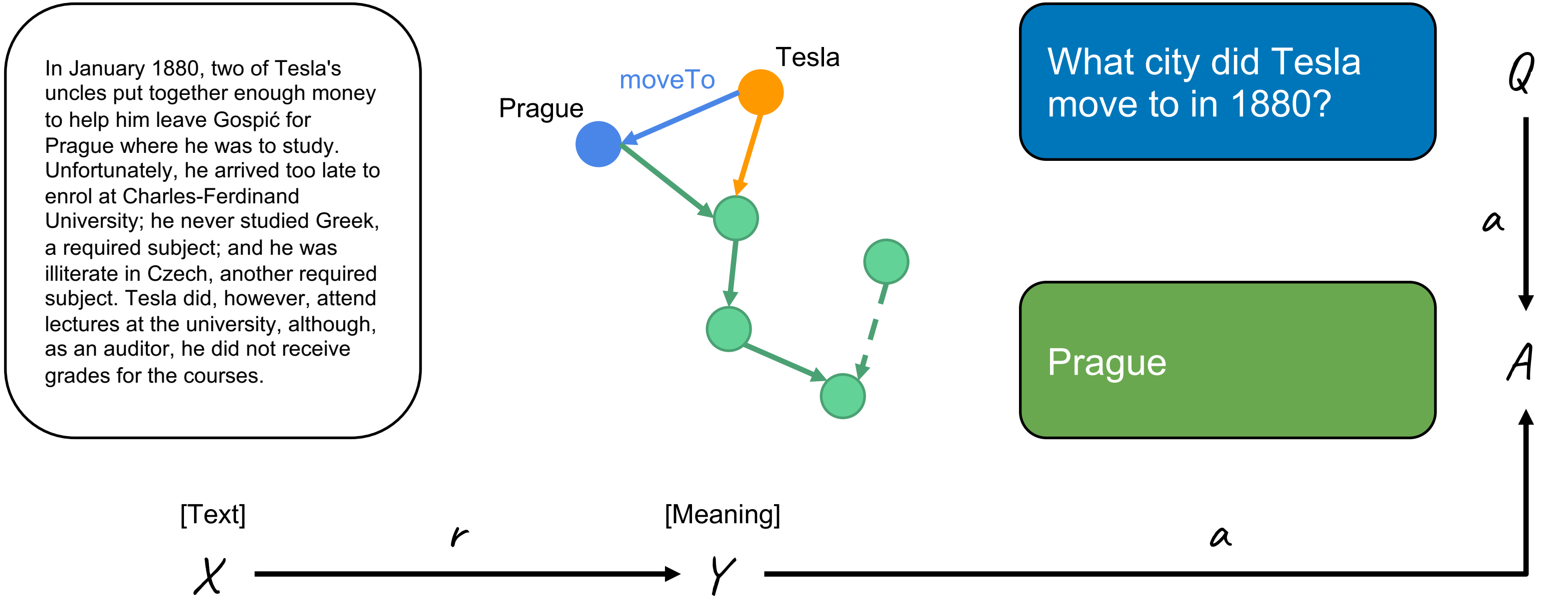
Hard to annotate

Structured Representations

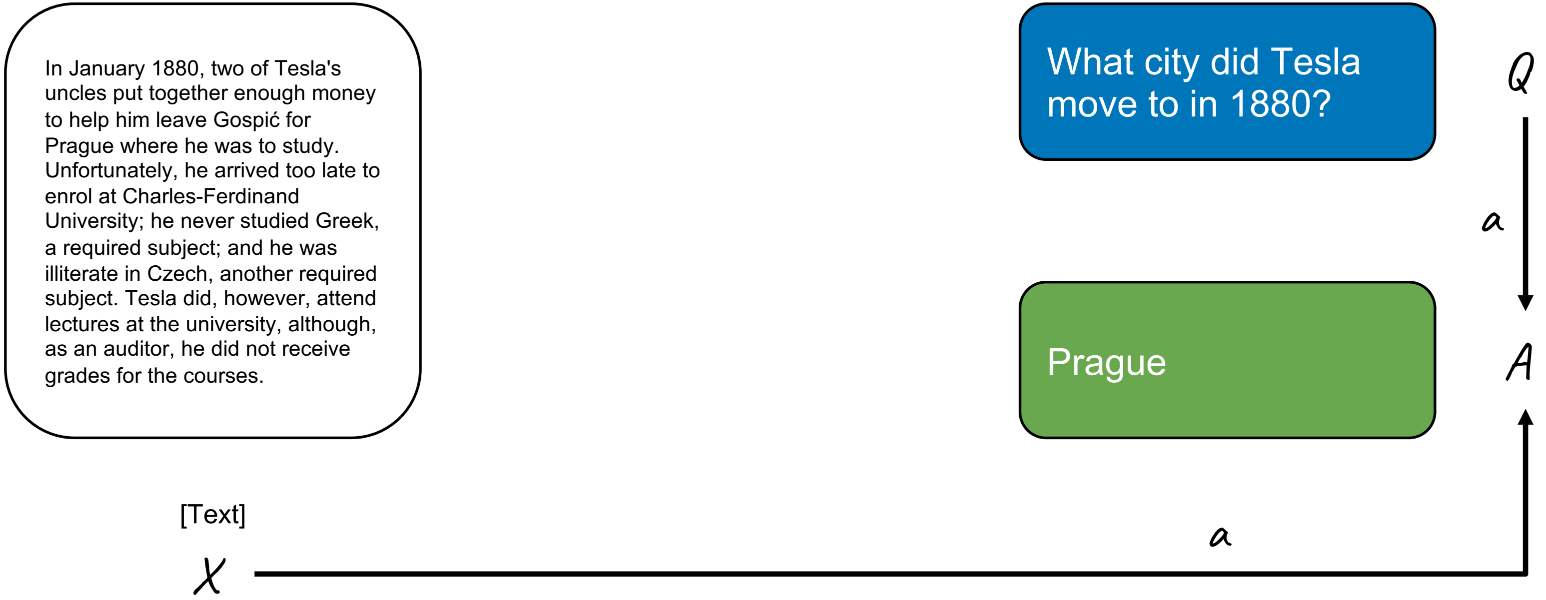
- Advantages
  - Fast access
  - Scalable
  - Interpretable
  - Supports reasoning
  - Universality of representations: independent of question

- Disadvantages
  - Less robust to variation in language
  - Cascading errors
  - Schema engineering
  - Annotation requires experts

Is there another way?



Omitting Symbolic Meaning Representations !!

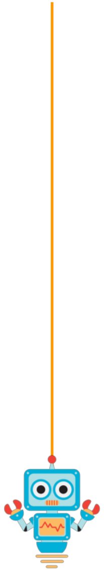


Machine Reading AT SCALE

A **machine** processes **a (very) large collection of texts** to satisfy an **information need**

Machine Reading

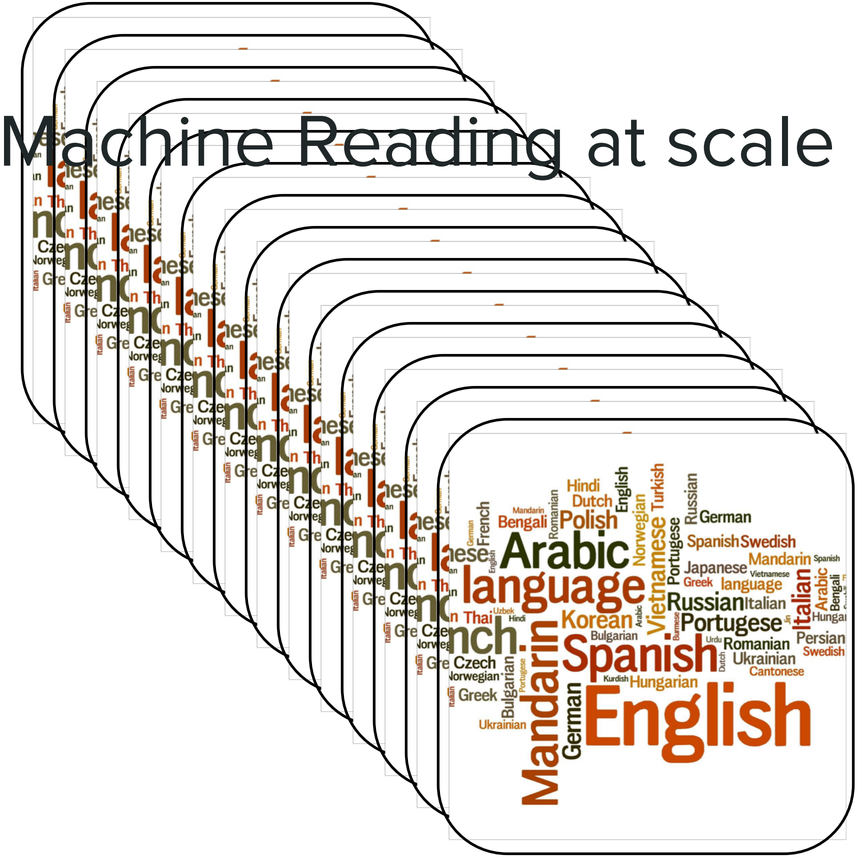
[Text]



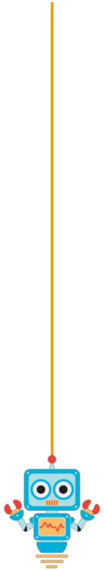
uses for

[Information Need]

Machine Reading at scale



[Large collection of Text]

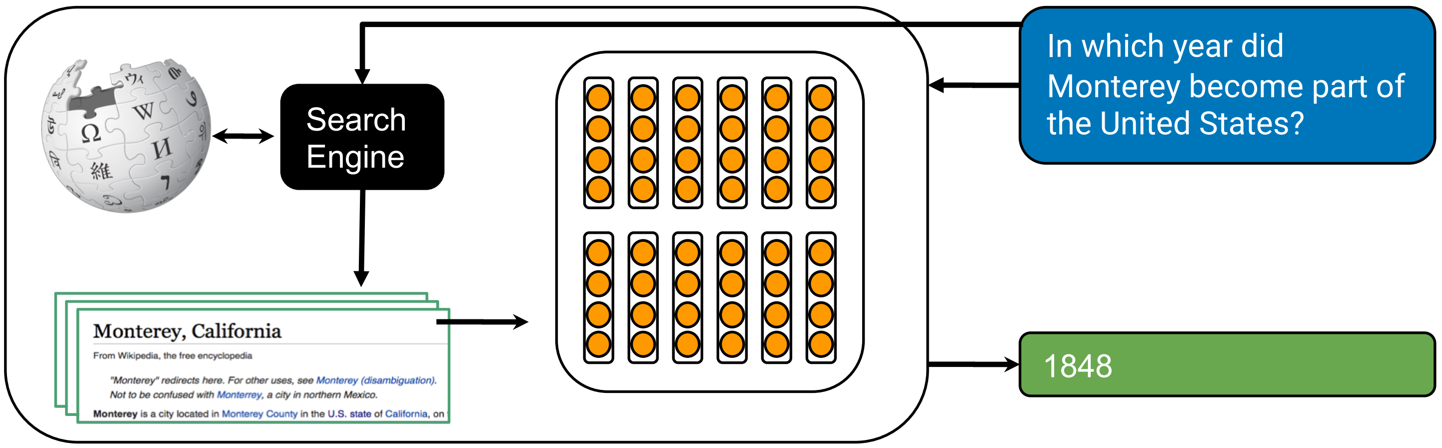


uses for

[Information Need]

Typical Machine Reading at Scale System

Dr.QA Chen et al., 2017

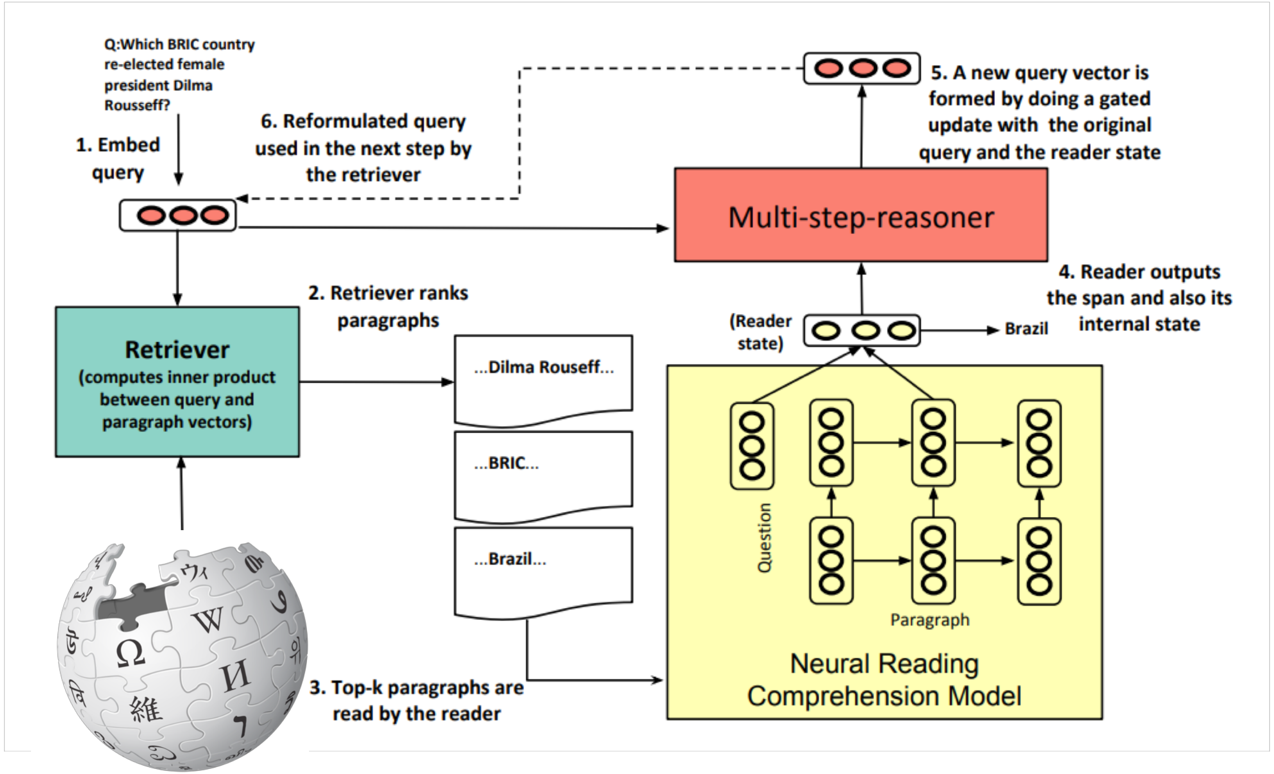


[Text] [Meaning] [Information Need]

No way to recover if the search engine is wrong!

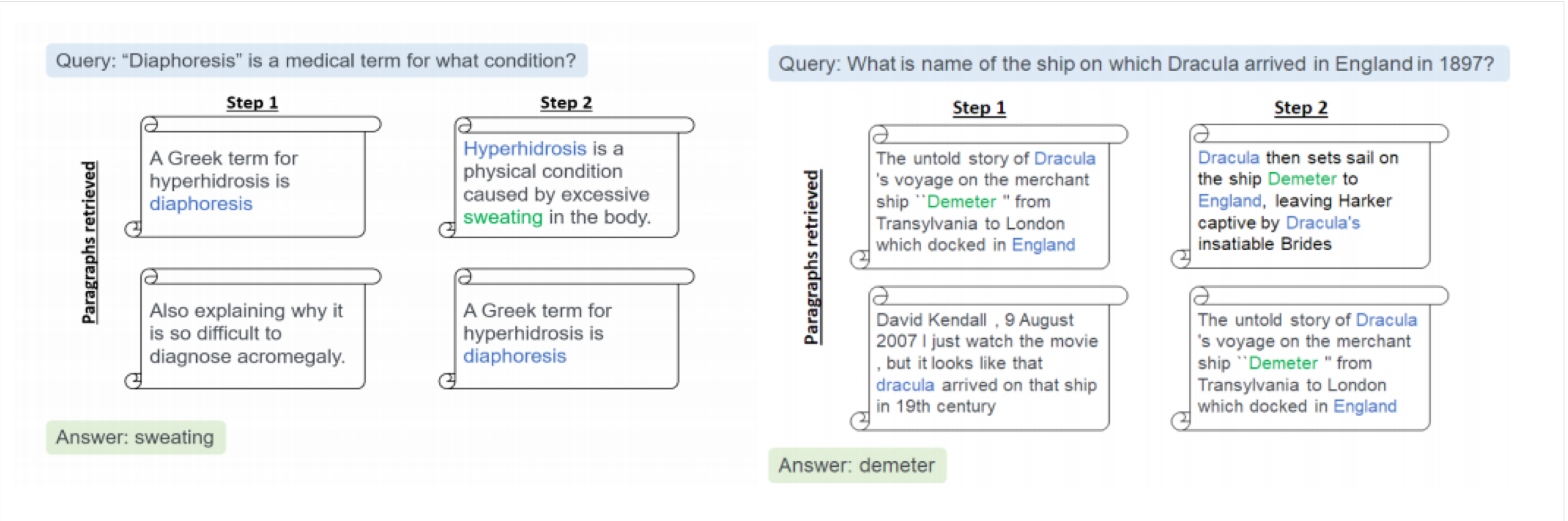
Current best: Multi-Step Retriever-Reader

Das et al., 2019



Current best: Multi-Step Retriever-Reader

Das et al., 2019

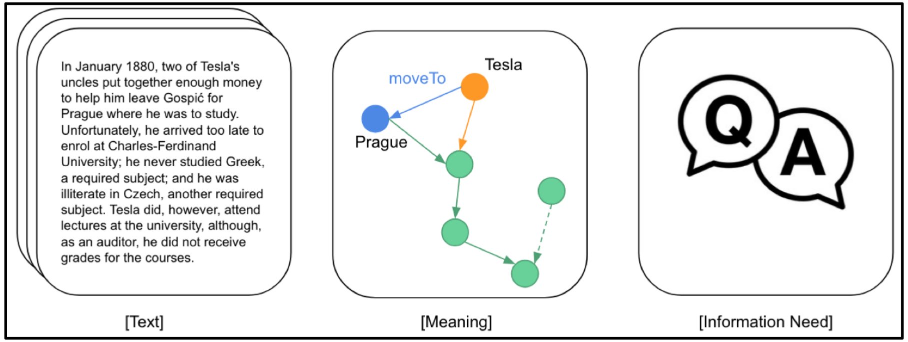


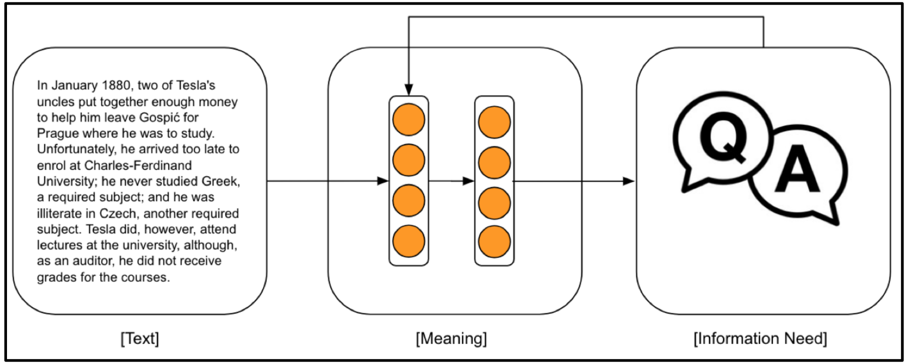
Between 40 and 60% of correct responses (for rather simple questions)

A Paradigm Shift

- Symbolic Meaning Representations
- ➡ Latent Vector Representations

- Feature Engineering & Domain Expertise
- ➡ Architecture Engineering & ML/DL Expertise





Pros and cons

End-to-end models	Symbolic systems
Neural Networks	*KBs, Inductive Logic Programming, etc.*
• Scale to very large datasets	• Small scale conditions
• Can be used by non domain experts	• Require heavy expert knowledge
• Robust to noise and ambiguity in data	• Very brittle with noisy, ambiguous data
• Game changers in multiple applications	• Limited applicative success
• Very data hungry (mostly supervised data)	
• Can't learn easily new tasks from old ones	ostractions
• Not interpretable	
• Relatively simple reasoning	

Great research opportunities!

Current Challenge: Reconciling Conflicting Information

So how much does the UK pay to the EU per week?

"Once we have settled our accounts, we will take back control of roughly **£350m** per week." *Boris Johnson*

"We are not giving £20bn a year or £350m a week to Brussels - Britain pays **£276m** a week to the EU budget because of the rebate." *BBC Reality Check*

"...When those are taken into account the figure is **£250m**." *Independent*

?

Trust into source, timeline, ...

Conclusion

- We’ve seen 2 approaches for building system to answer any question
- Most deployed systems still rely on traditional pipelines for the most part (+ some DL here and there)

- Why? Scale, reliability, interpretability

- Open questions:
    - All shortcomings of Machine Reading → Open domain QA. Need to solve them
    - Will pretrained contextual embeddings change everything forever?
    - Can we combine both symbolic and end-to-end approaches?

References Compositional Sequence Encoders

- Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. NAACL.
- McCann, B., Bradbury, J., Xiong, C., & Socher, R. (2017). Learned in translation: Contextualized word vectors. NIPS.
- Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving Language Understanding by Generative Pre-Training. arXiv.
- Howard, J. & Ruder, S. (2018). Universal Language Model Fine-tuning for Text Classification. ACL.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. & Polosukhin, I. (2017). Attention is all you need. NIPS.
- Cheng, J., Dong, L., & Lapata, M. (2016). Long short-term memory-networks for machine reading. EMNLP.
- Wang, W., Yang, N., Wei, F., Chang, B., & Zhou, M. (2017). Gated self-matching networks for reading comprehension and question answering. ACL.
- Yu, A. W., Dohan, D., Luong, M. T., Zhao, R., Chen, K., Norouzi, M., & Le, Q. V. (2018). QANet: Combining Local Convolution with Global Self-Attention for Reading Comprehension. ICLR.
- Yang, Z., Zhao, J., Dhingra, B., He, K., Cohen, W. W., Salakhutdinov, R., & LeCun, Y. (2018). GLoMo: Unsupervisedly Learned Relational Graphs as Transferable Representations. arXiv.
- Tai, K. S., Socher, R., & Manning, C. D. (2015). Improved semantic representations from tree-structured long short-term memory networks. ACL.
- Dyer, C., Kuncoro, A., Ballesteros, M., & Smith, N. A. (2016). Recurrent Neural Network Grammars. NAACL.

References Interaction

- Cho, K., Gulcehre, B. V. M. C., Bahdanau, D., Schwenk, F. B. H., & Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. EMNLP.
- Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. NIPS.
- Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. ICLR.
- Sukhbaatar, S., Weston, J., & Fergus, R. (2015). End-to-end memory networks. NIPS.
- Kumar, A., Irsoy, O., Ondruska, P., Iyyer, M., Bradbury, J., Gulrajani, I., ... & Socher, R. (2016). Ask me anything: Dynamic memory networks for natural language processing. ICML.
- Graves, A., Wayne, G., Reynolds, M., Harley, T., Danihelka, I., Grabska-Barwińska, A., ... & Badia, A. P. (2016). Hybrid computing using a neural network with dynamic external memory. Nature
- Grefenstette, E., Hermann, K. M., Suleyman, M., & Blunsom, P. (2015). NIPS.
- Henaff, M., Weston, J., Szlam, A., Bordes, A., & LeCun, Y. (2017). Tracking the world state with recurrent entity networks. ICLR.
- Rocktäschel, T., Grefenstette, E., Hermann, K. M., Kočiský, T., & Blunsom, P. (2016). Reasoning about entailment with neural attention. ICLR.
- Yu, A. W., Dohan, D., Luong, M. T., Zhao, R., Chen, K., Norouzi, M., & Le, Q. V. (2018). QANet: Combining Local Convolution with Global Self-Attention for Reading Comprehension. ICLR.

References

- Adversarial Examples for Evaluating Reading Comprehension Systems (Jia et al. 2017, EMNLP)
- Know What You Don't Know: Unanswerable Questions for SQuAD (Rajpurkar et al. 2018, ACL)
- Visual question answering: Datasets, algorithms, and future challenges (Kafle et al. 2017, Computer Vision and Image Understanding)
- Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering (Goyal et al. 2017, CVPR)
- Reading Wikipedia to Answer Open-Domain Questions (Chen et al. 2017, ACL)
- Event2Mind: Commonsense Inference on Events, Intents, and Reactions (Rashkin et al. 2018, arXiv)
- Semantically Equivalent Adversarial Rules for Debugging NLP Models (Ribeiro 2018, ACL)
- Understanding Neural Networks through Representation Erasure (Li et al. 2016, arXiv)
- HotFLip: White-Box Adversarial Examples for NLP (Ebrahimi et al. 2017, arXiv)
- Anchors: High-Precision Model-Agnostic Explanations (Ribeiro et al. 2018, AAAI)
- Deep contextualized word representations (Peters et al. 2018, NAACL)
- Learned in Translation: Contextualized Word Vectors (McCann et al. 2017, NIPS)
- Supervised Learning of Universal Sentence Representations from Natural Language Inference Data (Conneau et al. 2017, EMNLP)
- Efficient Estimation of Word Representations in Vector Space (Mikolov et al. 2013, NIPS)
- Simple and Effective Semi-Supervised Question Answering (Dhingra et al. NAACL 2018)
- Neural Domain Adaptation for Biomedical Question Answering (Wiese et al. 2017, CoNLL)
- Improving Language Understanding by Generative Pre-Training (Radford et al. 2018, arXiv)
- Neural Skill Transfer from Supervised Language Tasks to Reading Comprehension (Mihaylov et al. 2017, arXiv)
- Representing General Relational Knowledge in ConceptNet 5 (Speer and Havasi, LREC 2012)
- Learning to understand phrases by embedding the dictionary (Hill et al. 2016, TACL)
- Leveraging knowledge bases in lstms for improving machine reading (Yang et al. 2017, ACL)
- Knowledgeable Reader: Enhancing Cloze-Style Reading Comprehension with External Commonsense Knowledge. (Mihaylov and Frank, 2018, ACL)
- Reading Wikipedia to Answer Open-Domain Questions (Chen et al. 2017, ACL)
- Evidence aggregation for answer re-ranking in open-domain question answering (Wang et al. ICLR 2018)
- Marco Baroni and Gemma Boleda: https://www.cs.utexas.edu/~mooney/cs388/slides/dist-sem-intro-NLP-class-UT.pdf
- News article: https://www.independent.co.uk/infact/brexit-second-referendum-false-claims-eu-referendum-campaign-lies-fake-news-a8113381.html

References for Datasets

- Building a question answering test collection, *Voorhees & Tice* SIGIR 2000
- Besting the Quiz Master: Crowdsourcing Incremental Classification Games, *Boyd-Graber et al.* EMNLP 2012
- Semantic Parsing on Freebase from Question-Answer Pairs, *Berant et al.* EMNLP 2013
- Mctest: A challenge dataset for the open-domain macchine comprehension of text, *Richardson et al.* EMNLP 2013
- Teaching Machines to Read and Comprehend, *Hermann et al.* NIPS 2015
- WikiQA: A challenge dataset for open-domain question answering, *Yang et al.* EMNLP 2015
- Large-scale Simple Question Answering with Memory Networks, *Bordes et al. 2015* arXiv:1506.02075.
- The Goldilocks Principle: Reading Children's Books with Explicit Memory Representations, *Hill et al.* ICLR 2016
- SQuAD: 100,000+ Questions for Machine Comprehension of Text, *Rajpurkar et al.* EMNLP 2016
- [SQuAD 2.0] Know What You Don't Know: Unanswerable Questions for SQuAD, *Rajpurkar and Jia et al.* ACL 2018
- Towards AI-Complete Question Answering: A Set of Prerequisite Toy Tasks, *Weston et al.* ICLR 2016
- Constraint-Based Question Answering with Knowledge Graph, *Bao et al*. COLING 2016
- MovieQA: Understanding Stories in Movies through Question-Answering, *Tapawasi et al.* CVPR 2016
- Who did What: A Large-Scale Person-Centered Cloze Dataset, *Onishi et al.* EMNLP 2016
- MS MARCO: A Human Generated MAchine Reading COmprehension Dataset, *Nguyen et al.* NIPS 2016
- The LAMBADA dataset: Word prediction requiring a broad discourse context, *Paperno et al.* ACL 2016
- WIKIREADING: A Novel Large-scale Language Understanding Task over Wikipedia, *Hewlett et al.* ACL 2016
- TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension, *Joshi et al.* ACL 2017
- Crowdsourcing Multiple Choice Science Questions, *Welbl et al.* WNUT 2017
- RACE: Large-scale ReAding Comprehension Dataset From Examinations, *Lai et al.* EMNLP 2017
- NewsQA: a Machine Comprehension Dataset, *Trischler et al.* RepL4NLP 2017
- Science Exam Datasets by the Allen Institute for Artificial Intelligence: https://allenai.org/data/data-all.html
- SearchQA: A New Q&A Dataset Augmented with Context from a Search Engine, *Dunn et al*. https://arxiv.org/pdf/1704.05179.pdf
- Quasar: Datasets for Question Answering by Search and Reading. *Dhingra et al.* 2017 https://arxiv.org/abs/1707.03904
- Constructing Datasets for Multi-Hop Reading Comprehension across Documents, *Welbl et al.* TACL 2018
- The NarrativeQA Reading Comprehension Challenge, *Kocisky et al.* TACL 2018