

# 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](mailto:angelafan@fb.com)

PhD Student FAIR Paris



**Louis**

[louismartin@fb.com](mailto:louismartin@fb.com)

PhD Student FAIR Paris

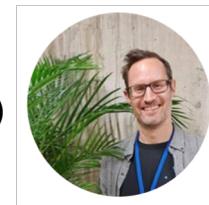


**Antoine**

[abordes@fb.com](mailto: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

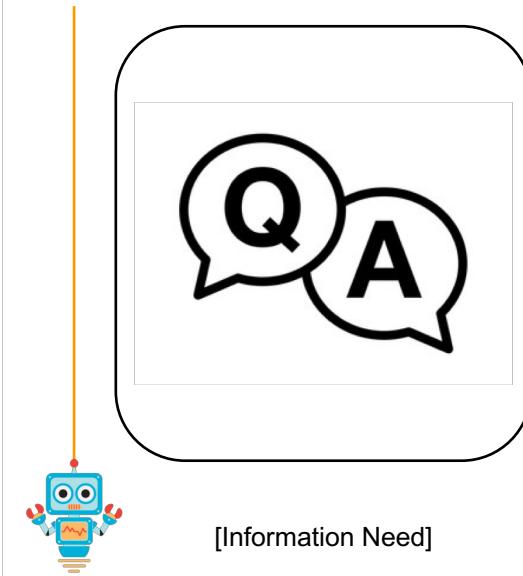
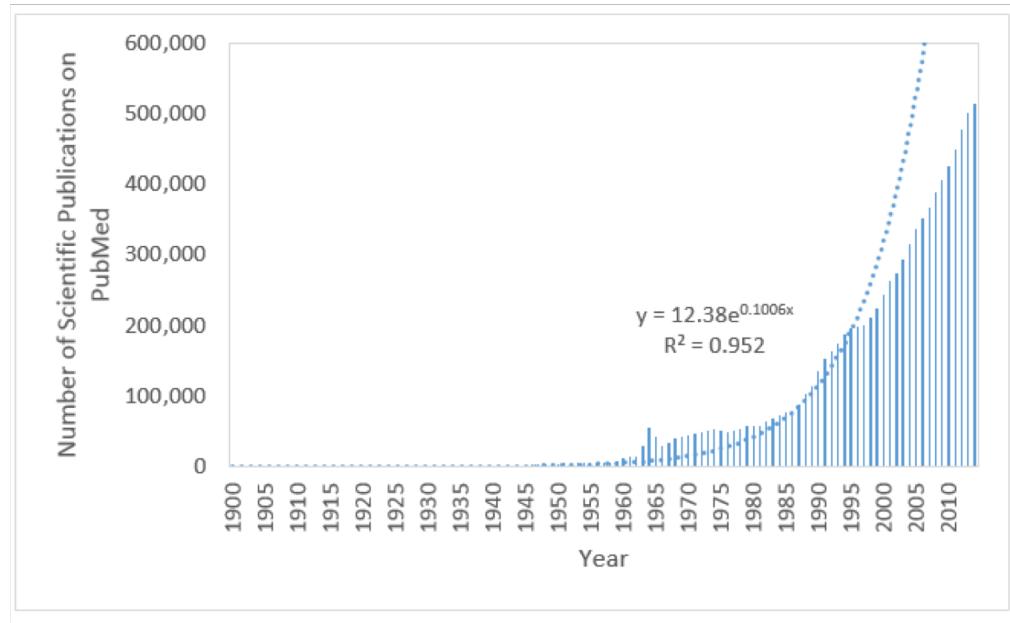
BY **ANTHONY CUTHBERTSON** ON 1/15/18 AT 8:00 AM



# Main (big) motivation

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

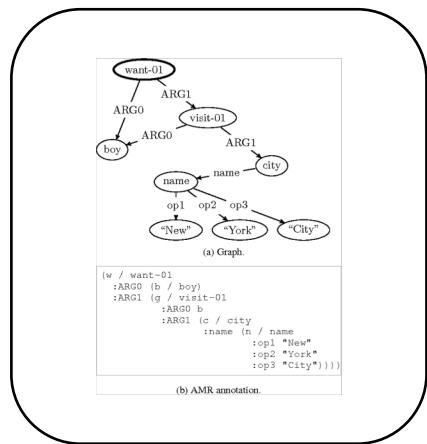
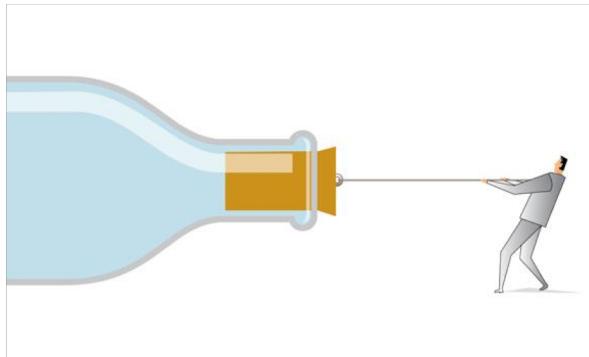
# Motivation 1: Information Overload



uses for

# Motivation 2: The Knowledge Acquisition Bottleneck

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



[Meaning]



uses for

# Applications: Question Answering

In January 1880, two of Tesla's uncles put together enough money to help him leave Gospic 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]

What city did Tesla move to in 1880?

Prague

[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]

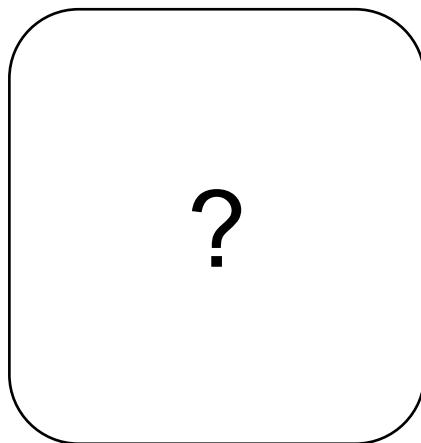


[Information Need]

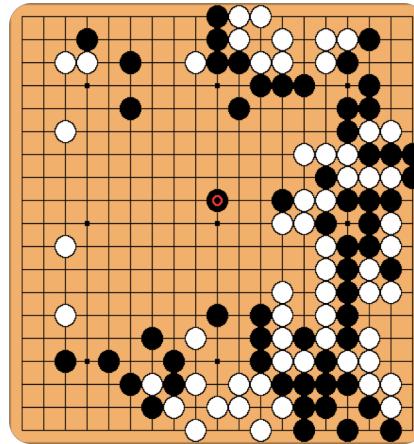
# Applications: Helping Agents to learn Faster

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 Gospic 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]



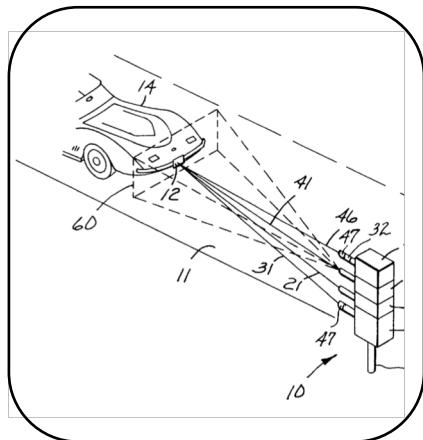
uses for

Q A

[Information Need]

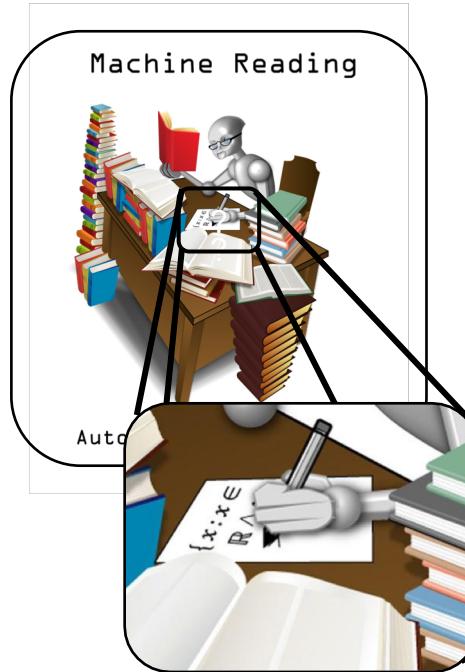
# Timeline of Machine Reading

Something else  
entirely!

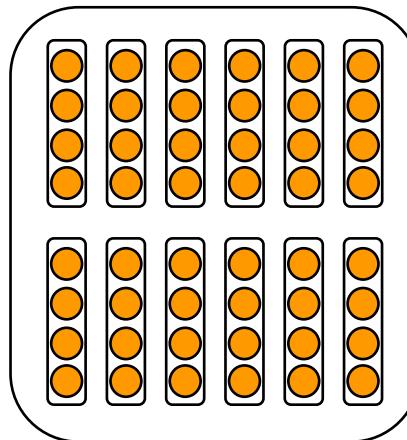


before 2006

Text to Meaning  
Representations

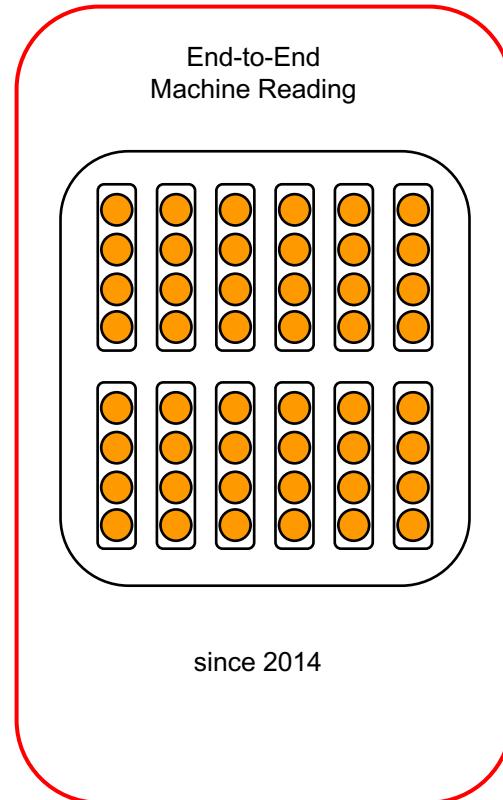


End-to-End  
Machine Reading



since 2014

# Today we cover:



since 2014

# Machine Reading

In January 1880, two of Tesla's uncles put together enough money to help him leave Gospic 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

Q A

[Information Need]

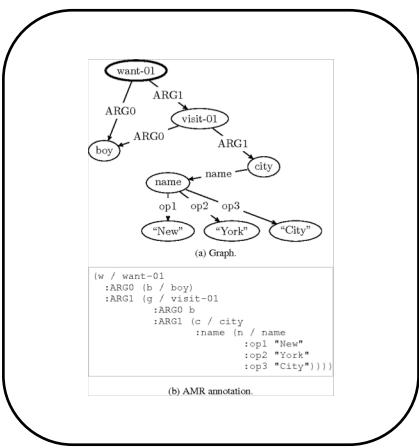
# Symbolic Approaches (until 2014 or so)

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]

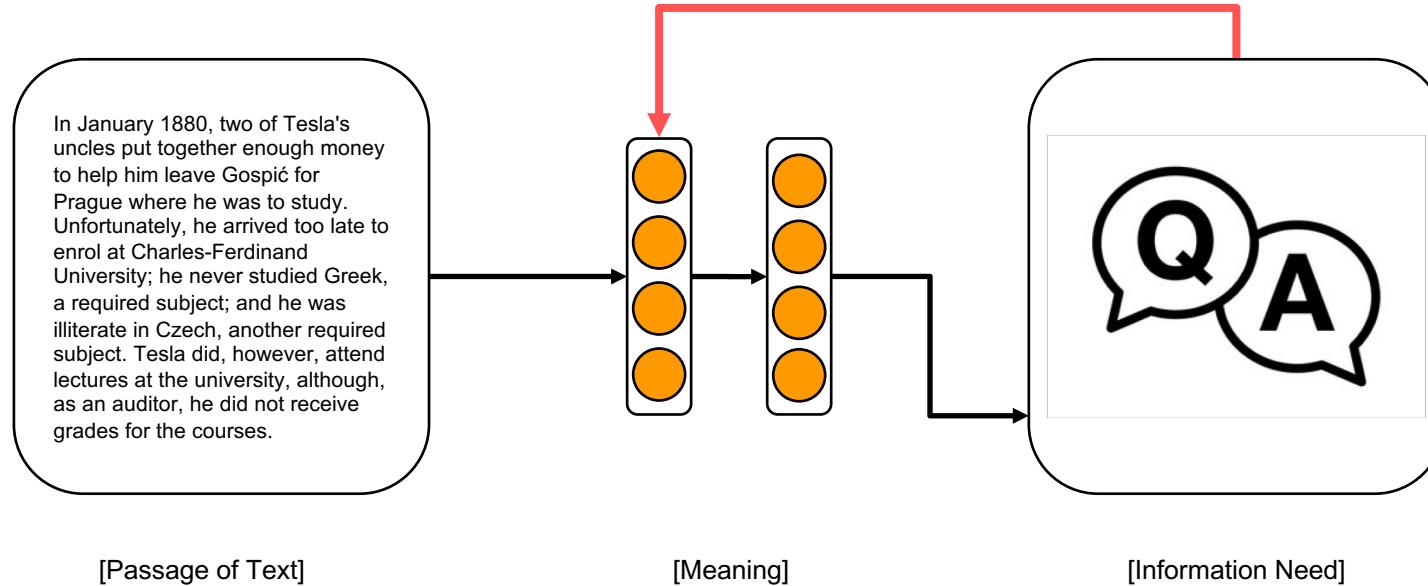


[Information Need]

uses for



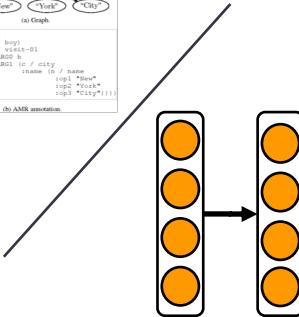
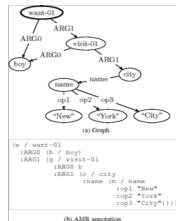
# End-to-End Approaches (since 2014 or so)



# 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]



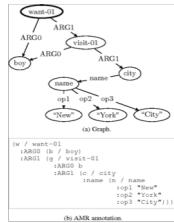
[Meaning]

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

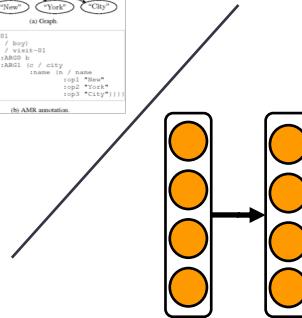
# What are the core challenges?

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]



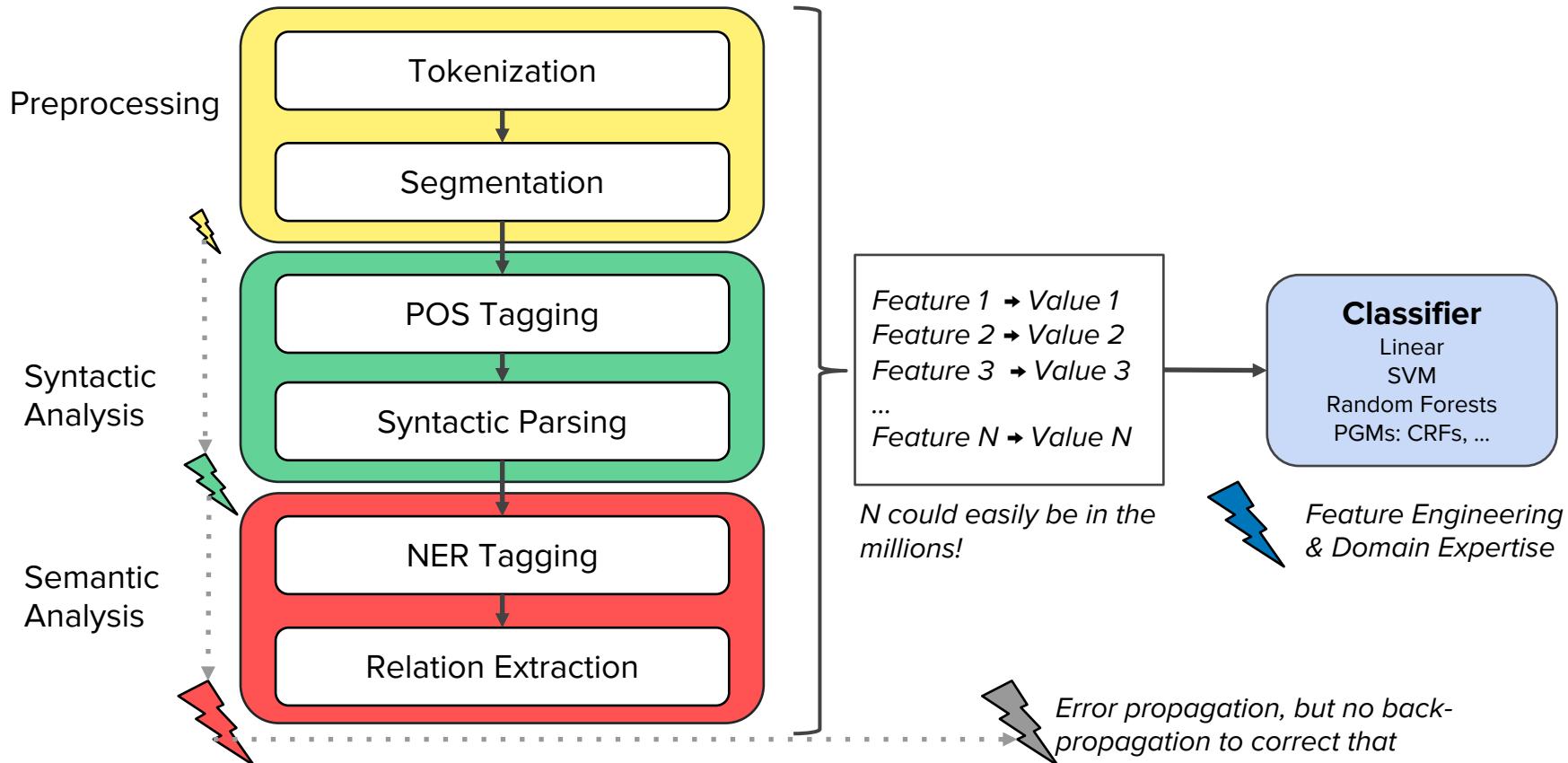
## [Meaning]



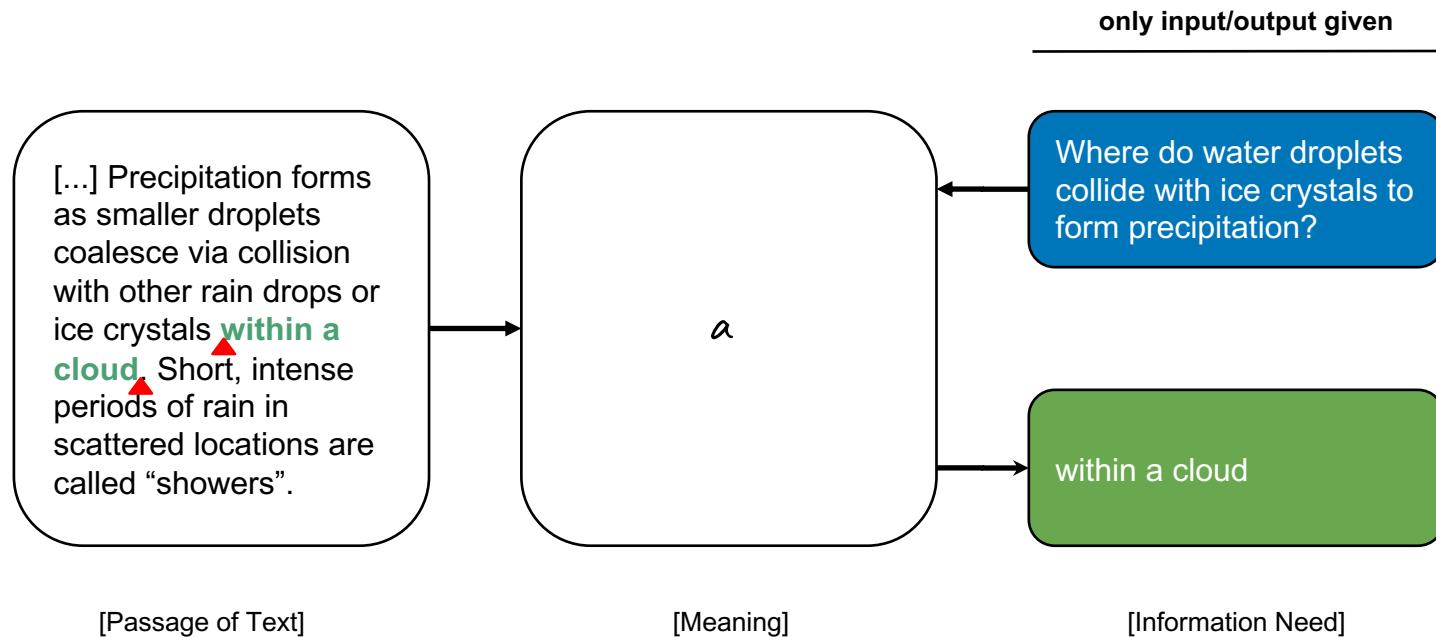
- Ambiguity
  - Variation
  - Coreference
  - Common Sense
  - Scale
  - ...

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

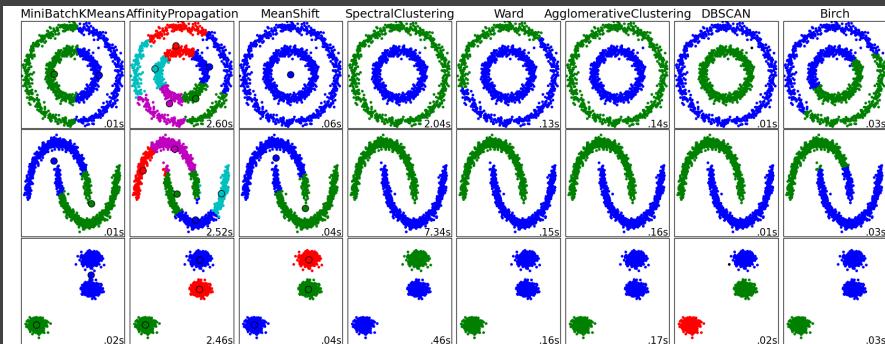
---

# Limits of Big Model + Big Data

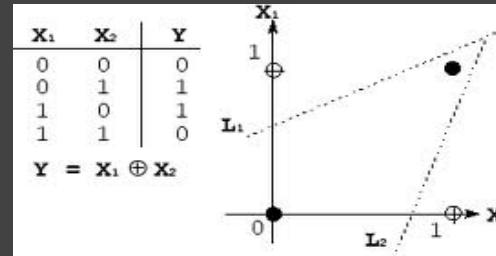
- End-to-end systems need a lot of cleaned data!
- Getting real conditions large-scale data is difficult:
  - Real large-scale data is complex, noisy, **unlabeled...**
  - Interpretation of success or failure is complex
- Complicates the design of innovative learning systems
- Some recent process with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

# A long history of artificial problems in ML

## Two moons and friends (clustering)



## XOR (neural networks)



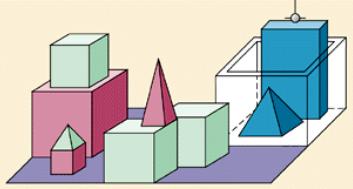
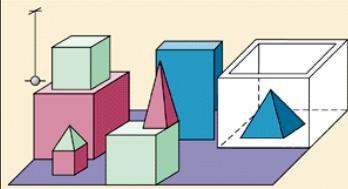
Many in the UCI repository  
(regression, classification)

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



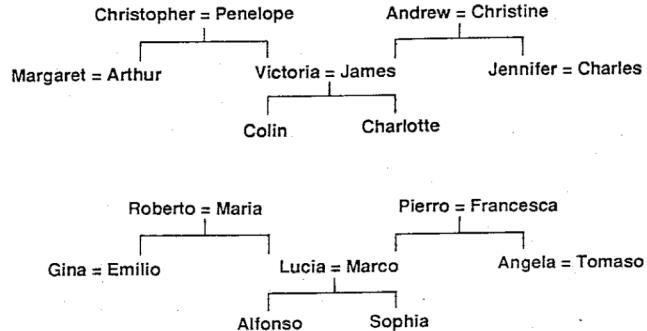
# And in early AI too

## 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.*

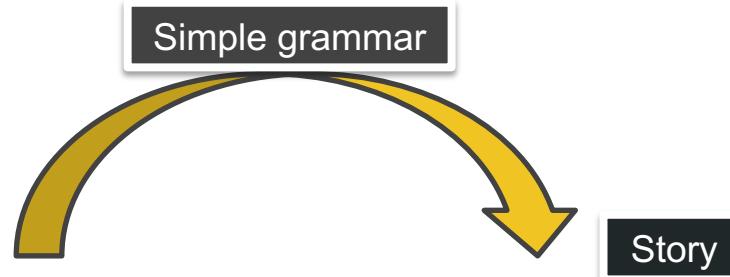
# Simulation commands

- 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

Jason went to the kitchen.

Jason picked up the milk.

Jason travelled to the office.

Jason left the milk there.

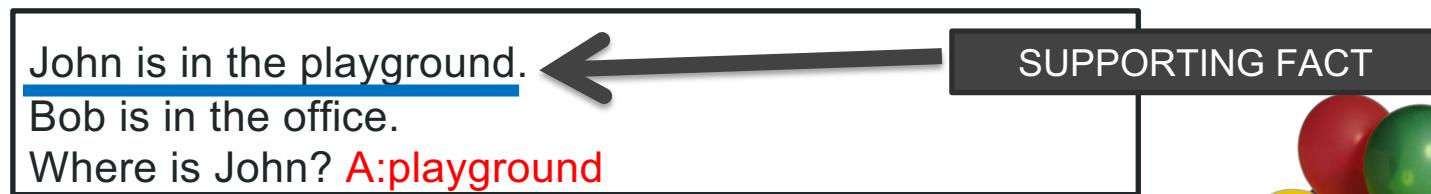
Jason went to the bathroom.

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.



## (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 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

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		

T11. Basic coreference	Source: <a href="https://paperswithcode.com/sota/question-answering-babi">https://paperswithcode.com/sota/question-answering-babi</a>			PASS
T12. Conjunction	5	14	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

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 with Bing queries	crowdsourced	3,047	Yang et al. (2015)
<b>bAbI</b>	<b>20 complex reasoning tasks with controlled language</b>	<b>Automatically Generated</b>	<b>20,000</b>	<b>Weston et al. (2015)</b>
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)
<b>SQuAD (1.0 + 2.0)</b>	<b>Extractive QA</b>	<b>Crowdsourced</b>	<b>107,702</b>	<b>Rajpurkar et al (2016), Rajpurkar and Jia et al (2018)</b>
ComplexQuestions	NL question + KB	Search API & Human verification	2,100	Bao et al. (2016)
MovieQA	Multiple choice QA, text & video.	crowdsourced	14,944	Tapaswi 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	Kočiský 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 occurs as smaller drops coalesce via collision with other rain droplets to form ice crystals **within the cloud**. Short, intense periods of rain in scattered locations are called “showers.”

**Task:** Given a text passage, predict the text spans that answer questions.

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 <a href="https://github.com/google-research/bert">https://github.com/google-research/bert</a>	86.673	89.147
2 Mar 05, 2019	BERT + N-Gram Masking + Synthetic Self-Training (single model) Google AI Language <a href="https://github.com/google-research/bert">https://github.com/google-research/bert</a>	85.150	87.715
3 Jan 15, 2019	BERT + MMFT + ADA (ensemble) Microsoft Research Asia	85.082	87.615
4 Jan 10, 2019	BERT + Synthetic Self-Training (ensemble)	84.292	86.967
5 Dec 16, 2018	PAML+BERT (ensemble model) PINGAN GammaLab	83.457	86.122

Very popular leaderboard!

<https://stanford-qa.com>

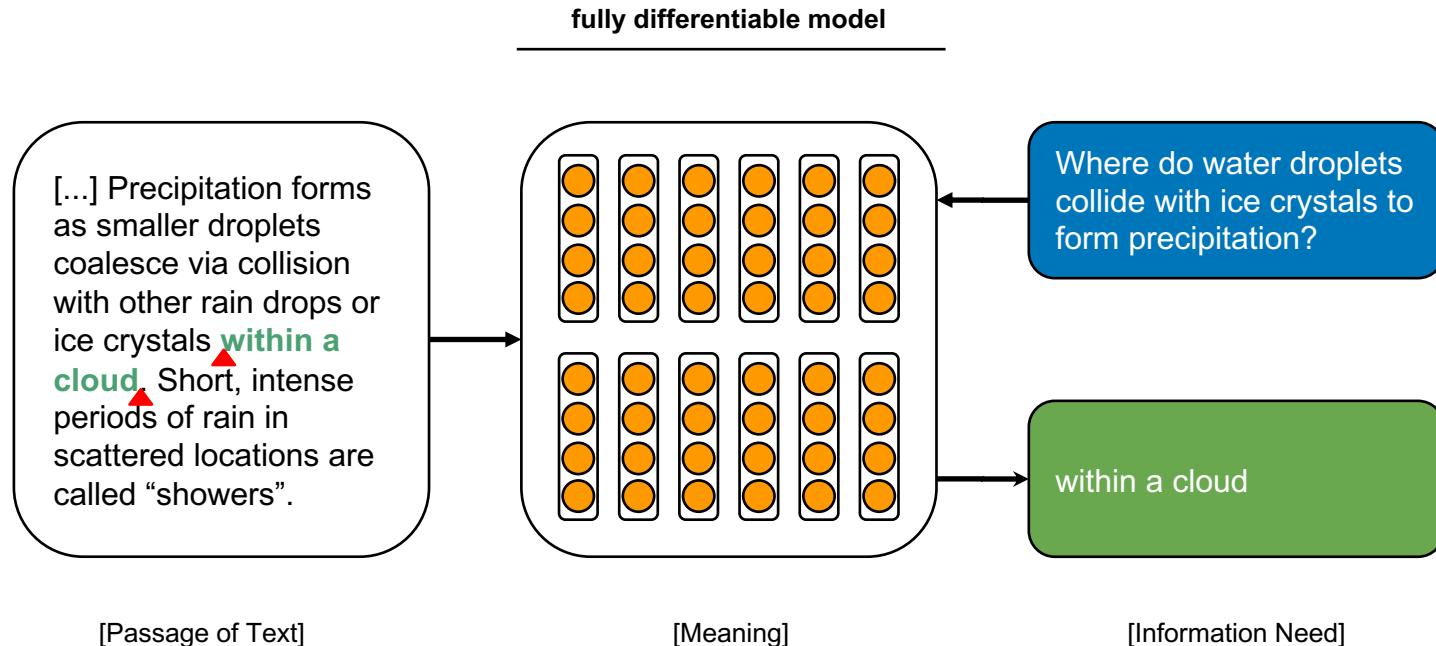
## Answer

Water droplets freeze into ice crystals to form snowflakes?

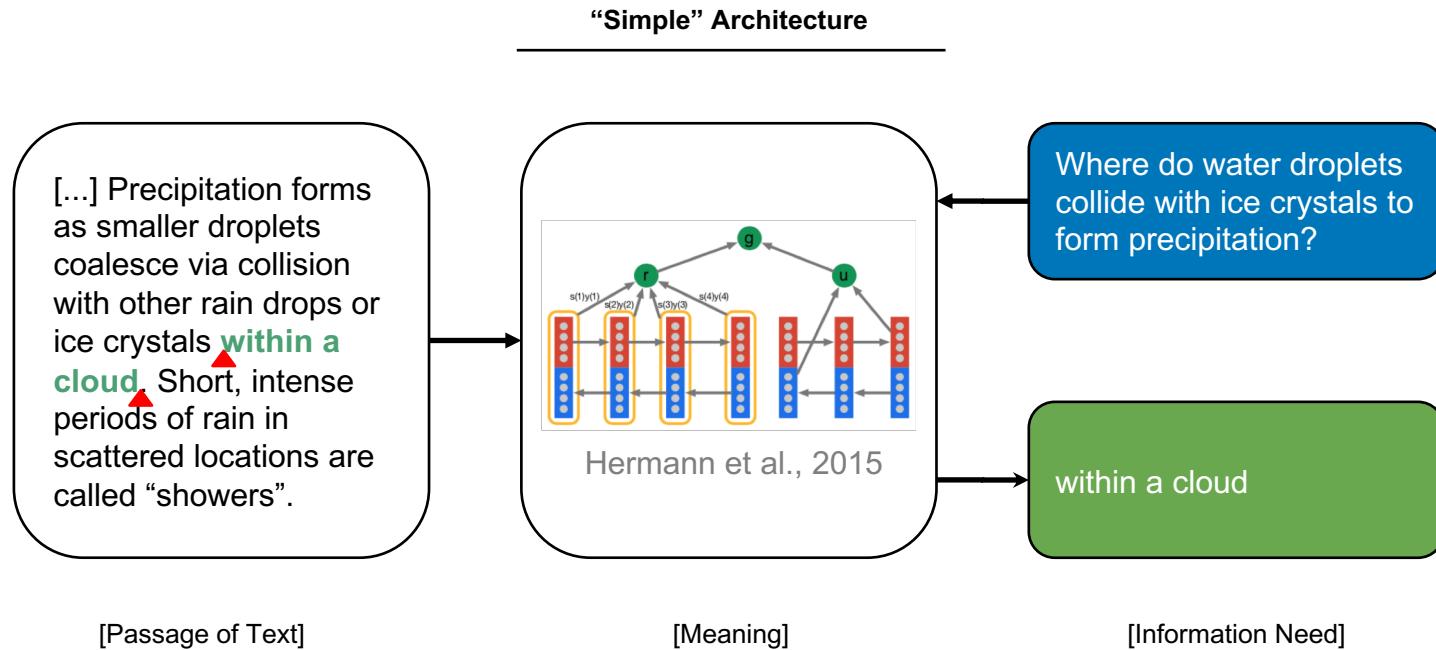
# Machine Reading / Models

---

# End-to-end Machine Reading for Question Answering



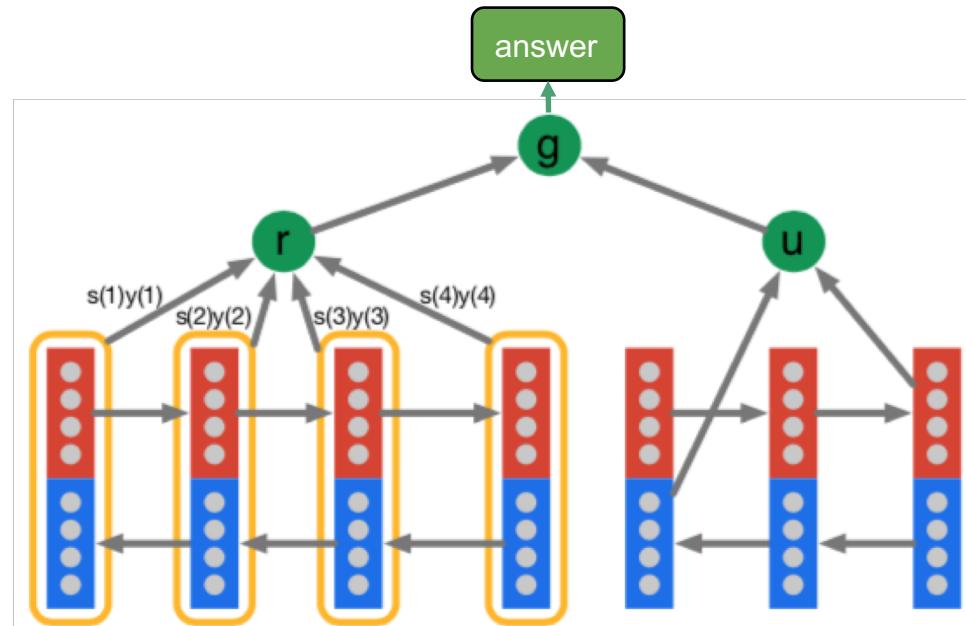
# End-to-end Machine Reading for Question Answering



# The Attentive Reader Model: Overview

Hermann et al., NIPS'15

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



Precipitation

forms

as

smaller

How

do

water

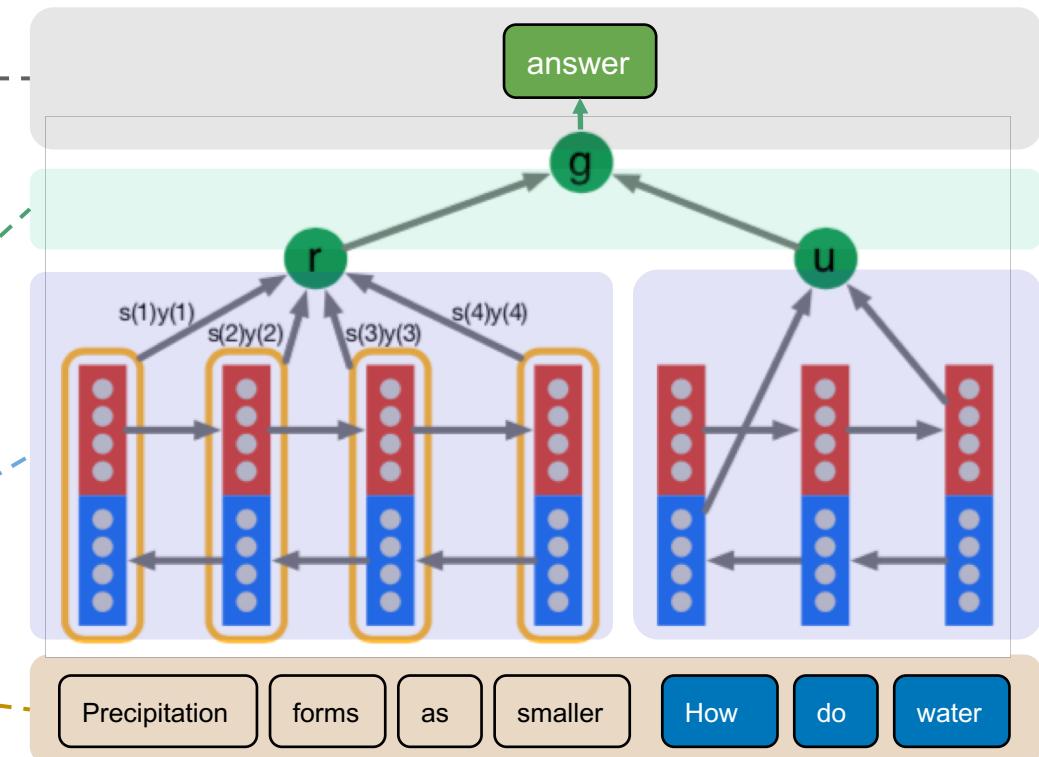
# The Attentive Reader Model: Overview

**Answer Selection:**  
answer prediction

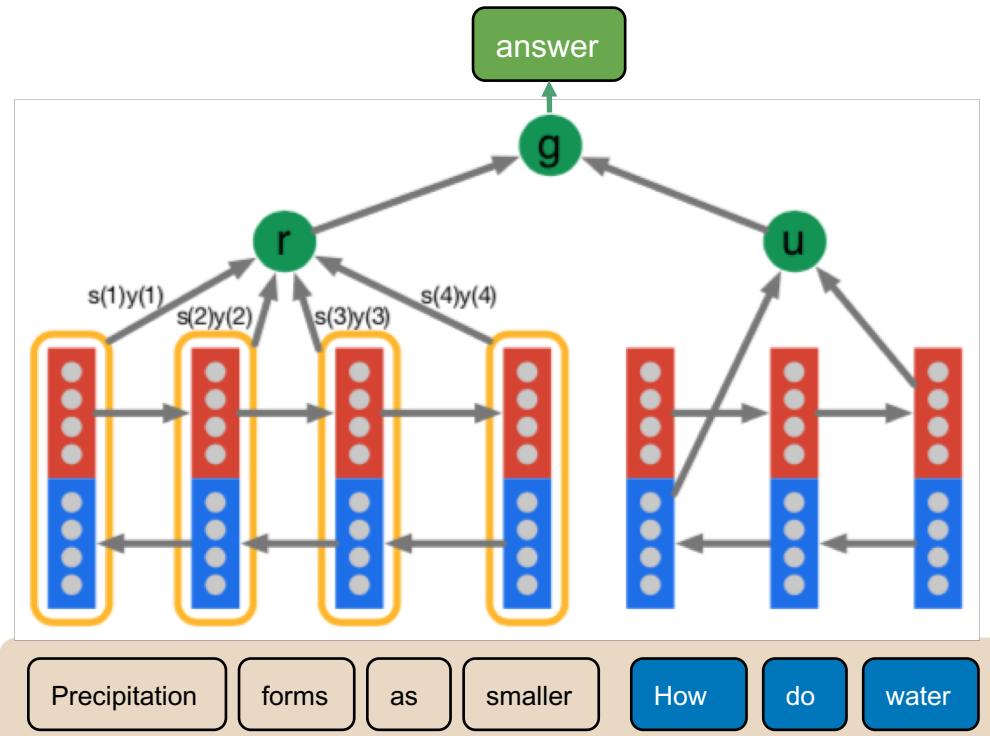
**Sequence Interaction:**  
Matching text with question

**Composition:** incorporating  
context around words

**Input:** Representing symbols as  
vectors

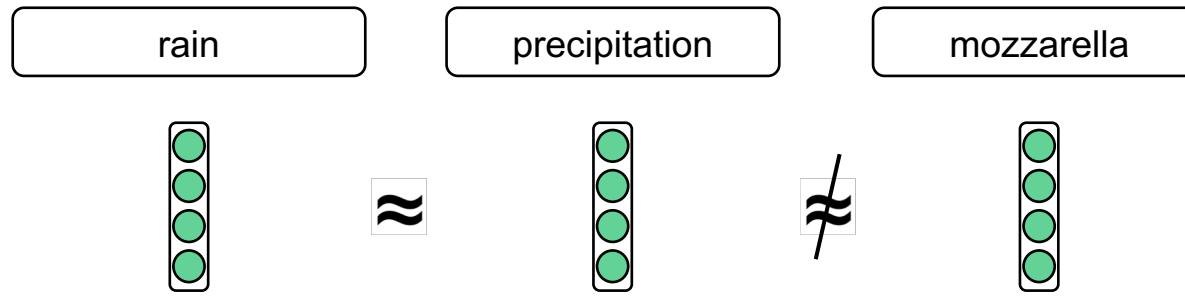


# The Attentive Reader Model: Overview



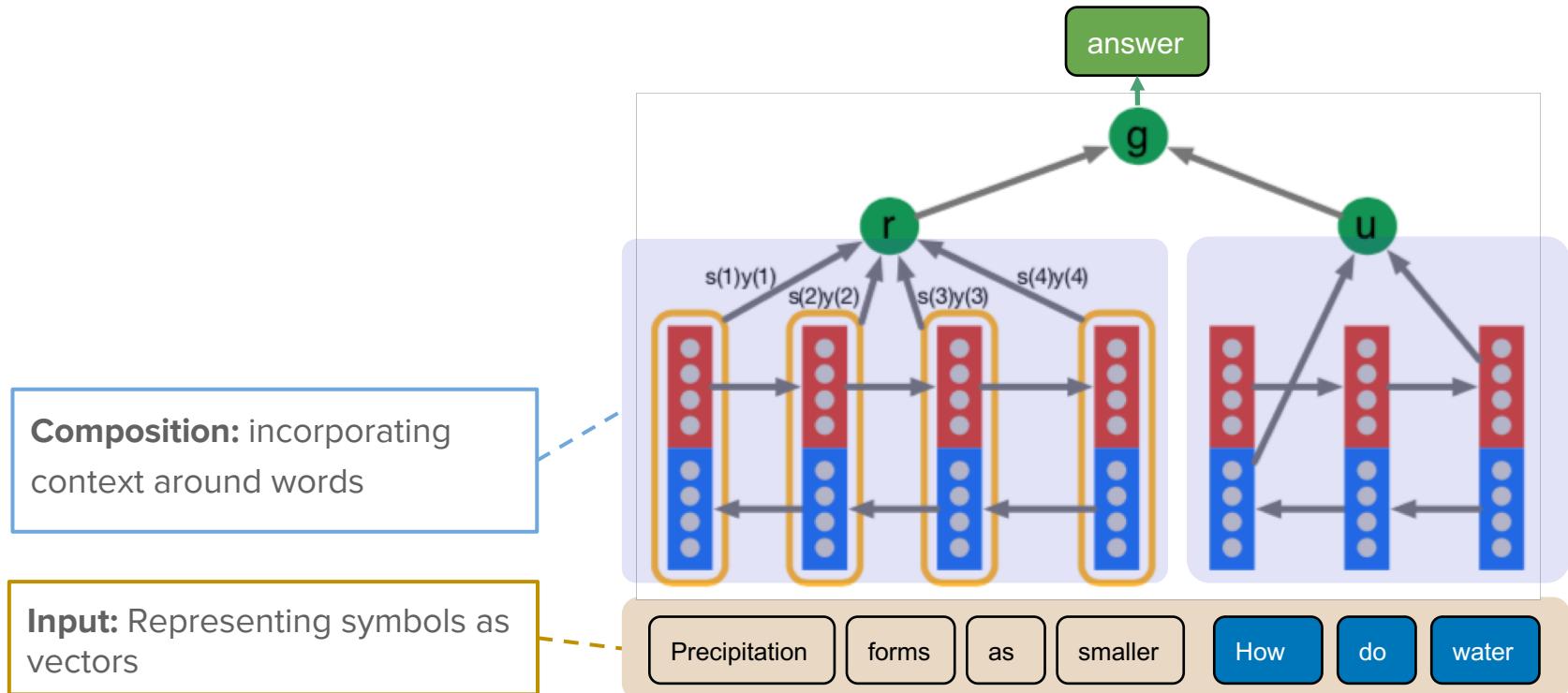
Modified visualization from Hermann et. al. NIPS'15

# Representations for words: Embeddings

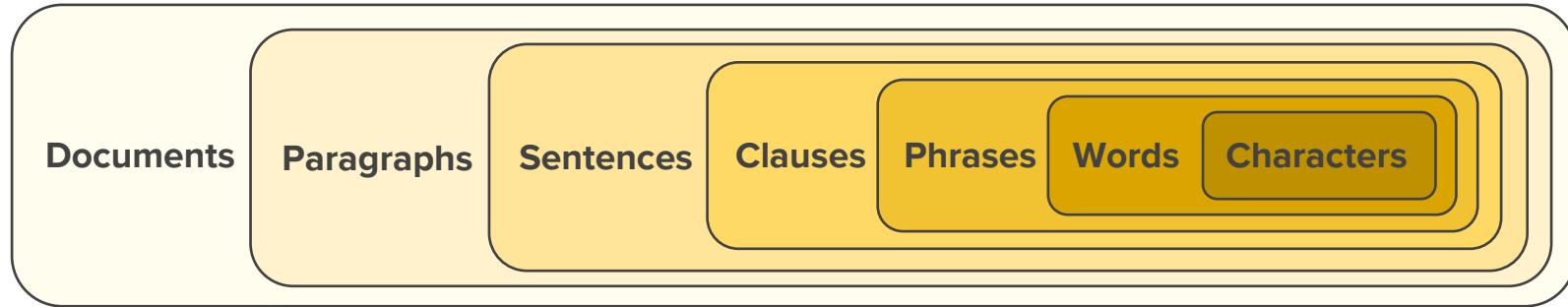


Similar meaning of words → similar vector representations – see previous lectures!

# The Attentive Reader Model: Overview



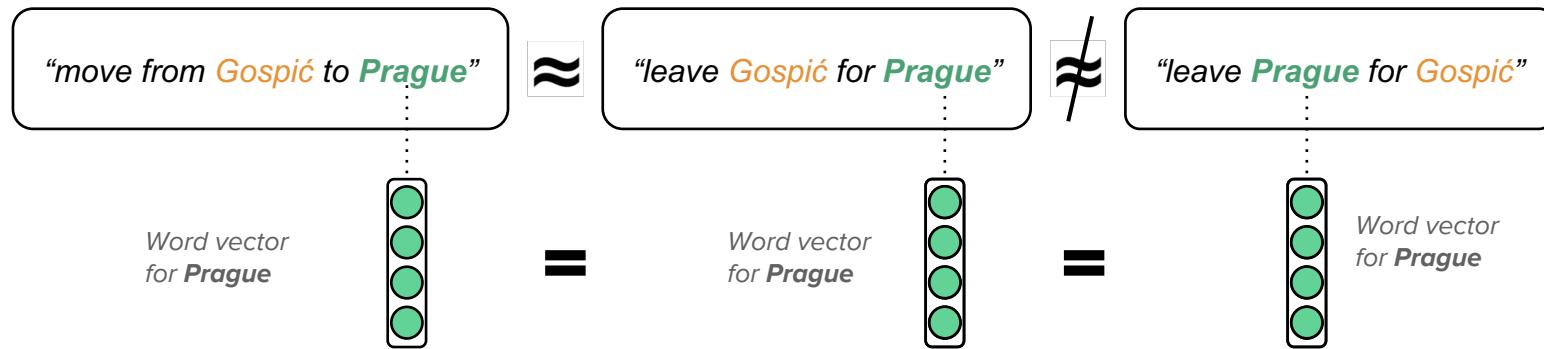
# Language is compositional



## Challenges

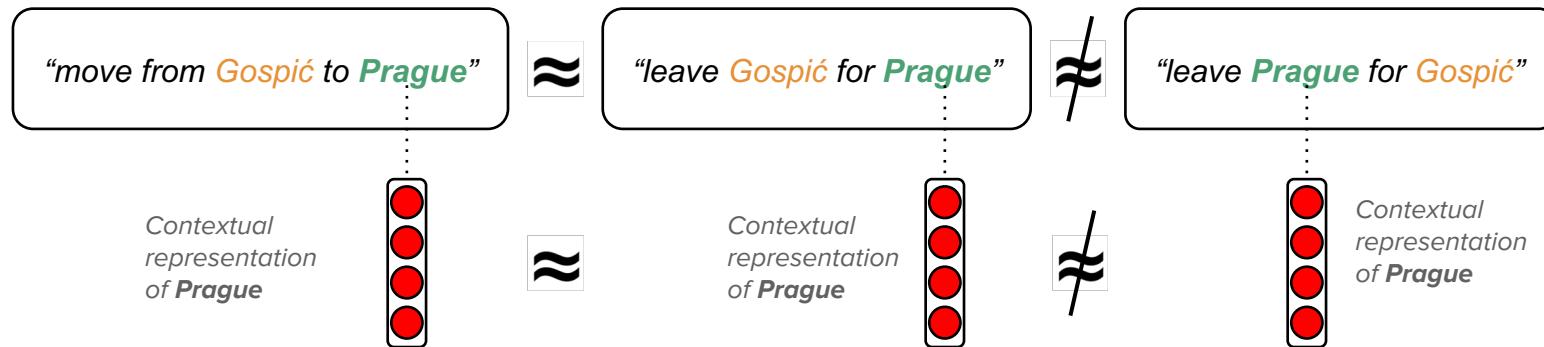
- Inductive bias: which composition function to use?
  - sequence, tree or more general graph structures?
  - varies for different levels
- Capturing long-range dependencies
  - **co-reference** (tracking entities)
  - effective information flow: ease of learning

# Representing Words in Context



- Word representations should vary depending on context

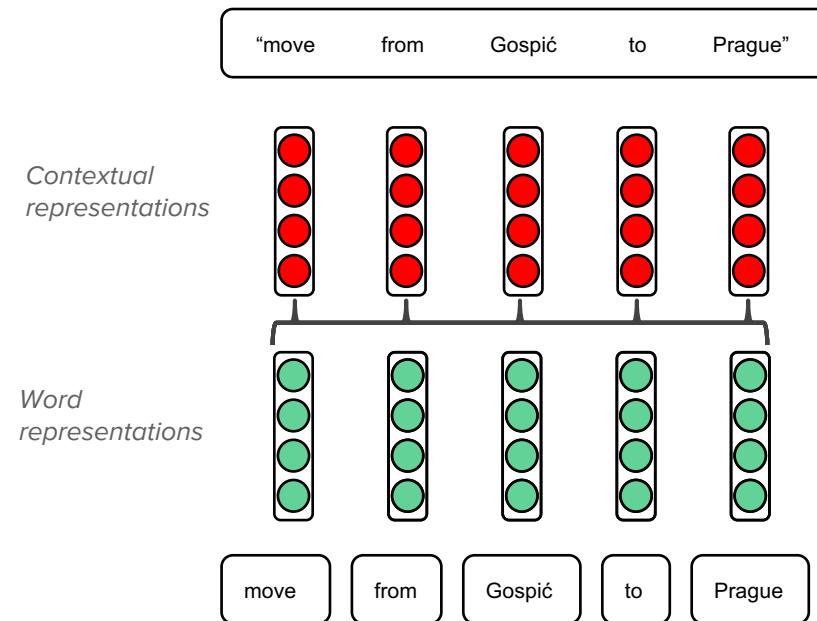
# Representing Words in Context



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

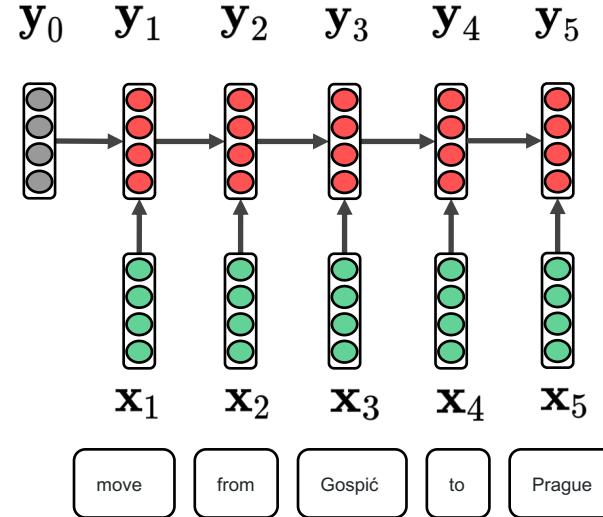
# Representing Words in Context

- 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



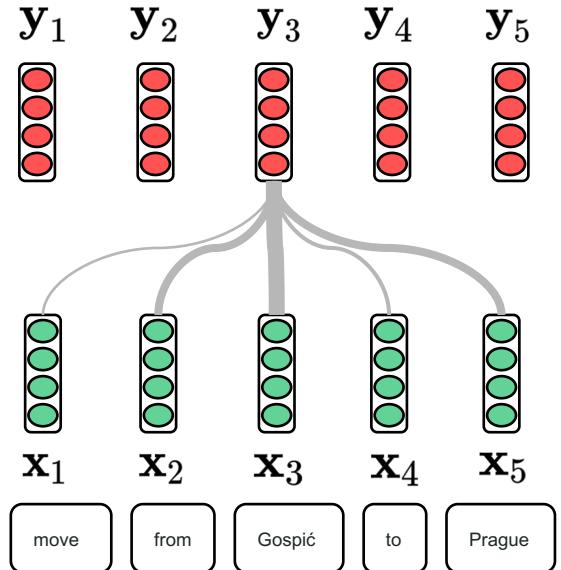
$$\mathbf{y}_t = f(\mathbf{x}_t, \mathbf{y}_{t-1})$$

**Tree-variants:**

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

# Self-Attention Layer

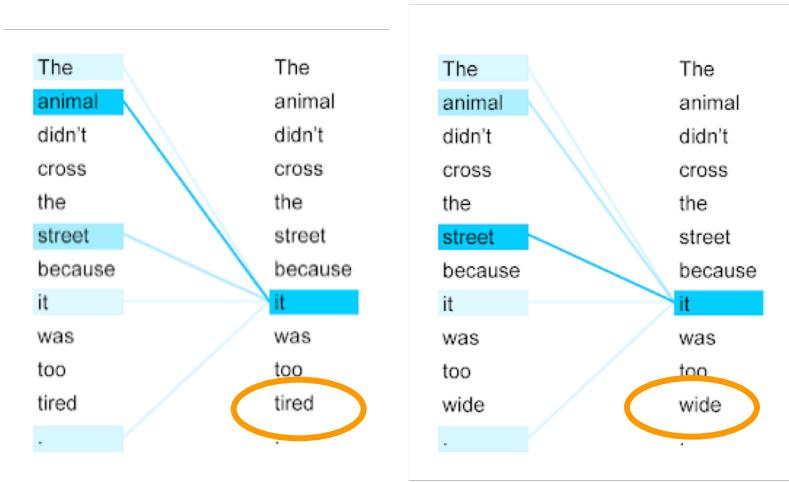
- **Idea:** latent graph on text
- **Inductive bias:**
  - relationships between word pairs
- compute  $K$  separate weighted word representation(s) of the context for each word  $t$
- **Advantages**
  - can capture long-range dependencies
  - Parallelizable and fast
- **Disadvantages**
  - careful setup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



$$\begin{aligned} \mathbf{y}_t &= f(\mathbf{x}_1, \dots, \mathbf{x}_T) \\ \tilde{\mathbf{x}}_t^k &= \sum_{j=1}^T \alpha_{j,t}^k \mathbf{x}_j \quad k = 1, \dots, K \\ f(\mathbf{x}_1, \dots, \mathbf{x}_T) &= \text{nonlinear}(\tilde{\mathbf{x}}_t^1, \dots, \tilde{\mathbf{x}}_t^K) \\ \alpha_t^k &: k^{\text{th}} \text{ self-attention weights for token } t \end{aligned}$$

# 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

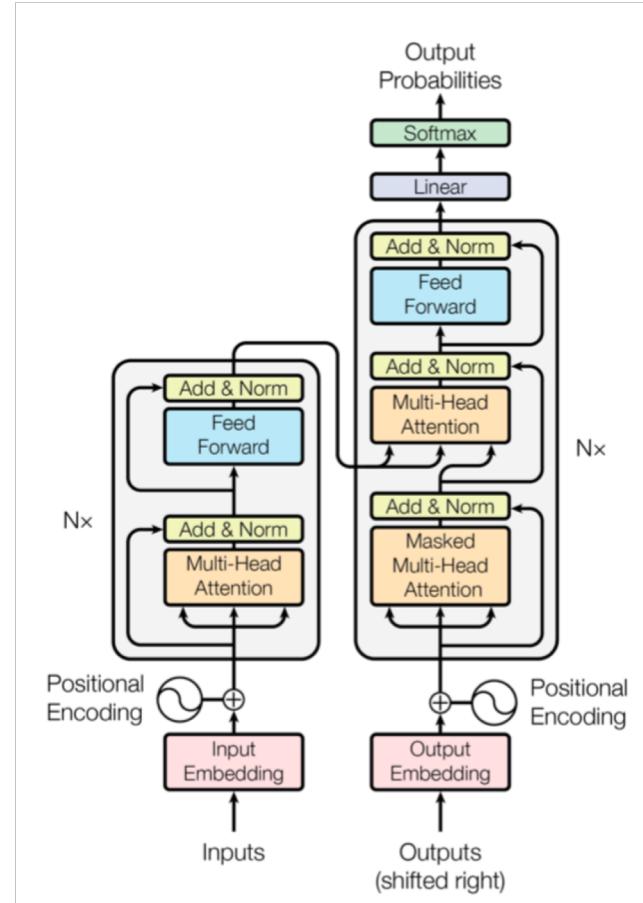
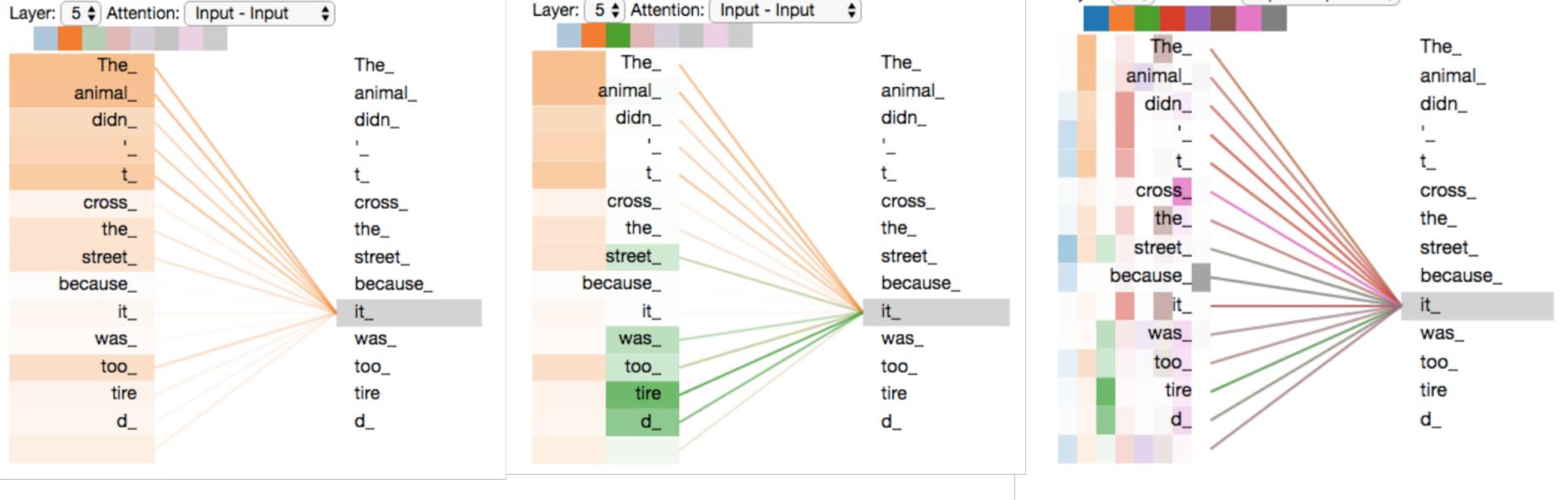


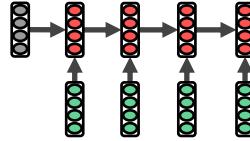
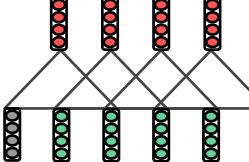
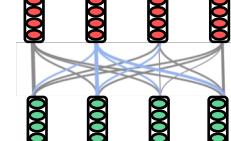
Figure from Vaswani et al., NIPS'17

## Multi-head attention

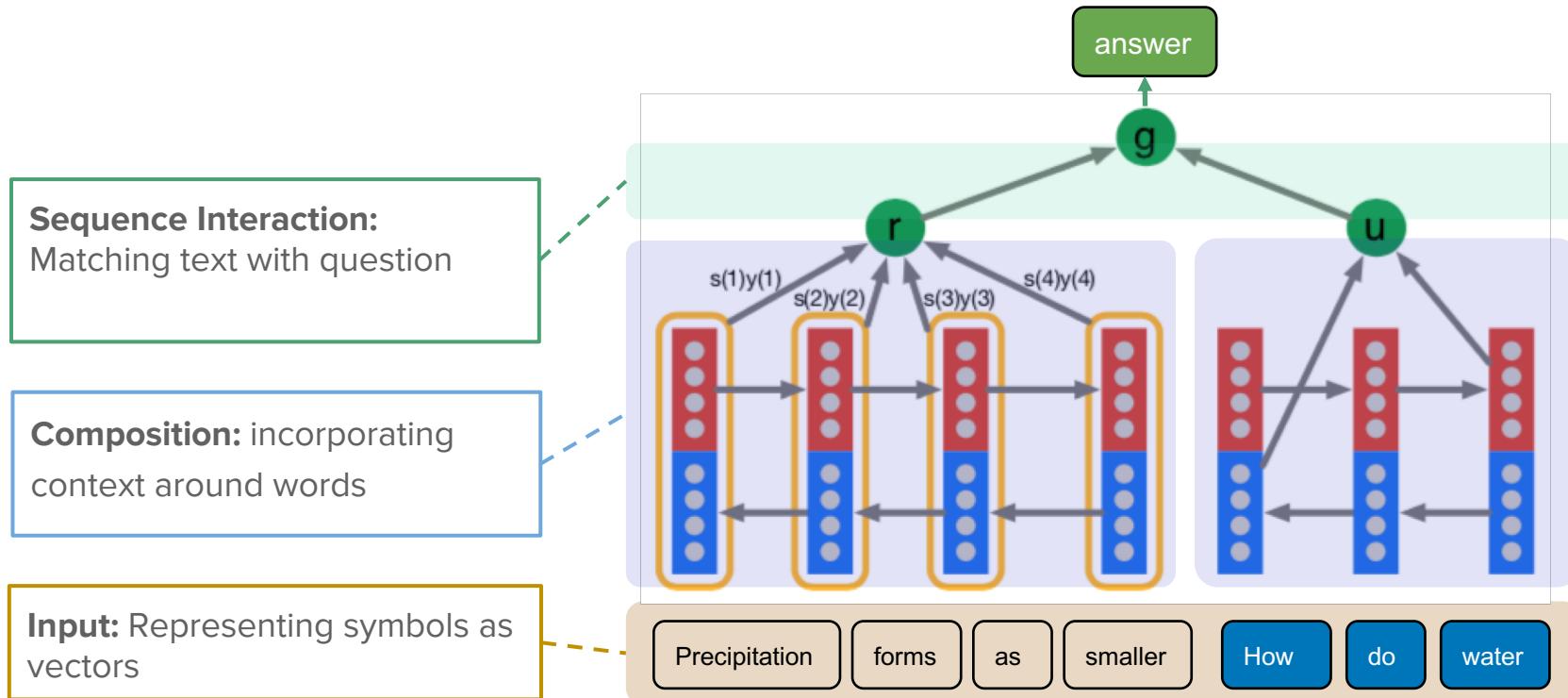


# Compositional Sequence Encoders - Overview

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

Architecture	RNN (LSTM, GRU)	CNN	Self-Attention
Illustration			
Function $\mathbf{y}_t =$	$f(\mathbf{x}_t, \mathbf{y}_{t-1})$	$f(\mathbf{x}_{t-k}, \dots, \mathbf{x}_{t+k})$	$f(\mathbf{x}_1, \dots, \mathbf{x}_T)$
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



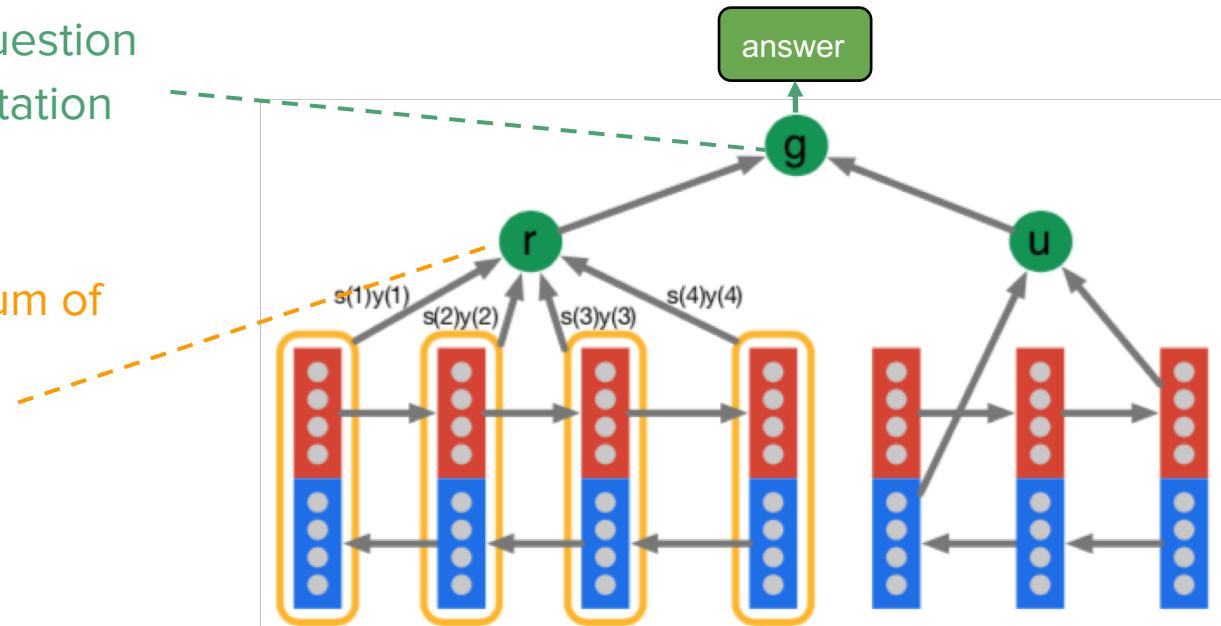
# Modelling sequence interactions

- **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



Precipitation

forms

as

smaller

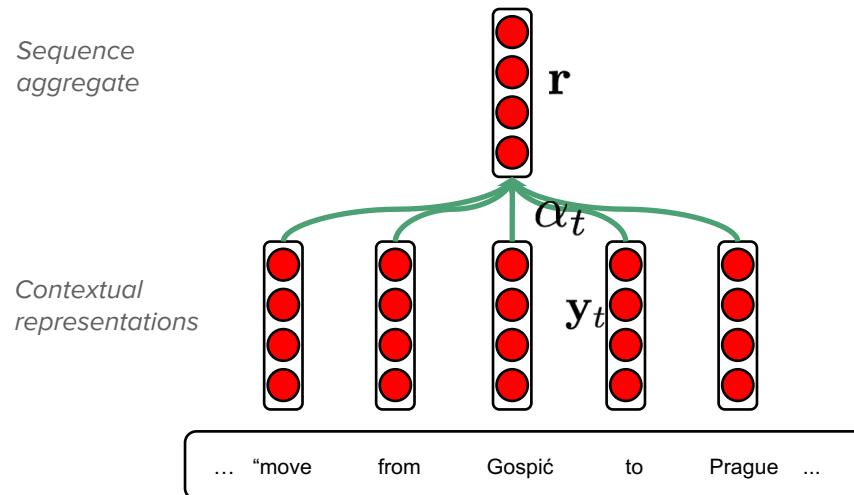
How

do

water

# 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

# 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**

# Example: Learned attention patterns

by ent423 ,ent261 correspondent updated 9:49 pm et ,thu march 19 ,2015 ( ent261 ) a ent114 was killed in a parachute accident in ent45 ,ent85 ,near ent312 ,a ent119 official told ent261 on wednesday .he was identified thursday as special warfare operator 3rd class ent23 ,29 ,of ent187 , ent265 .`` ent23 distinguished himself consistently throughout his career .he was the epitome of the quiet professional in all facets of his life ,and he leaves an inspiring legacy of natural tenacity and focused

...

ent119 identifies deceased sailor as X ,who leaves behind a wife

by ent270 ,ent223 updated 9:35 am et ,mon march 2 ,2015 ( ent223 ) ent63 went familial for fall at its fashion show in ent231 on sunday ,dedicating its collection to `` mamma '' with nary a pair of `` mom jeans '' in sight .ent164 and ent21 , who are behind the ent196 brand ,sent models down the runway in decidedly feminine dresses and skirts adorned with roses ,lace and even embroidered doodles by the designers ' own nieces and nephews .many of the looks featured saccharine needlework phrases like `` i love you ,

...

X dedicated their fall fashion show to moms

Attentive Reader on QACNN/DailyMail dataset

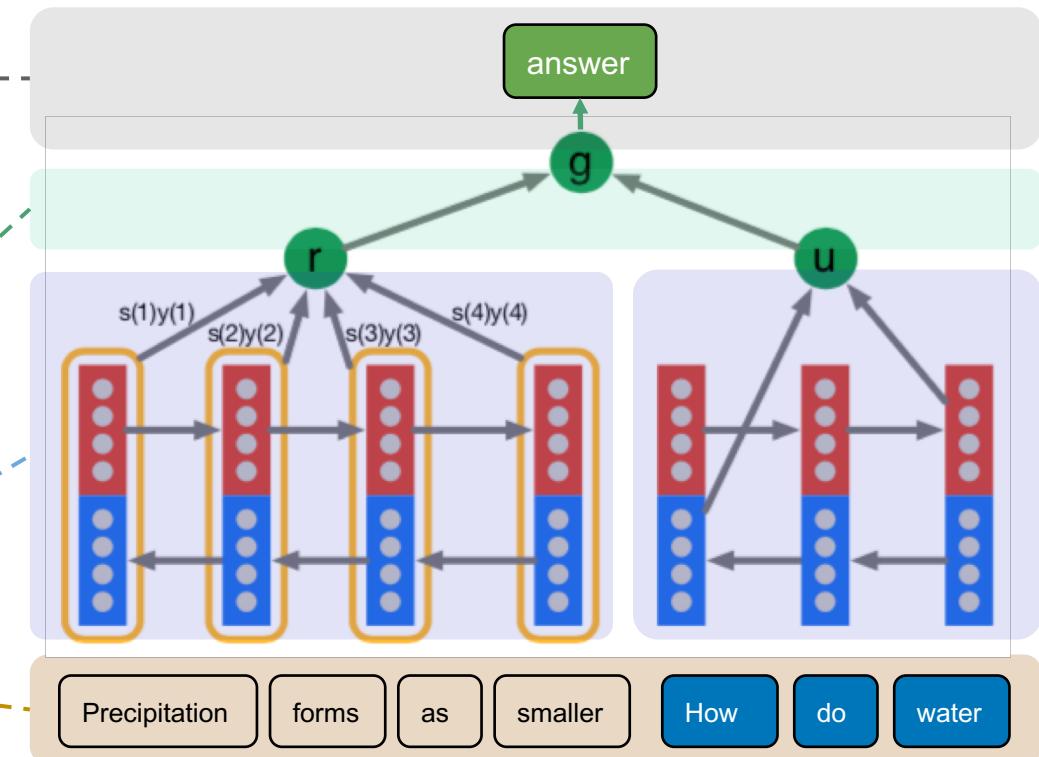
# The Attentive Reader Model: Overview

**Answer Selection:**  
answer prediction

**Sequence Interaction:**  
Matching text with question

**Composition:** incorporating  
context around words

**Input:** Representing symbols as  
vectors

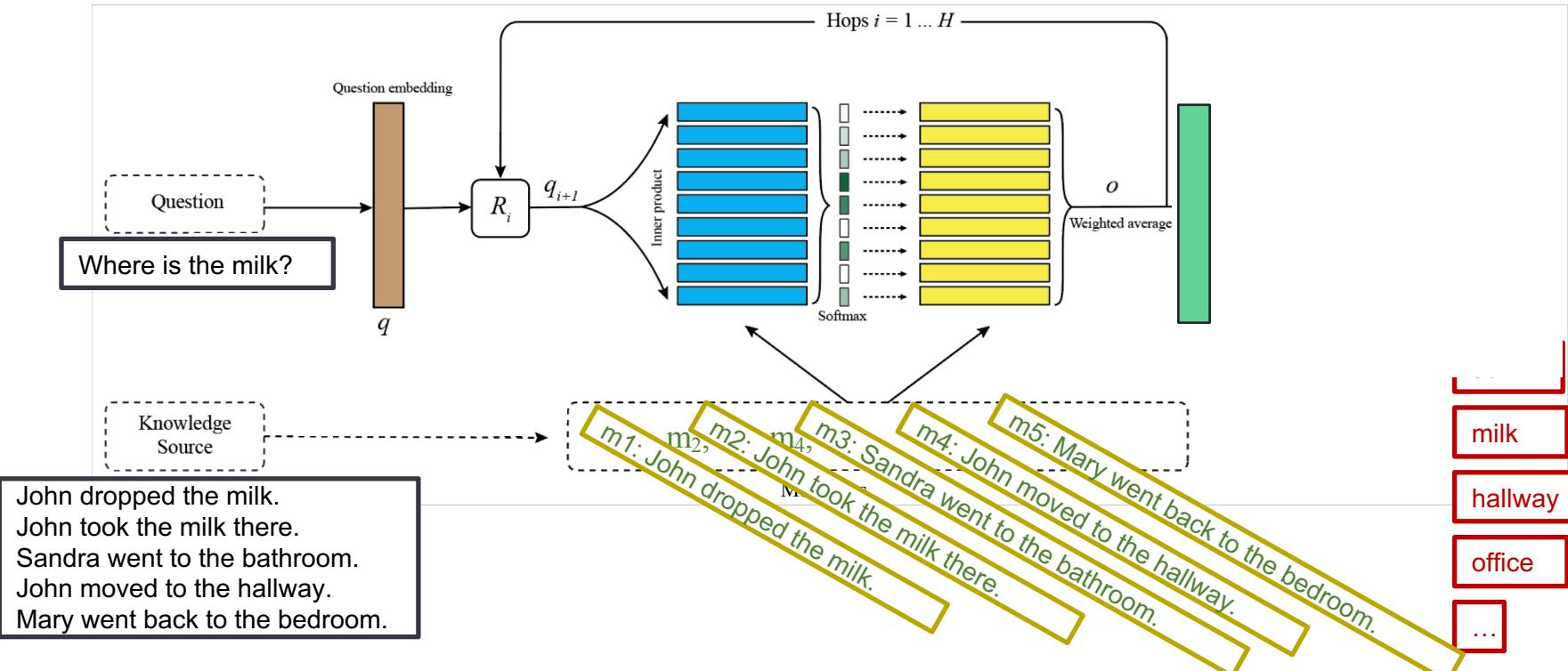


# Answer prediction

- 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



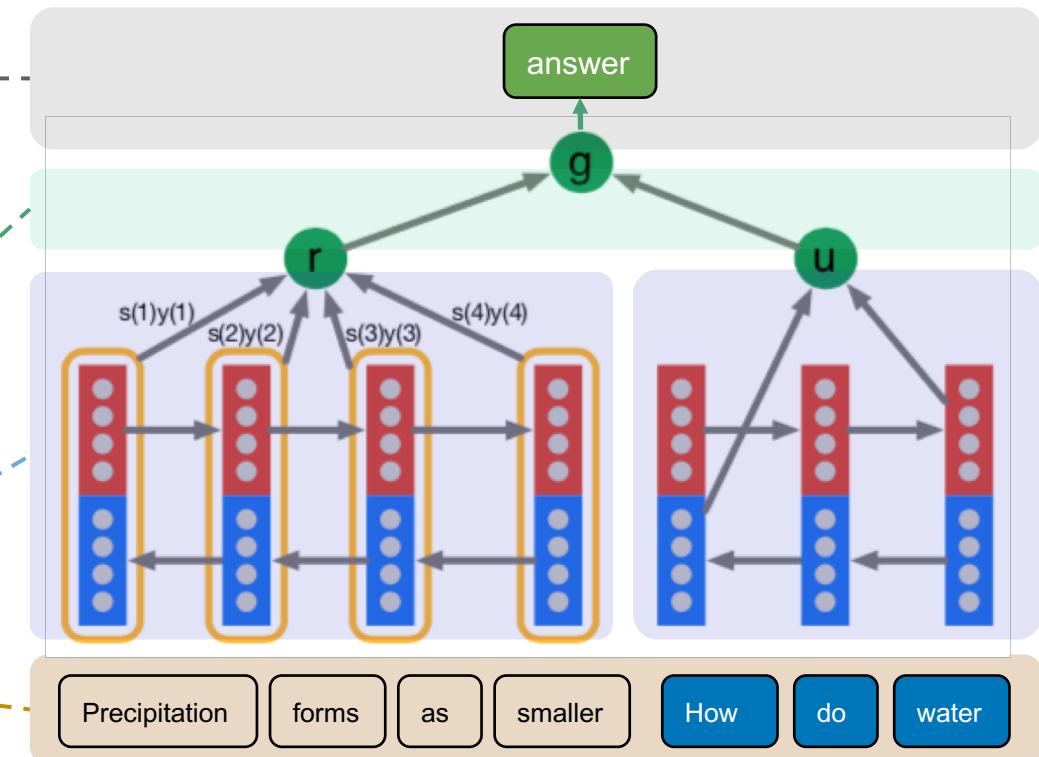
# The Attentive Reader Model: Overview

**Answer Selection:**  
answer prediction

**Sequence Interaction:**  
Matching text with question

**Composition:** incorporating  
context around words

**Input:** Representing symbols as  
vectors



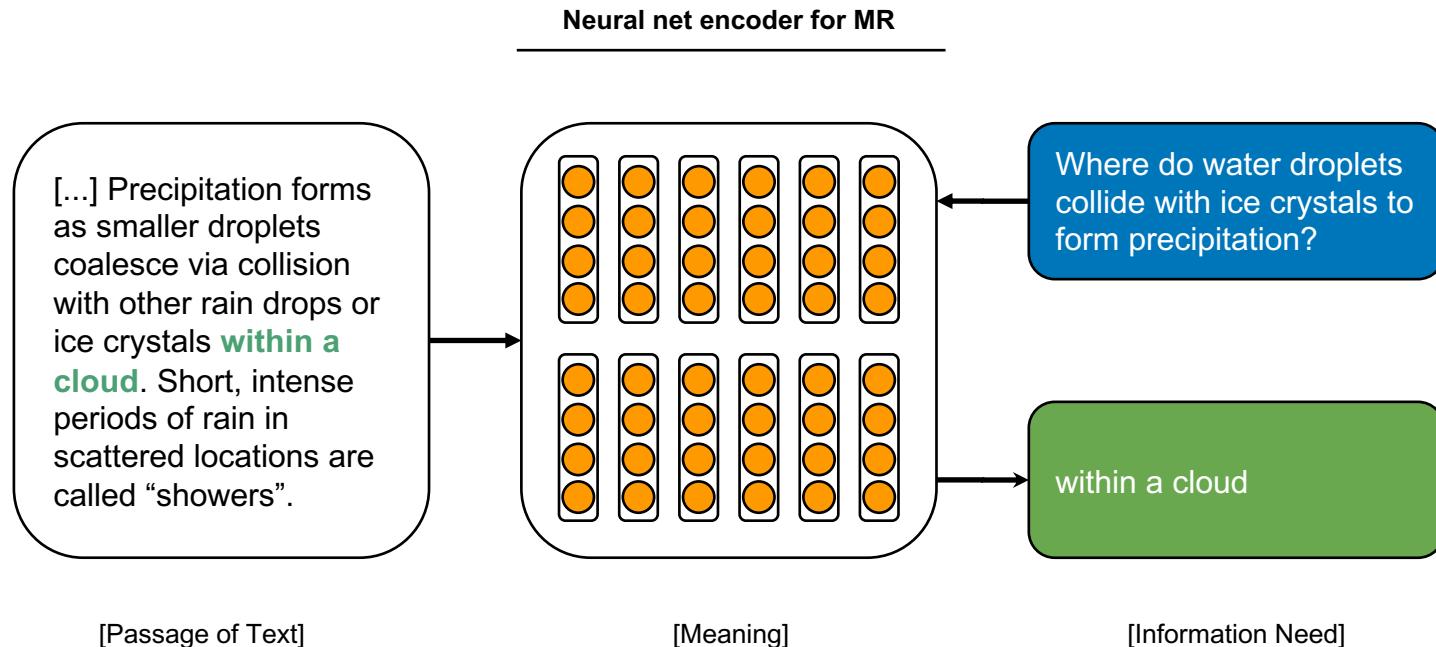
# Conclusion

- 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!

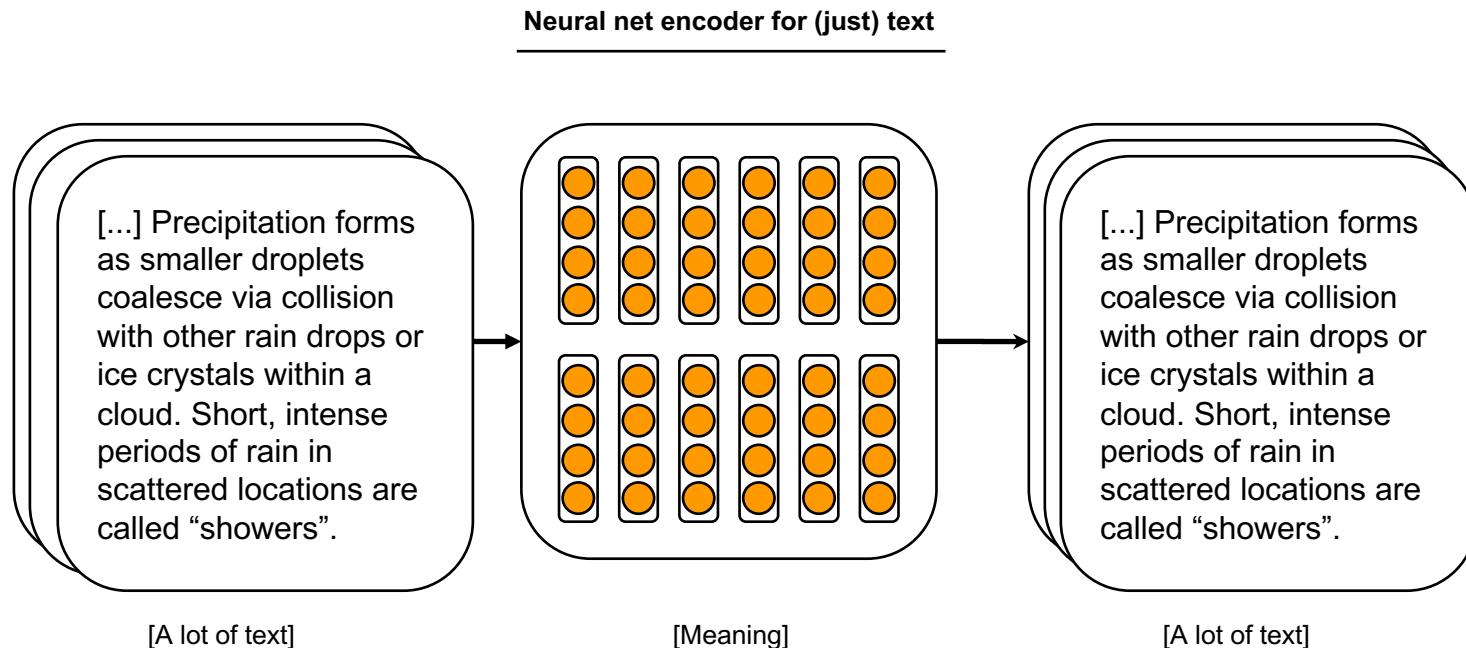
# Machine Reading / Current Trend

---

# Supervised training

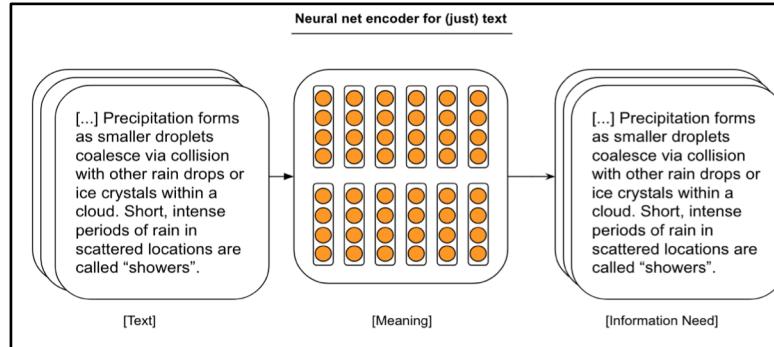


# Unsupervised pretrained representations

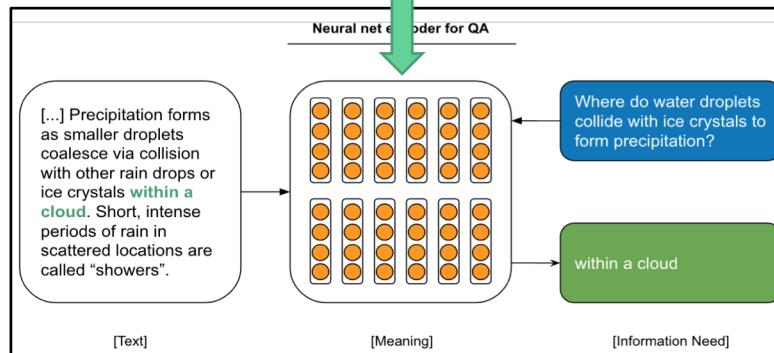


# Lifting over pretrained representations

## Pretrained Language Model



Transfer



## Machine Reading

# How is this different from pretrained word embeddings?

## Pretrained **Word** Embeddings (word2vec)

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

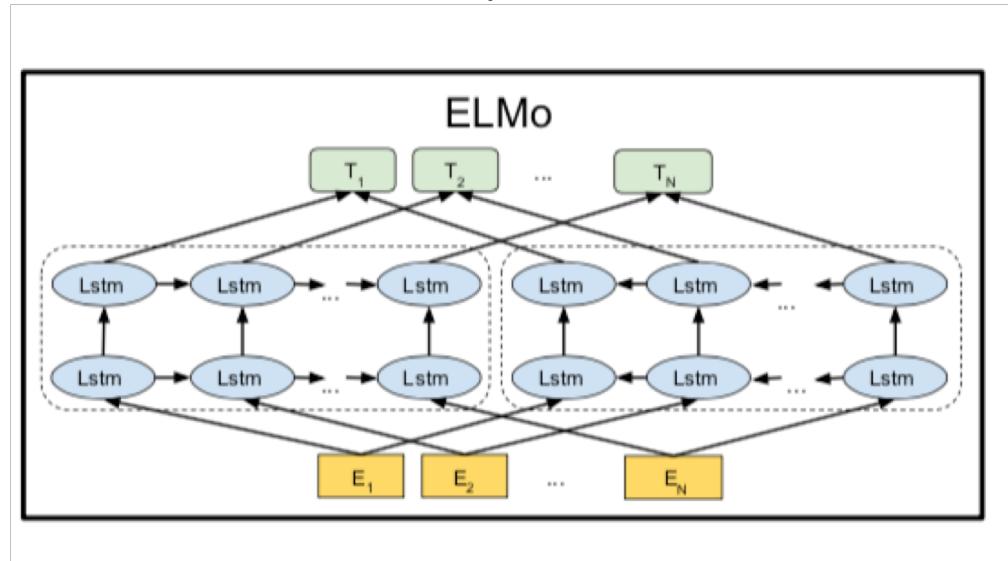
## Pretrained **Contextualized** Embeddings (e.g. ELMo, BERT)

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

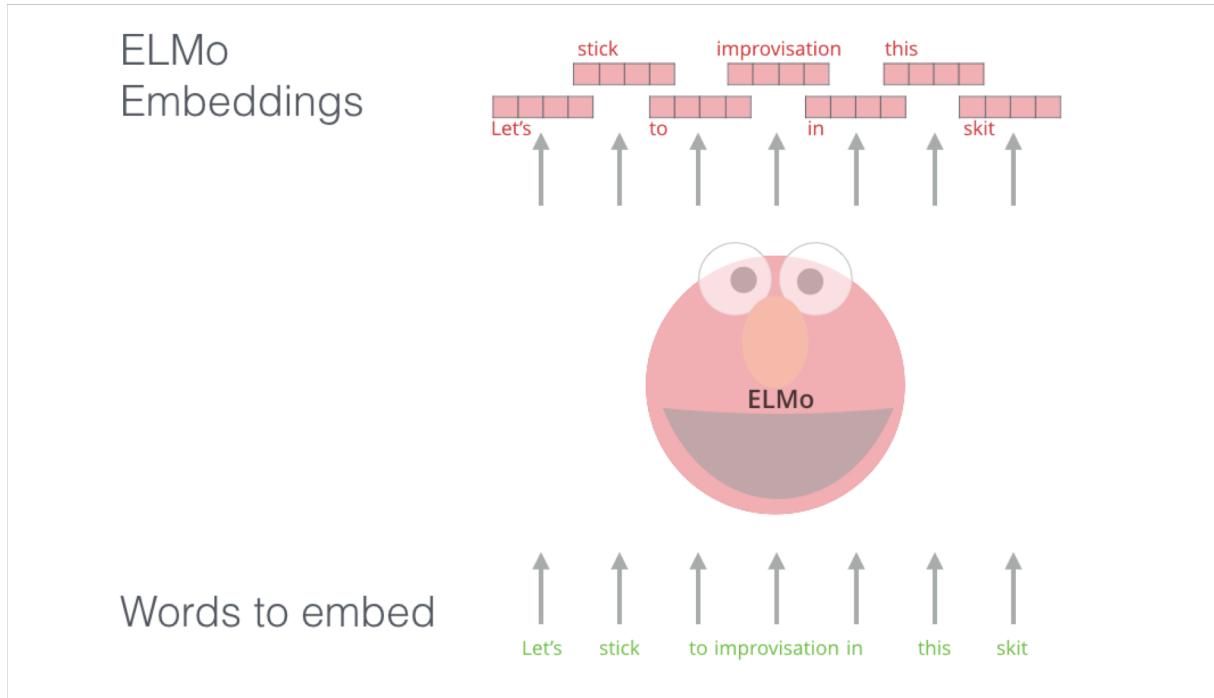
# ELMo: Embeddings from Language Models

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 forward and backward representations into final contextual embeddings

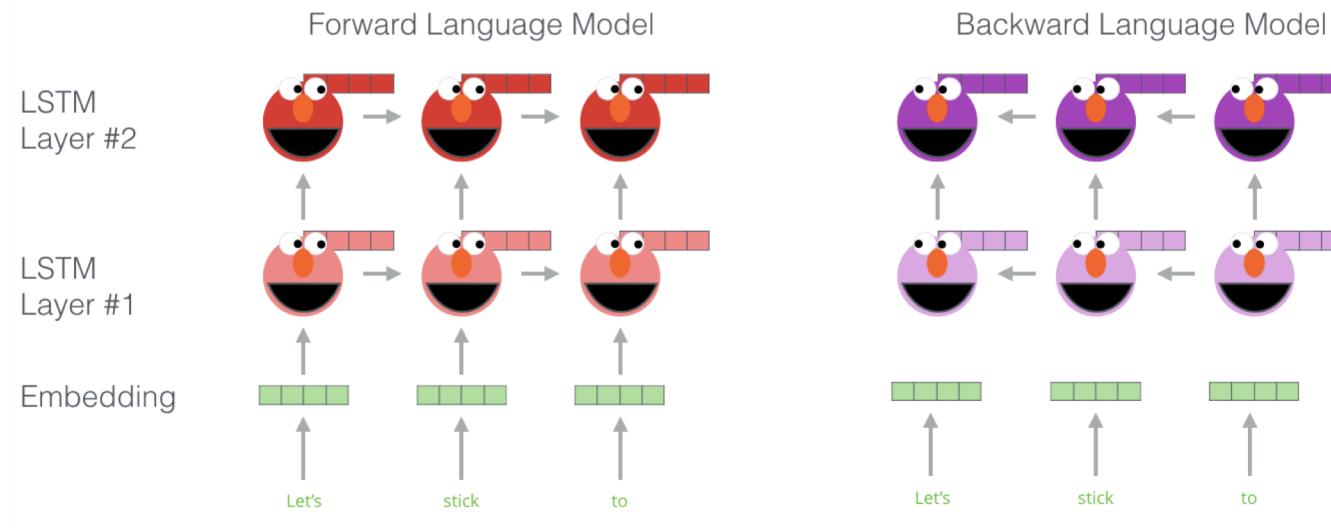


# ELMo: Embeddings from Language Models



# ELMo: Embeddings from Language Models

Embedding of “stick” in “Let’s stick to” - Step #1



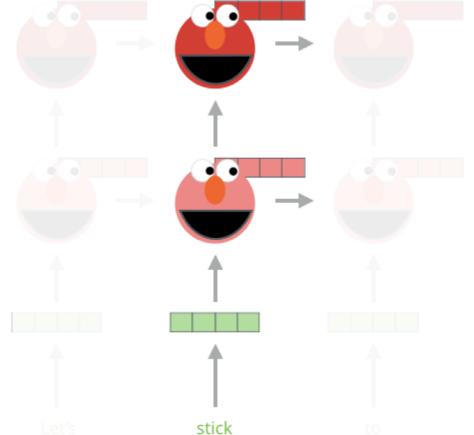
# ELMo: Embeddings from Language Models

Embedding of “stick” in “Let’s stick to” - Step #2

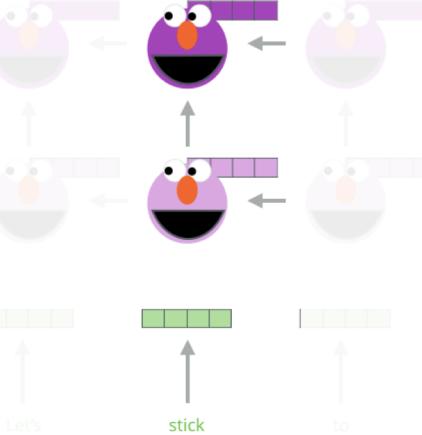
1- Concatenate hidden layers



Forward Language Model



Backward Language Model



2- Multiply each vector by a weight based on the task

$$\text{red vector} \times s_2$$

$$\text{purple vector} \times s_1$$

$$\text{green vector} \times s_0$$

3- Sum the (now weighted) vectors



ELMo embedding of “stick” for this task in this context

# 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 <u>play</u> 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 <u>play</u> .
	Olivia De Havilland signed to do a Broadway <u>play</u> for Garson {...}	{...} they were actors who had been handed fat roles in a successful <u>play</u> , and had talent enough to fill the roles competently , with nice understatement .

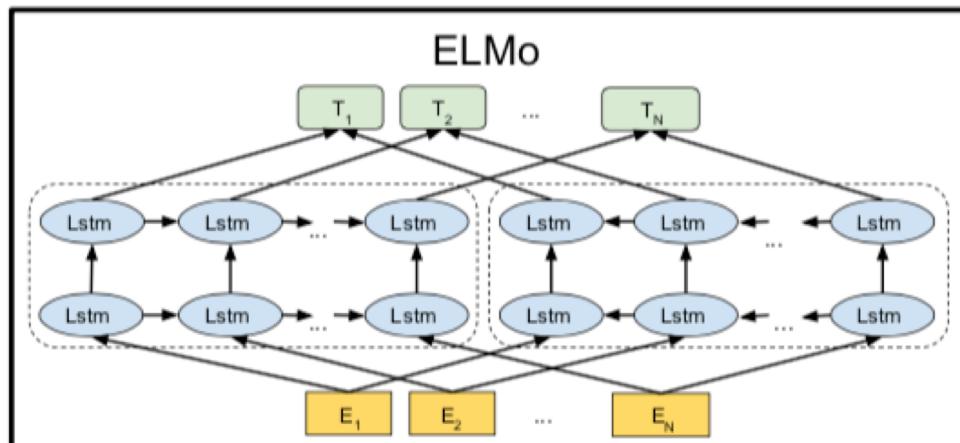
# Problems with ELMo

- 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!

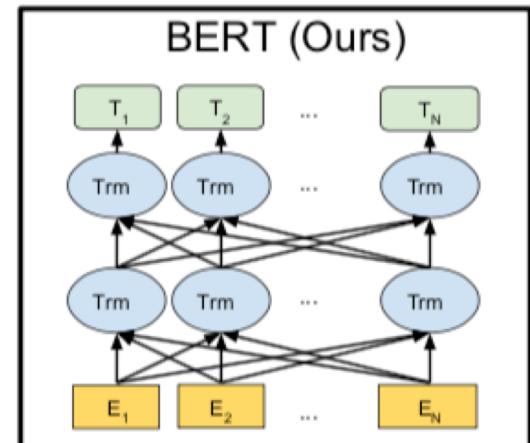
# BERT - Bidirectional Encoder Representations from Transformers

Devlin et al., NAACL'19

Solutions: use Transformer + encoder layers instead of decoder layers



(OpenAI GPT)

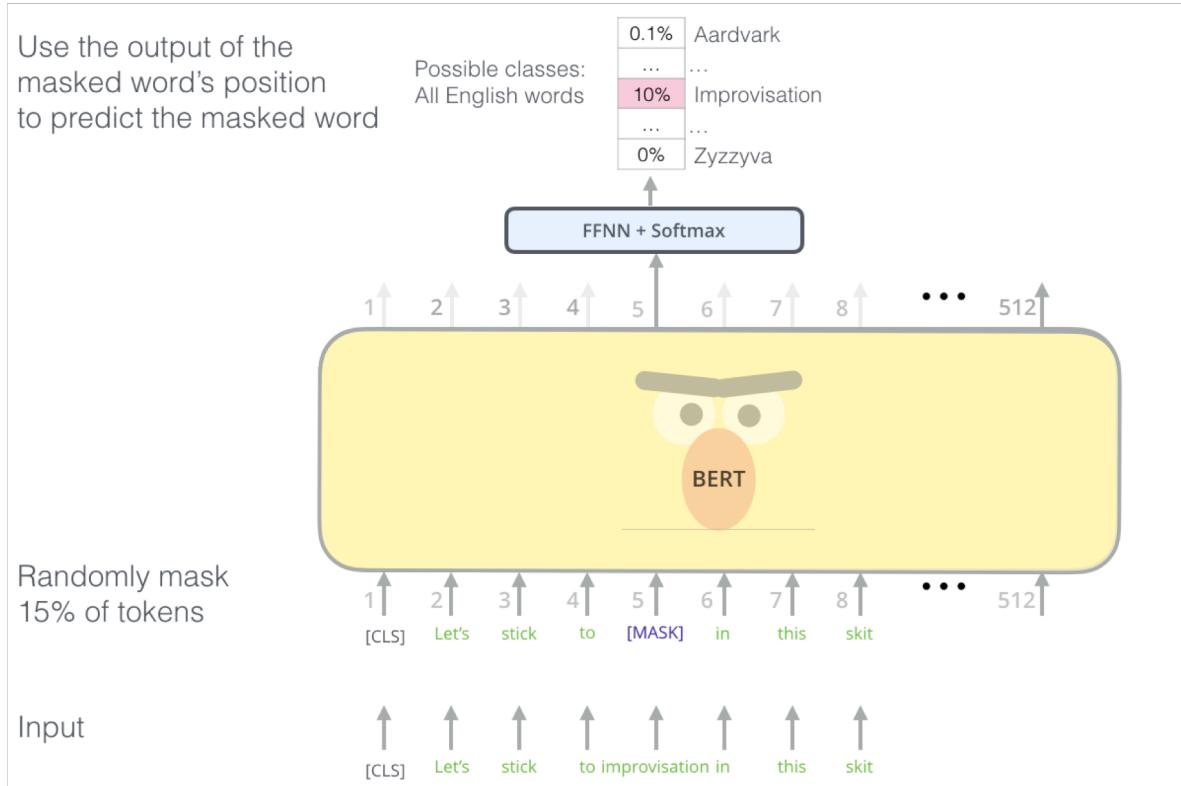


Innovation with multiple pretraining tasks

# 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>”

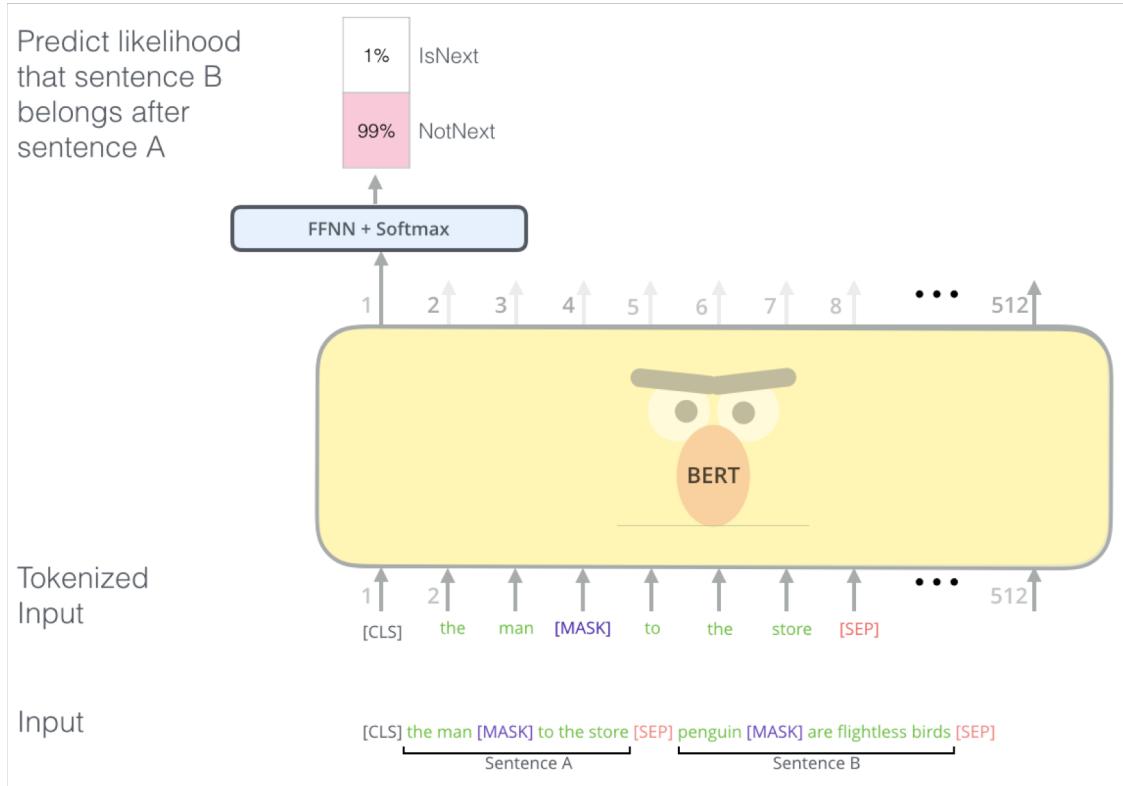
# BERT – Pretraining 1: masked language modeling



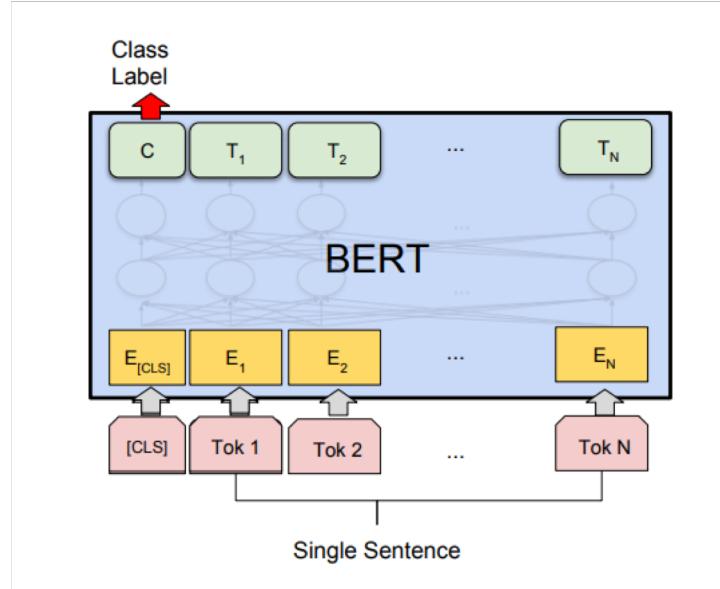
## BERT – Pretraining 2: next sentence prediction

- 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

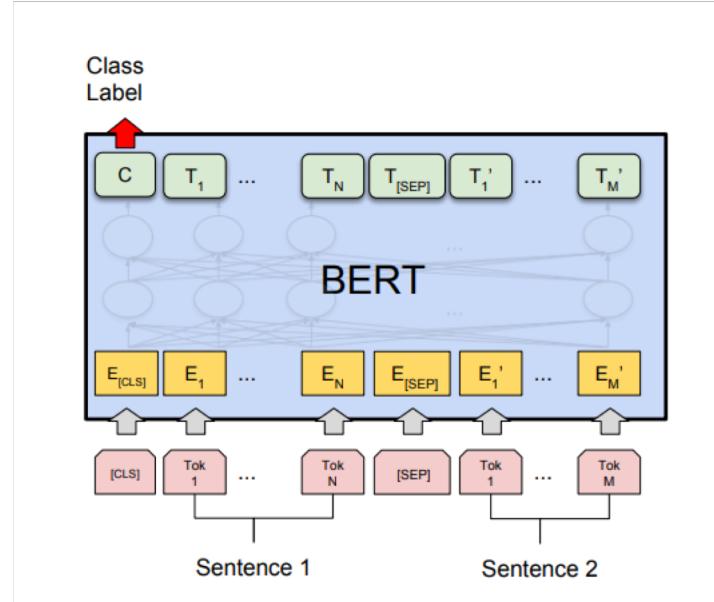
# BERT – Pretraining 2: next sentence prediction



# BERT – Fine-tuning for Classification

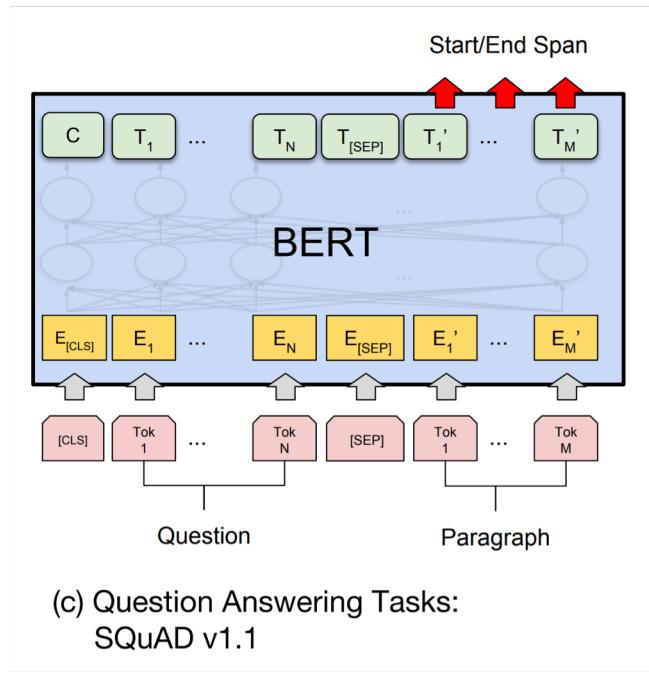


**Single sentence classification**  
Sentiment analysis, spam detection, etc.



**Pair of sentences classification**  
Entailment, paraphrase detection, etc.

# BERT – Fine-tuning for Machine Reading



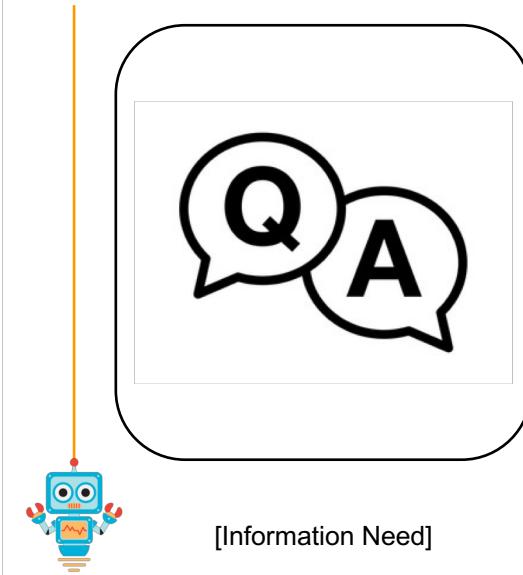
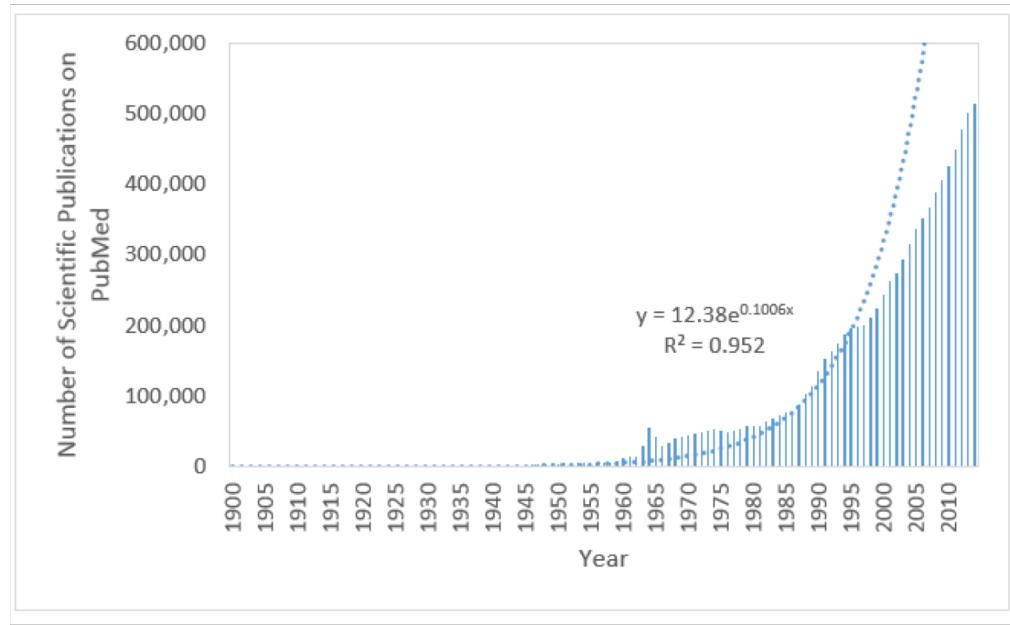
System	Dev		Test	
	EM	F1	EM	F1
Leaderboard (Oct 8th, 2018)				
Human	-	-	82.3	91.2
#1 Ensemble - nlnet	-	-	86.0	91.7
#2 Ensemble - QANet	-	-	84.5	90.5
#1 Single - nlnet	-	-	83.5	90.1
#2 Single - QANet	-	-	82.5	89.3
Published				
BiDAF+ELMo (Single)	-	85.8	-	-
R.M. Reader (Single)	78.9	86.3	79.5	86.6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

# Open Domain Question Answering

---

Is Machine Reading actually useful?

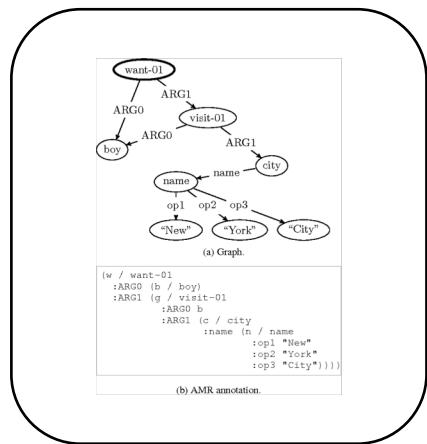
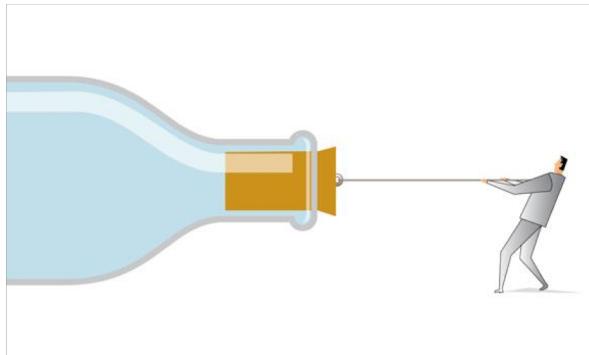
# Motivation 1: Information Overload



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



[Meaning]



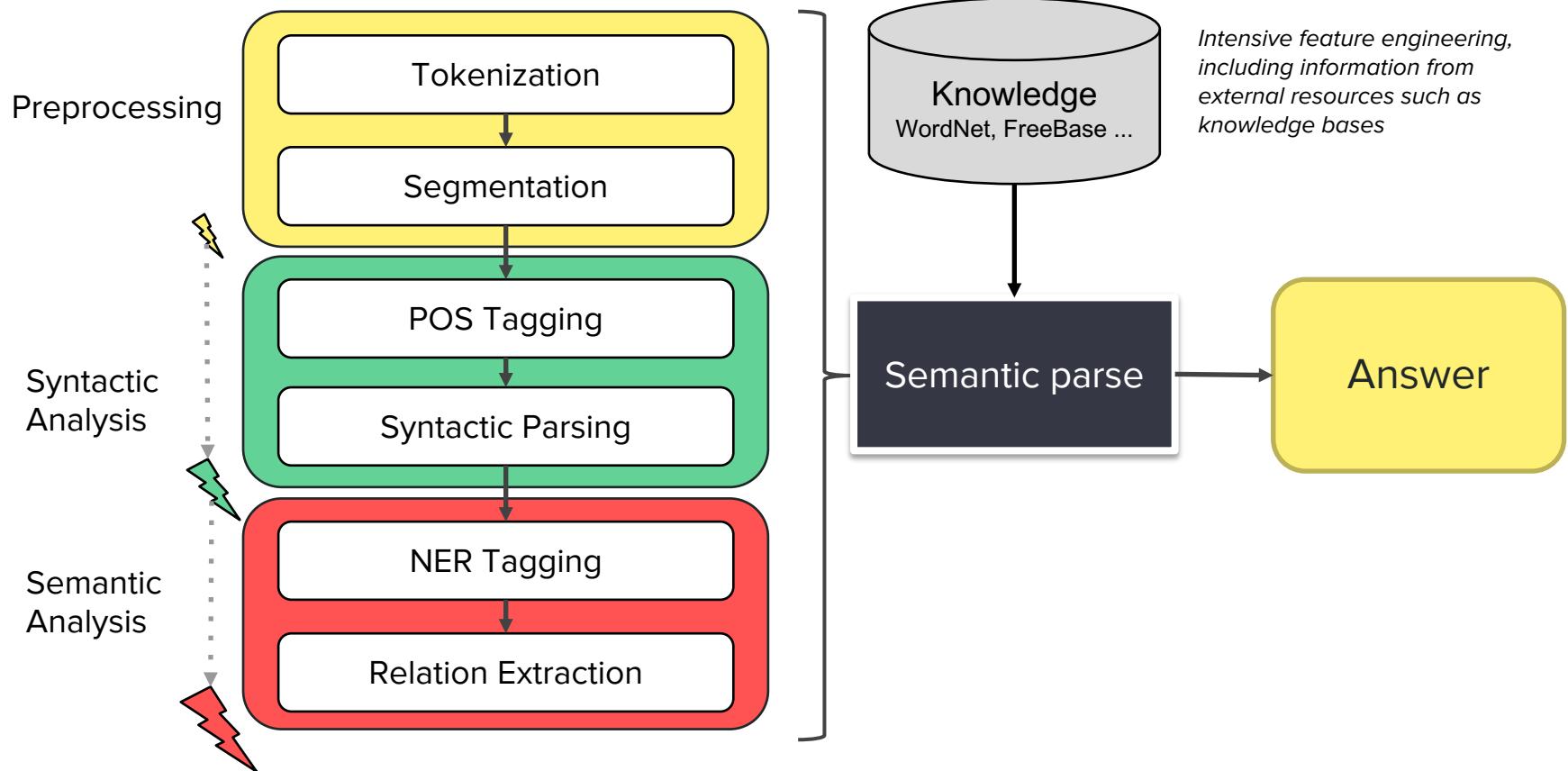
uses for



# 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  
mozzarella in his car

[Text]

$$\begin{aligned} \exists x_0 \text{named}(x_0, \text{ewan}, \text{person}) \wedge \\ \exists x_1 \text{mozzarella}(x_1) \wedge \\ \exists x_2 \text{car}(x_2) \wedge \text{of}(x_2, x_0) \wedge \text{in}(x_1, x_2) \wedge \\ \exists e \text{event}(e) \wedge \text{forget}(e) \wedge \text{agent}(e, x_0) \wedge \\ \text{patient}(e, x_1) \end{aligned}$$

[Meaning]

[Information Need]



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

# Knowledge Bases

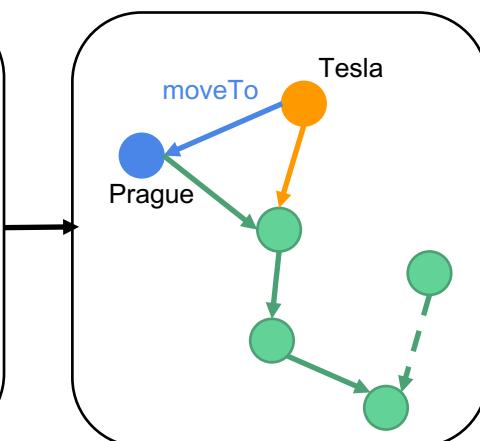
- KB: structured repository of knowledge (usually relational DB)
- Goal: encode knowledge so that it can be queried by semantic parses efficiently
- Scale can be huge: billions of facts, millions of entities
- KB can be generic or specific
- Examples: Cyc, WikiData, DBPedia, Google KG, GeneOntology, 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 Gospic 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]

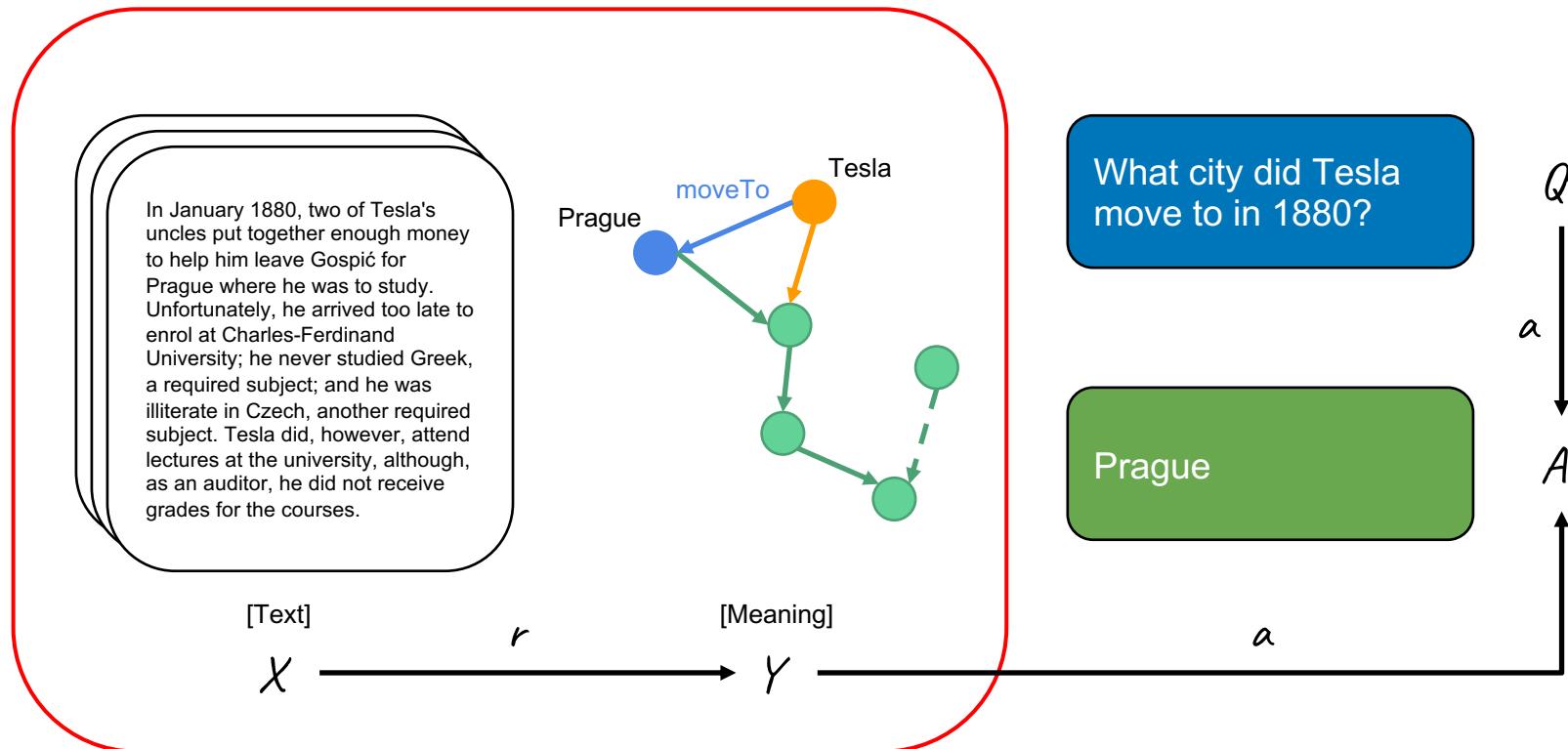


[Meaning]



[Information Need]

# Knowledge Graph Construction



# Knowledge Graph Construction

In January 1880, two of Tesla's uncles put together enough money to help him leave Gospic 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.

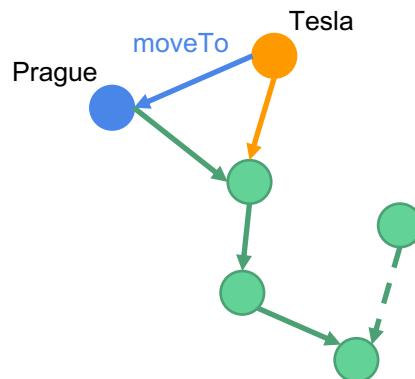
[Text]

$X$

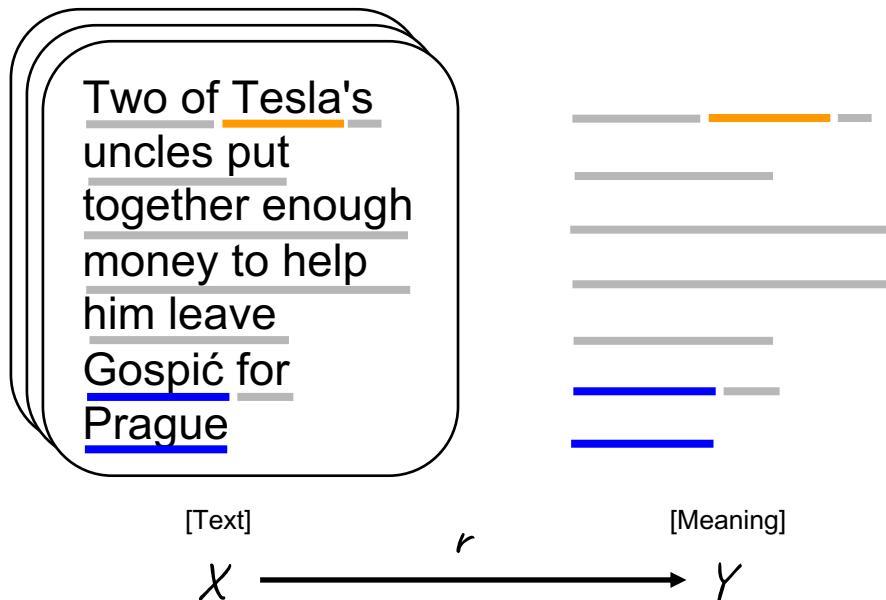
$r$

[Meaning]

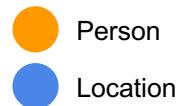
$Y$



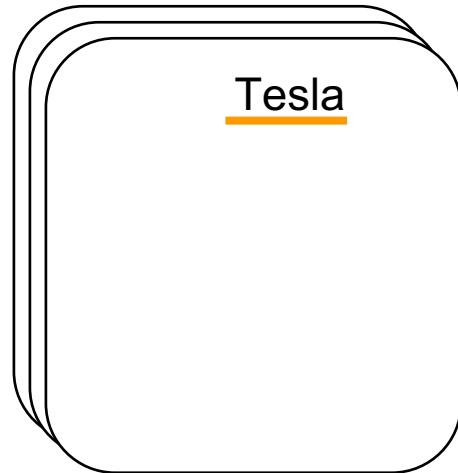
# Entity Extraction



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



# Challenge: Ambiguity

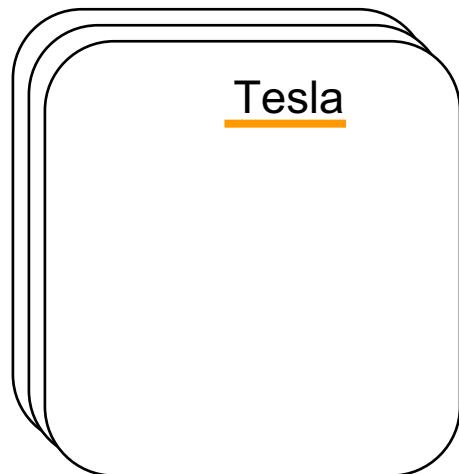


○ Person?

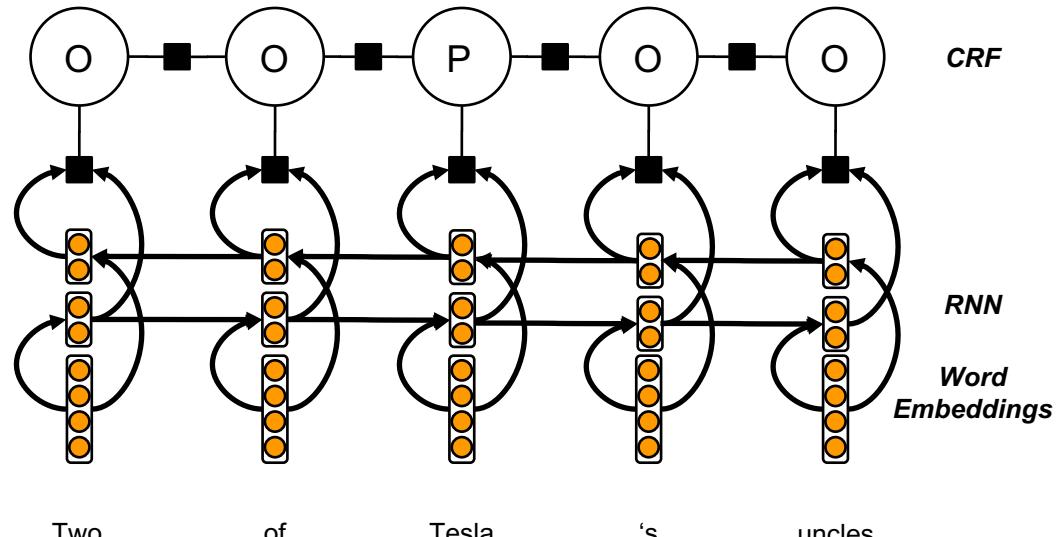
○ Brand?

# Conditional Random Fields with RNN Potentials

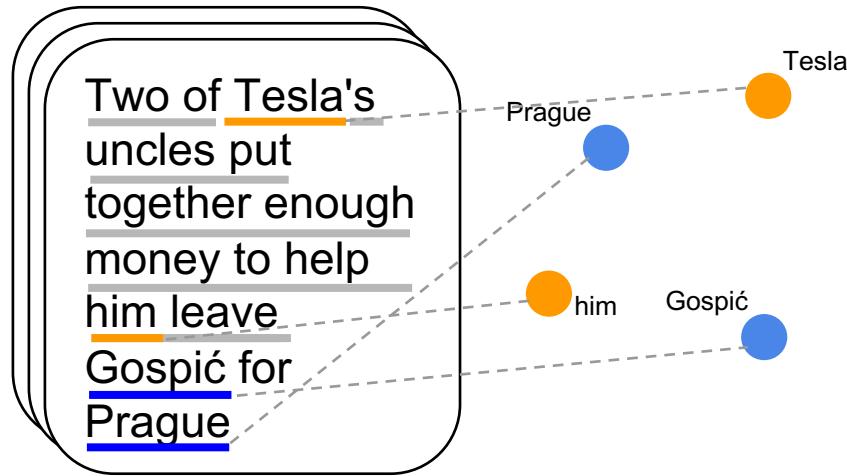
Huang et al., 2015



- Person?
- Brand?

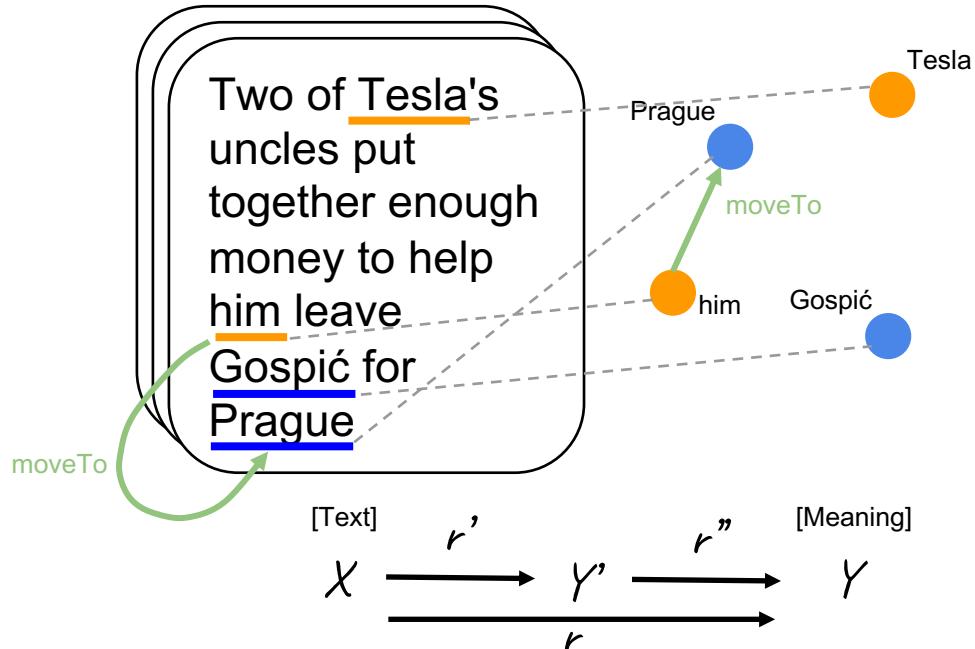


# Instantiate Nodes



- Yellow circle: Person
- Blue circle: Location

# Relation Extraction



- Neural Classification
- Distant Supervision

# Challenge: Variation

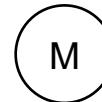
Two of Tesla's uncles put together enough money to help **him leave Gospic for Prague**

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

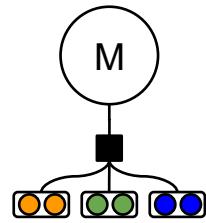
Two of Tesla's uncles put together enough money to help **him settle in Prague**

# Relation Classification

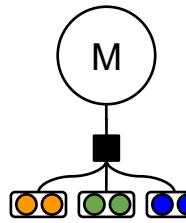
[Current SOTA neural RE model]



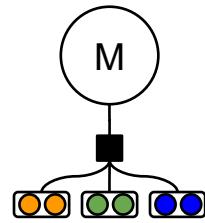
(Tesla, moveTo, Prague)



him leave  
Gospic for  
Prague



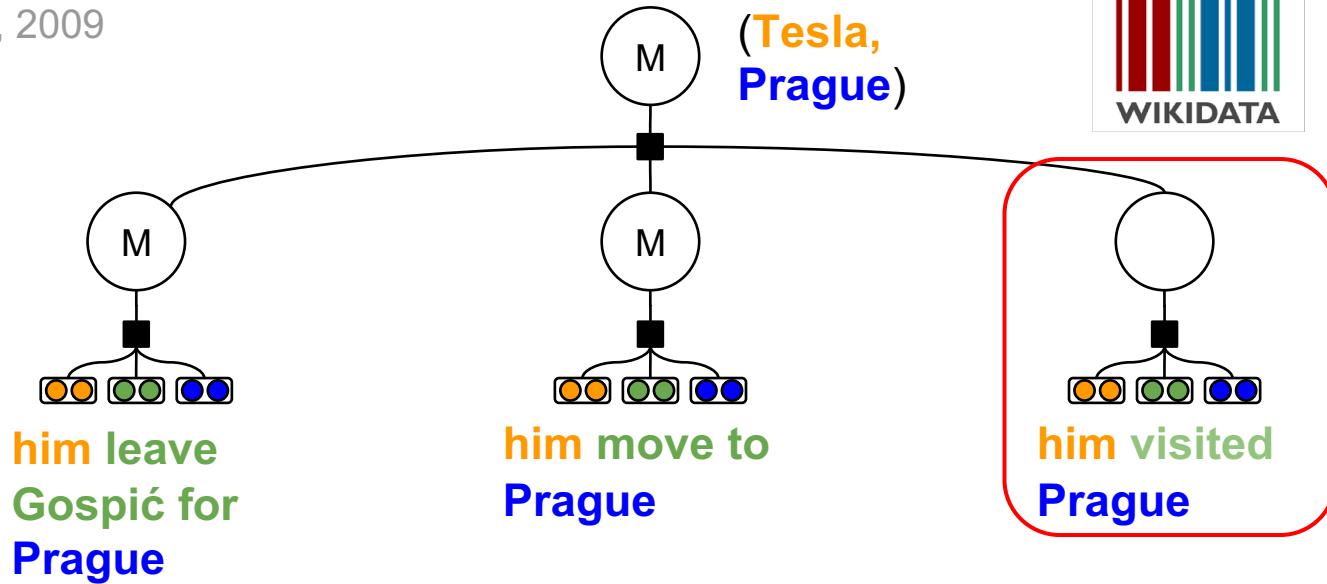
him move to  
Prague



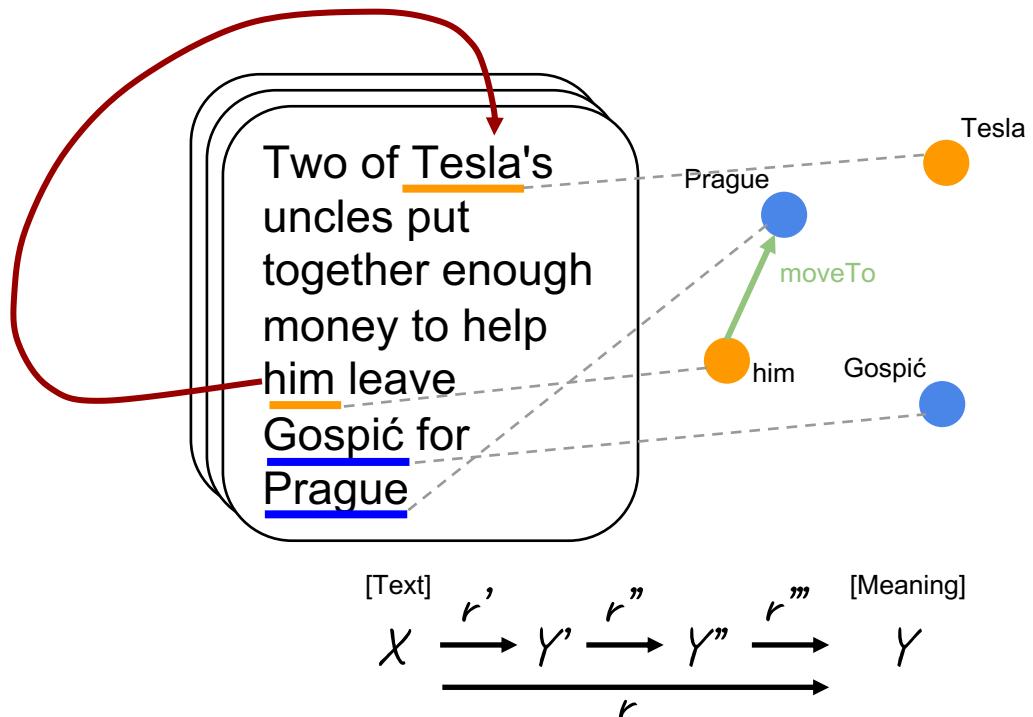
him settle in  
Prague

# Distant Supervision & Multiple Instance Learning

Mintz et al., 2009

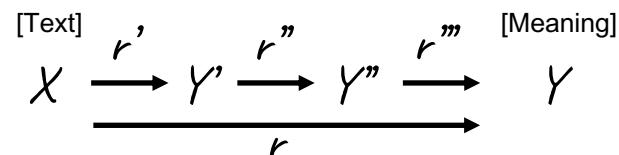
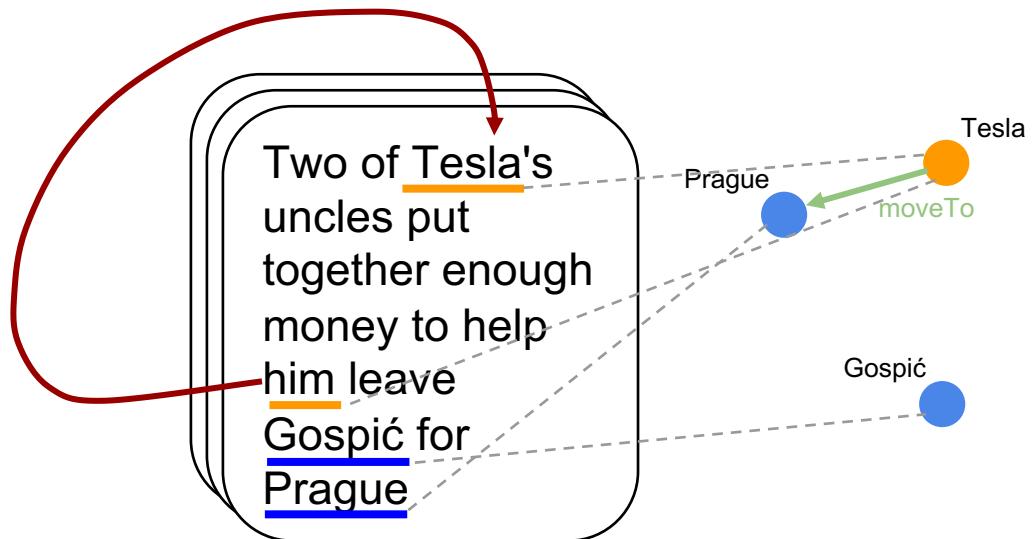


# Coreference Resolution



- Neural Classification
- Latent Variables

# Collapsing Nodes



# Challenge: Common Sense

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

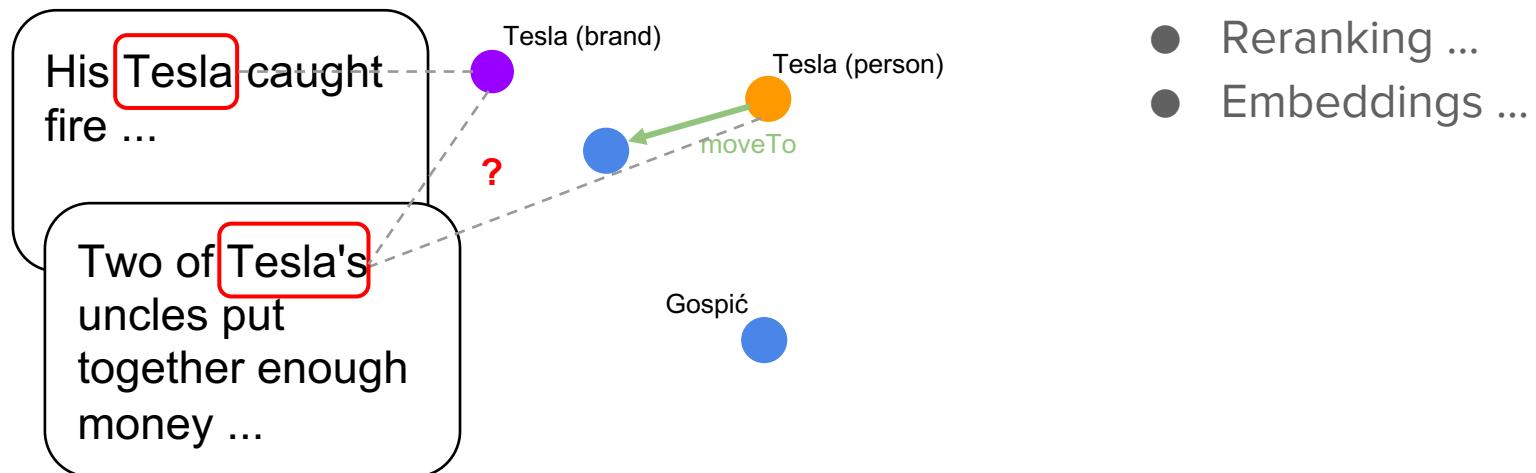
Surface

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

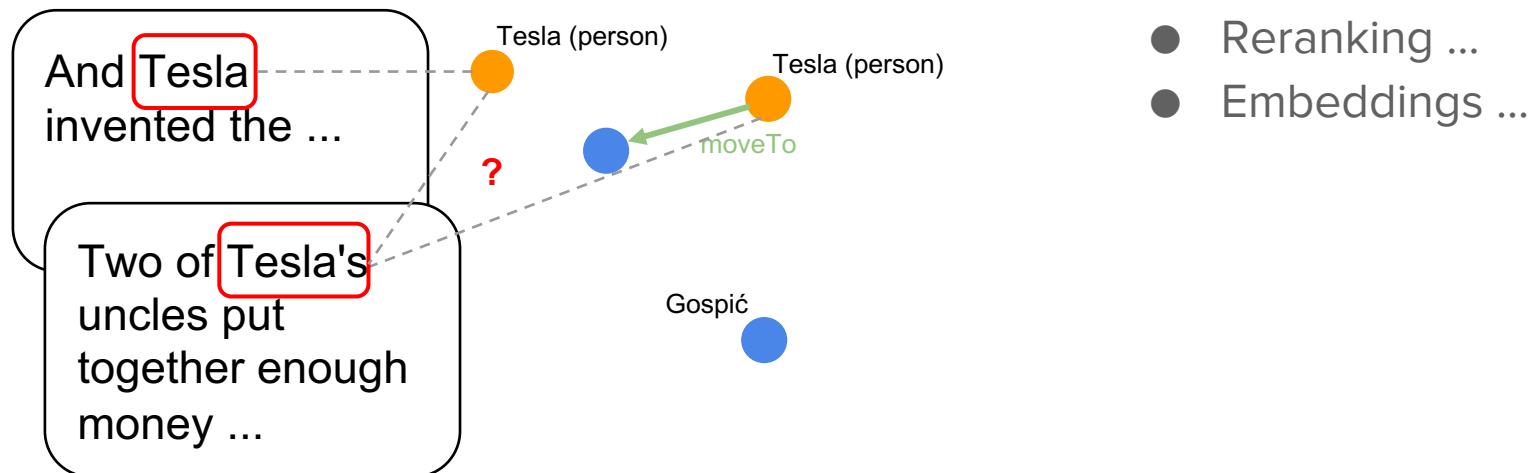
Common Sense

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

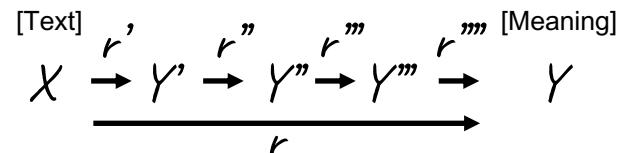
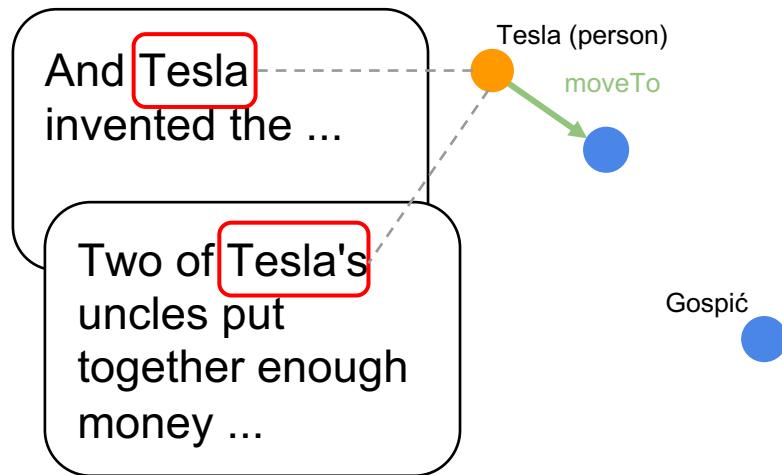
# Entity Linking



# Entity Linking

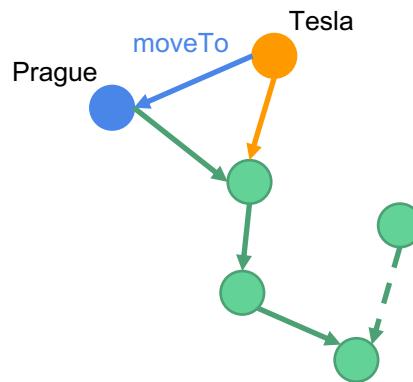


# Collapsing



# Strengths

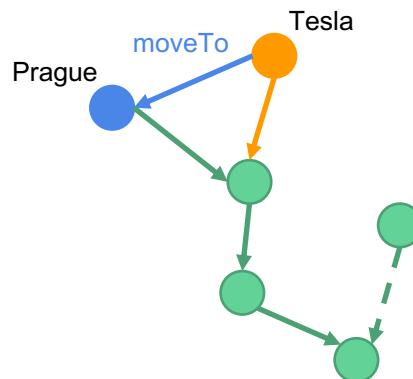
In January 1880, two of Tesla's uncles put together enough money to help him leave Gospić for Prague where he was to study. Unfortunately, he arrived too late to enrol at Charles-Ferdinand University; he never studied Greek, a required subject; and he was illiterate in Czech, another required subject.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



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

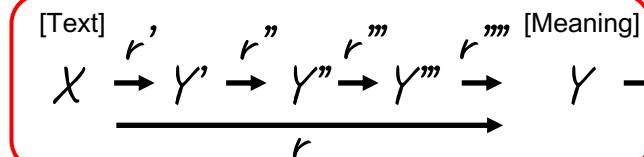
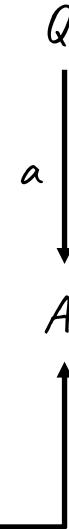
# Weakness: Cascading errors

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.



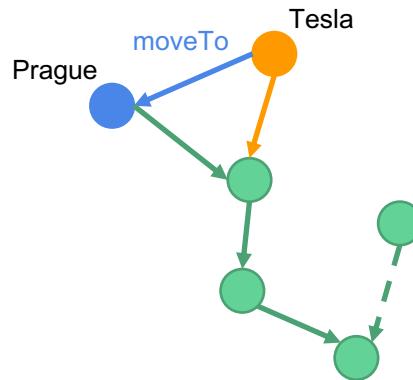
What city did Tesla move to in 1880?

Prague



# Weakness: Cascading errors

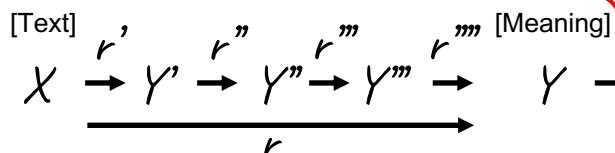
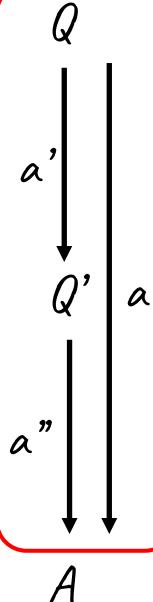
In January 1880, two of Tesla's uncles put together enough money to help him leave Gospić for Prague where he was to study. Unfortunately, he arrived too late to enrol at Charles-Ferdinand University; he never studied Greek, a required subject; and he was illiterate in Czech, another required subject. 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?

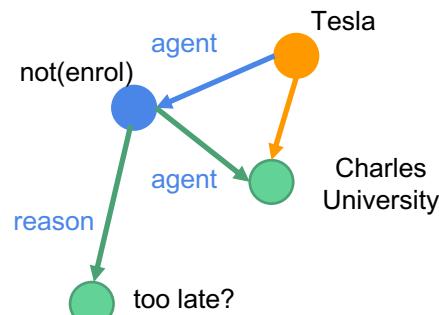
moveTo(Tesla,X)?

Prague



# Weakness: Engineering Schemas and Formalisms

Unfortunately, he arrived too late to enrol at Charles University



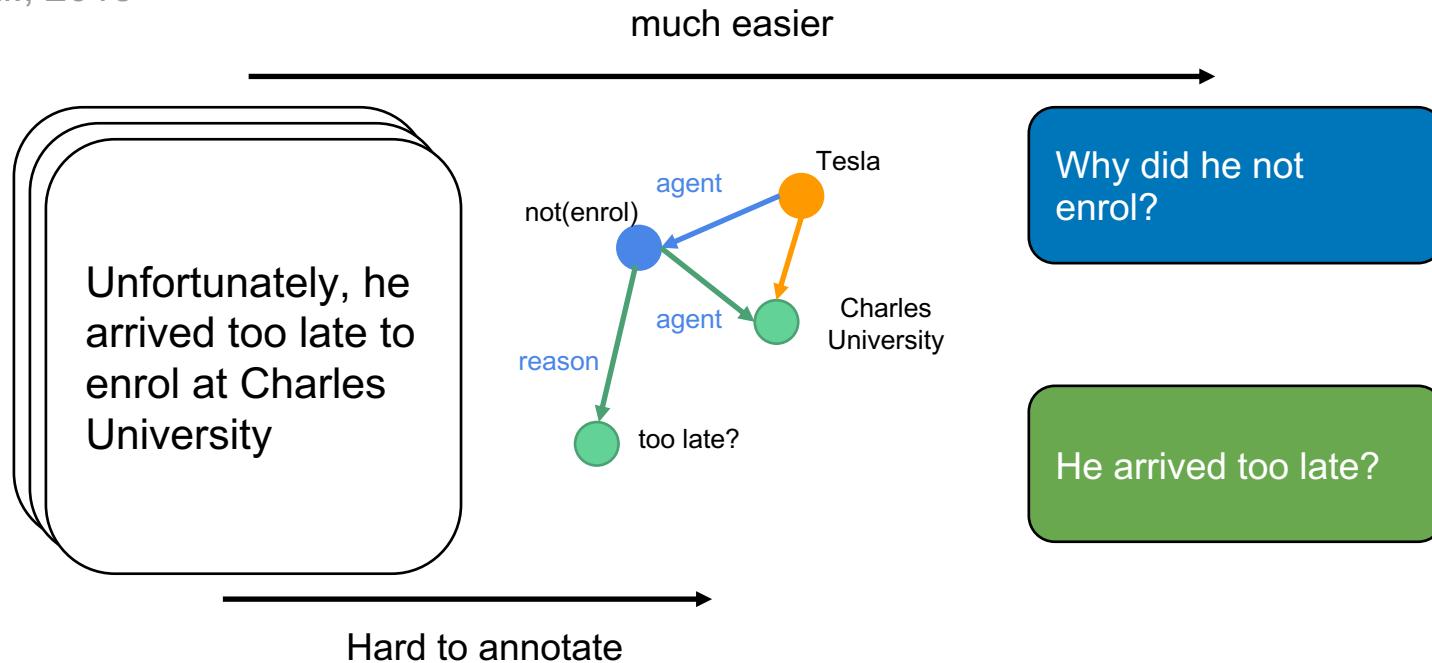
Why did he not enrol?

He arrived too late?

getting this right is hard

# Weakness: Annotation

He et al., 2015



# Structured Representations

- Advantages

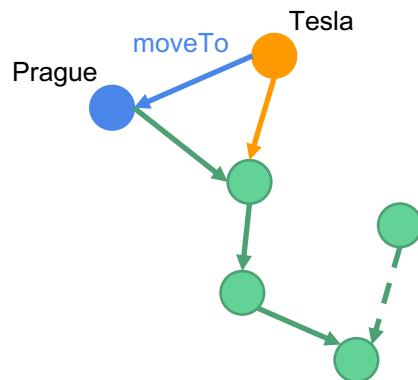
- Fast access
- Scalable
- Interpretable
- Supports reasoning
- Universality of representations: independent of question

- Disadvantages

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

# Is there another way?

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]

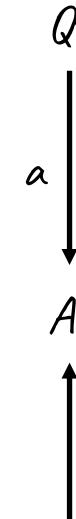
$X$

$r$

[Meaning]

$Y$

$a$



# Omitting Symbolic Meaning Representations !!

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]

X

—

What city did Tesla move to in 1880?

Prague

a

Q

a

A

# Machine Reading AT SCALE

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

# Machine Reading



[Text]

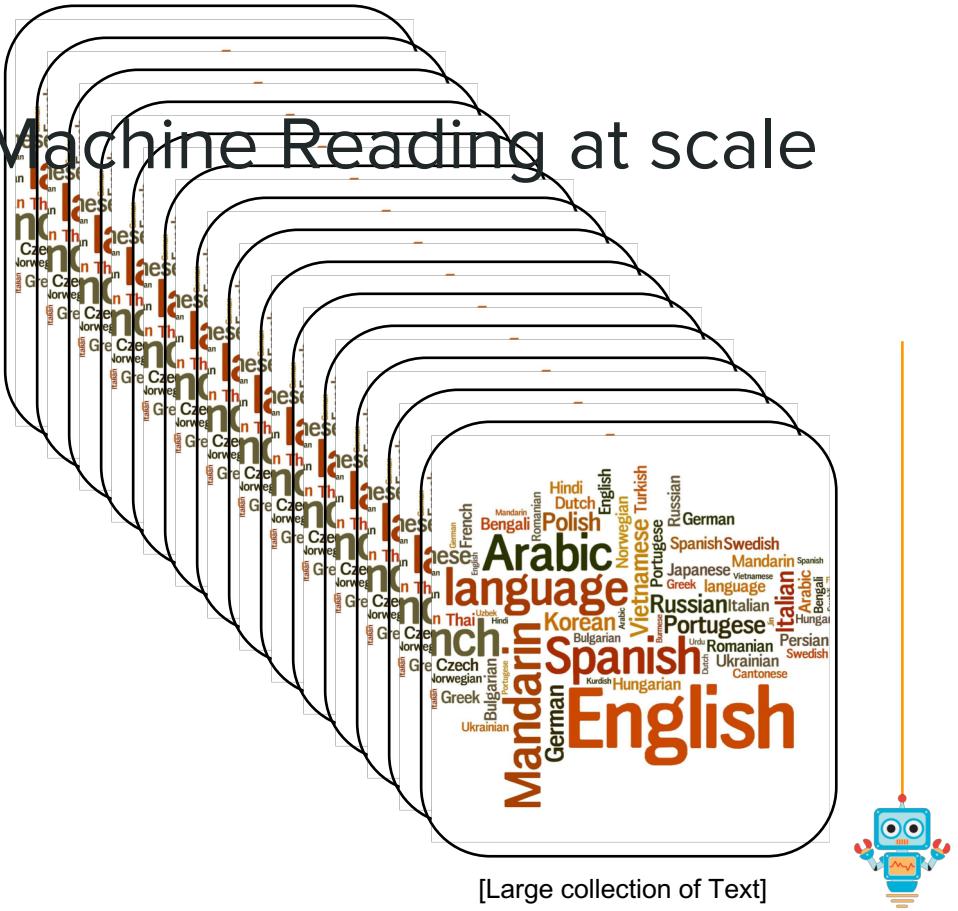


uses for



[Information Need]

# Machine Reading at scale

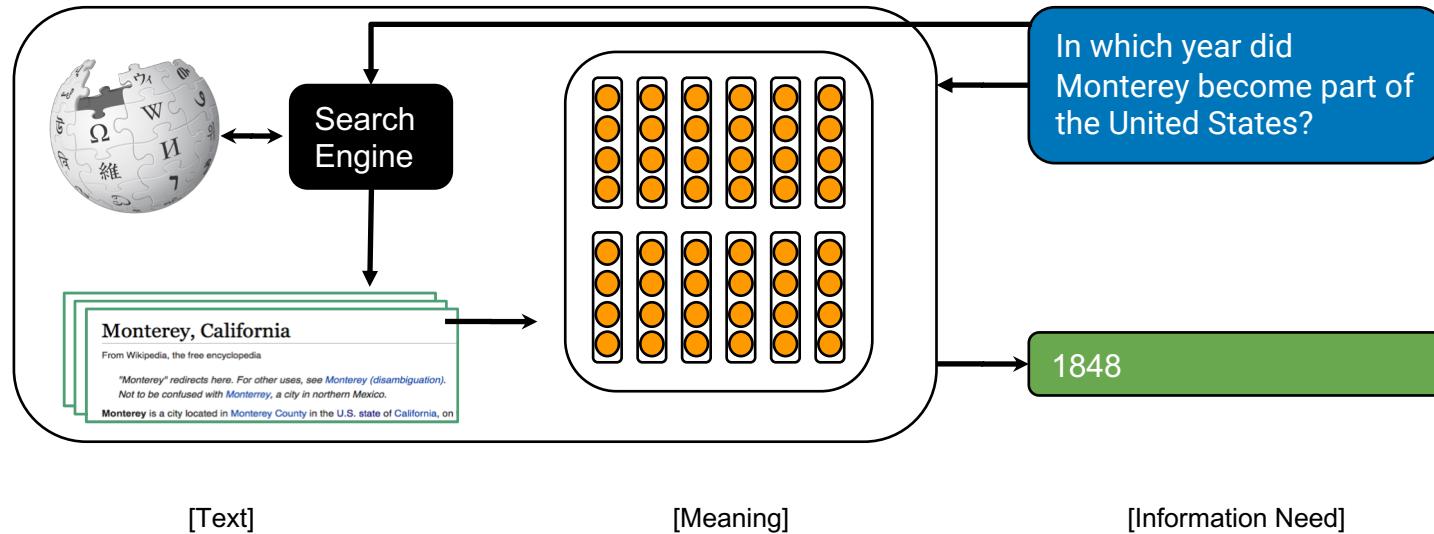


uses for



# Typical Machine Reading at Scale System

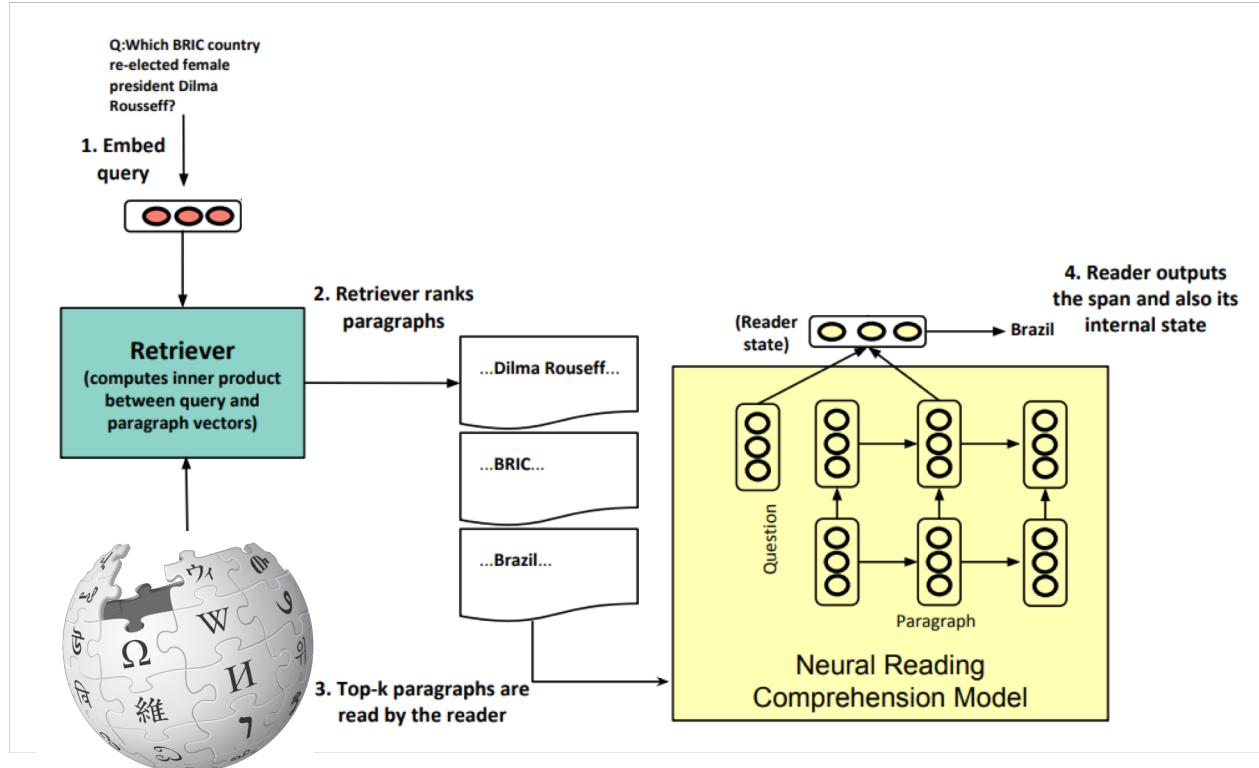
Dr.QA Chen et al., 2017



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

The diagram illustrates a Multi-Step Retriever-Reader system. It consists of two main sections, each with a query at the top, followed by two steps of retrieved paragraphs, and an answer at the bottom.

**Example 1: Query - "Diaphoresis" is a medical term for what condition?**

- Step 1:** A Greek term for hyperhidrosis is diaphoresis
- Step 2:** Hyperhidrosis is a physical condition caused by excessive sweating in the body.

**Answer:** sweating

**Example 2: Query - What is name of the ship on which Dracula arrived in England in 1897?**

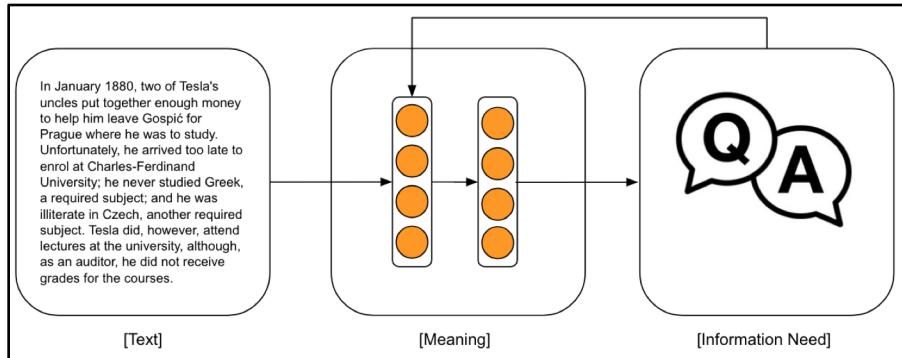
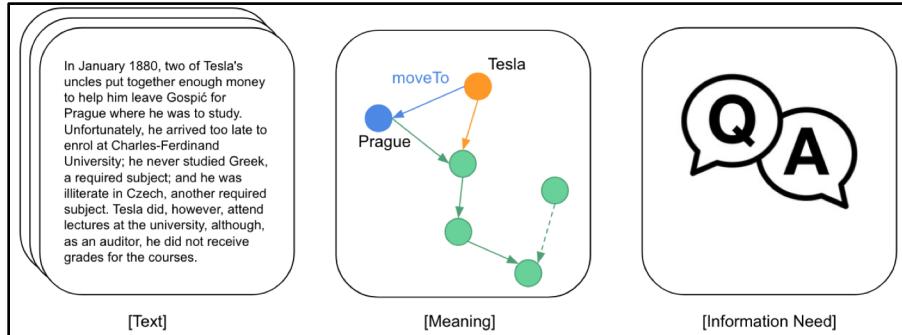
- Step 1:** The untold story of Dracula's voyage on the merchant ship "Demeter" from Transylvania to London which docked in England
- Step 2:** Dracula then sets sail on the ship Demeter to England, leaving Harker captive by Dracula's insatiable Brides

**Answer:** demeter

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
<p><i>Neural Networks</i></p> <ul style="list-style-type: none"><li>• Scale to very large datasets</li><li>• Can be used by non domain experts</li><li>• Robust to noise and ambiguity in data</li><li>• Game changers in multiple applications</li><li>• Very data hungry (mostly supervised data)</li><li>• Can't learn easily new tasks from old ones</li><li>• Not interpretable</li><li>• Relatively simple reasoning</li></ul>	<p><i>KBs, Inductive Logic Programming, etc.</i></p> <ul style="list-style-type: none"><li>• Small scale conditions</li><li>• Require heavy expert knowledge</li><li>• Very brittle with noisy, ambiguous data</li><li>• Limited applicative success</li></ul> <p>Great research opportunities!</p>

# 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, ...

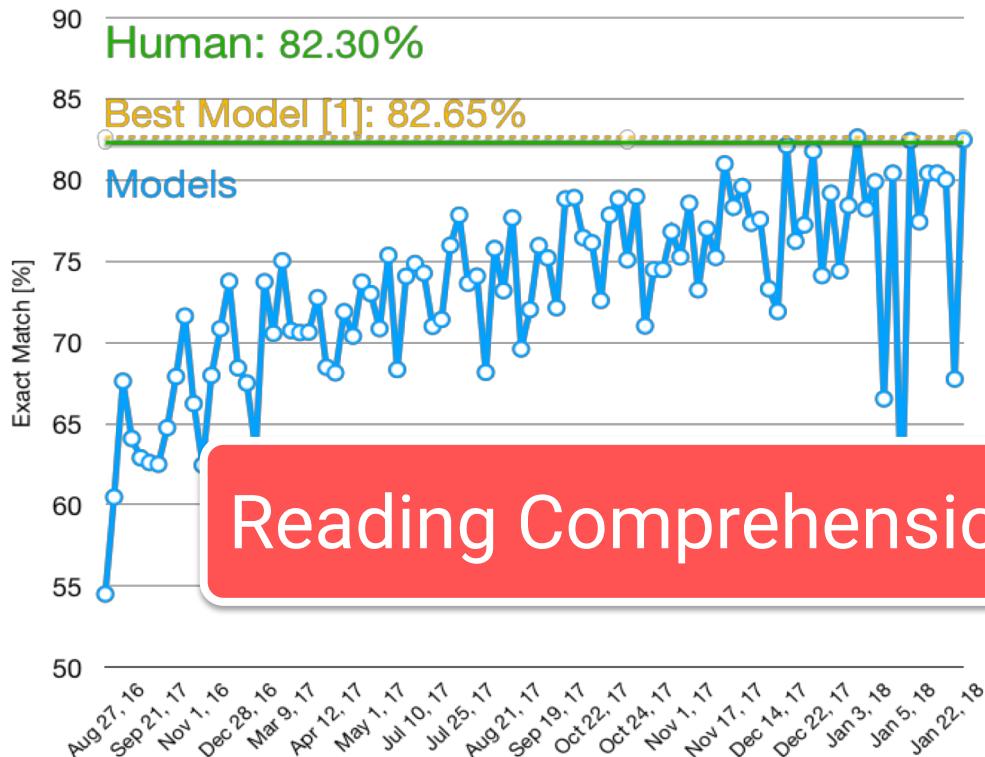
# Conclusion

- 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?

# Machine Reading & QA / 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 XXXIV?



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.

# Challenge 2: Solvability

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

What was the name of the 1937 treaty?

[UNANSWERABLE]

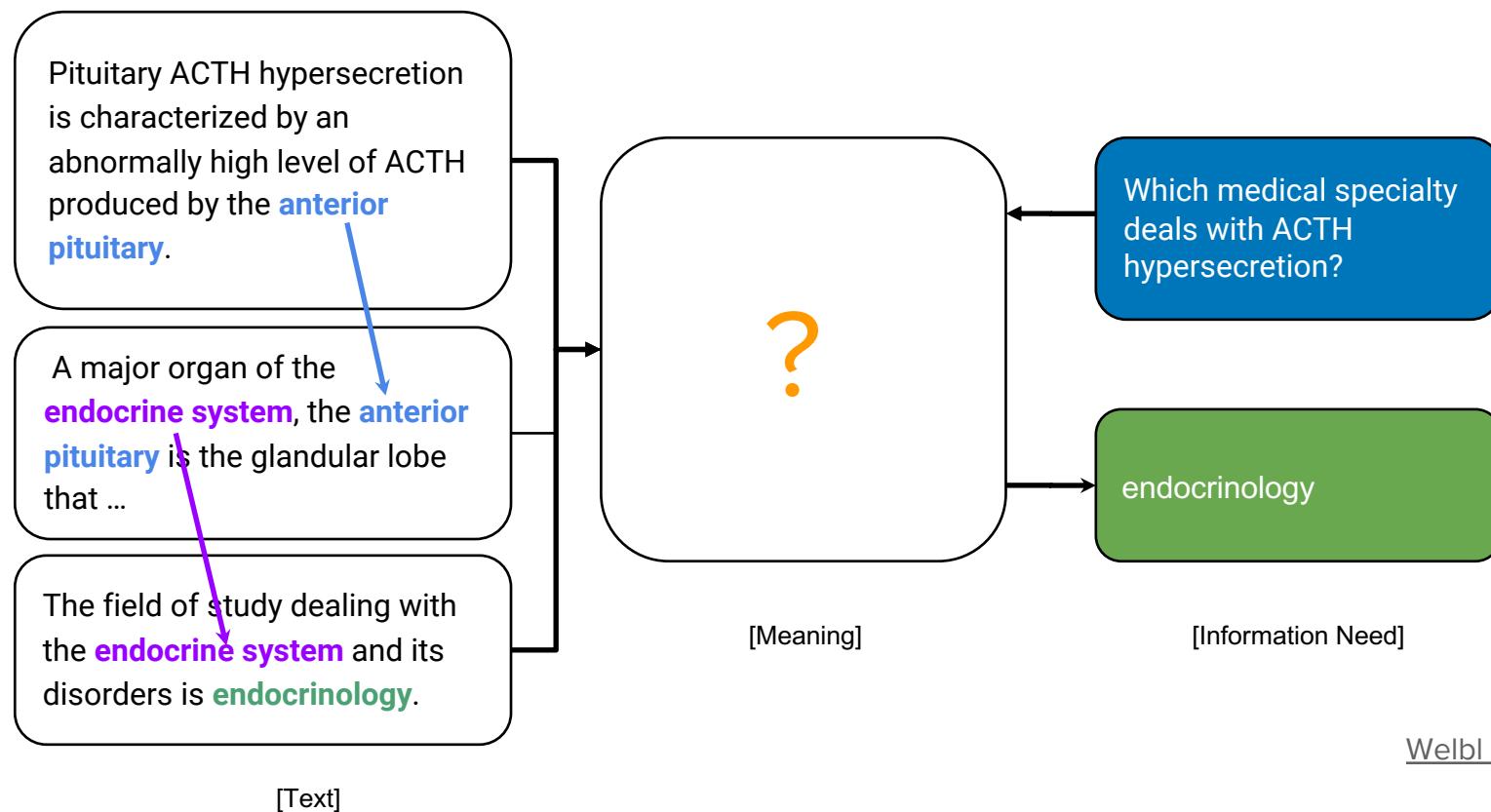
... Other legislation followed, including the **Migratory Bird Conservation Act** of 1929, a **1937 treaty** prohibiting the hunting of right and gray whales, and the **Bald Eagle Protection Act** of 1940.

System	SQuAD 1.1 test		SQuAD 2.0 dev		SQuAD 2.0 test	
	EM	F1	EM	F1	EM	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	<b>78.6</b>	<b>85.8</b>	<b>65.1</b>	<b>67.6</b>	<b>63.4</b>	<b>66.3</b>
Human	82.3	91.2	86.3	89.0	86.9	89.5
Human–Machine Gap	3.7	5.4	<b>21.2</b>	<b>21.4</b>	<b>23.5</b>	<b>23.2</b>

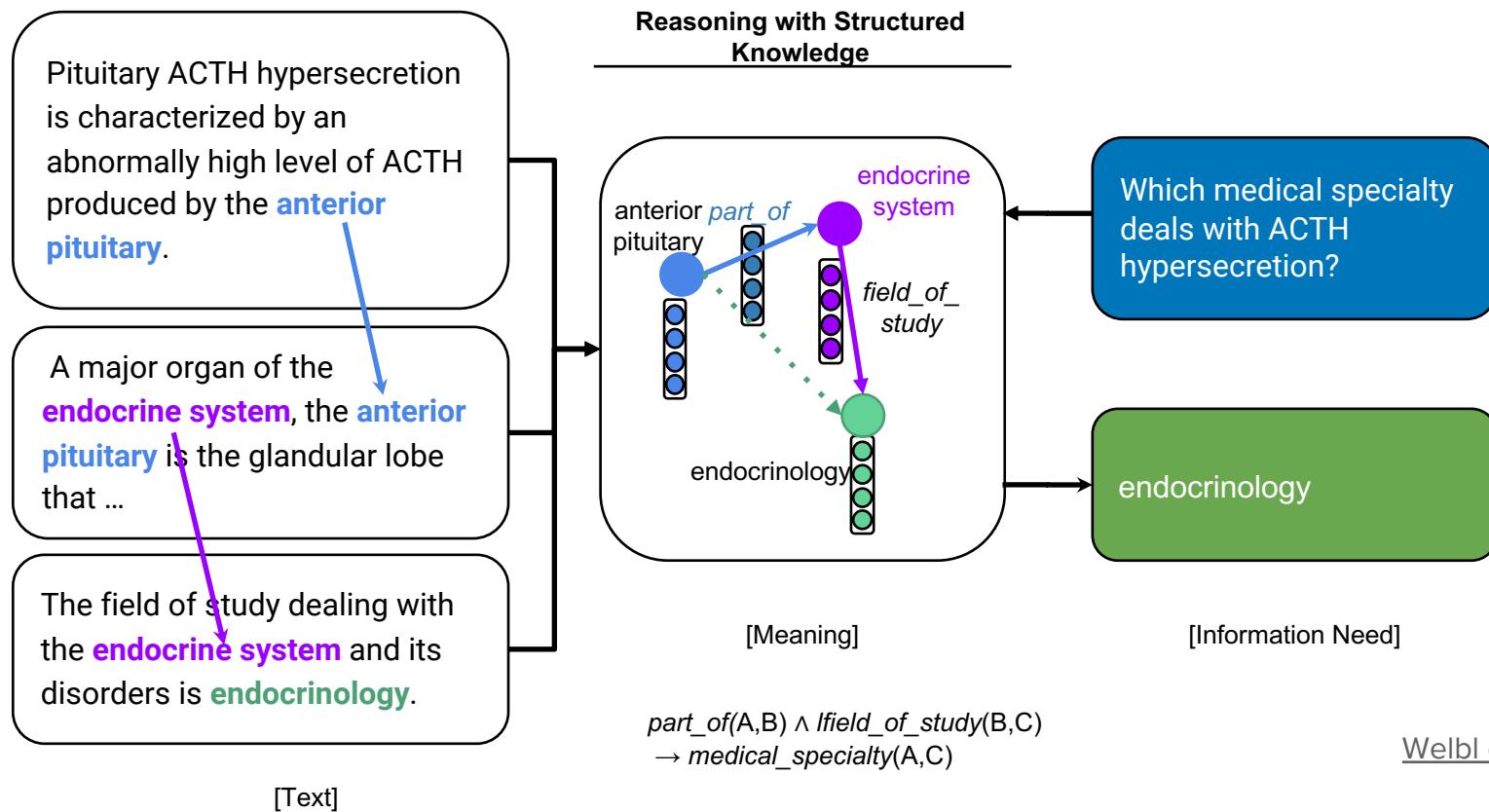
# Challenge 3: Limited Supervision

- 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)
  - **unsupervised pretraining, e.g. ELMo, Peters et al. (2018)**

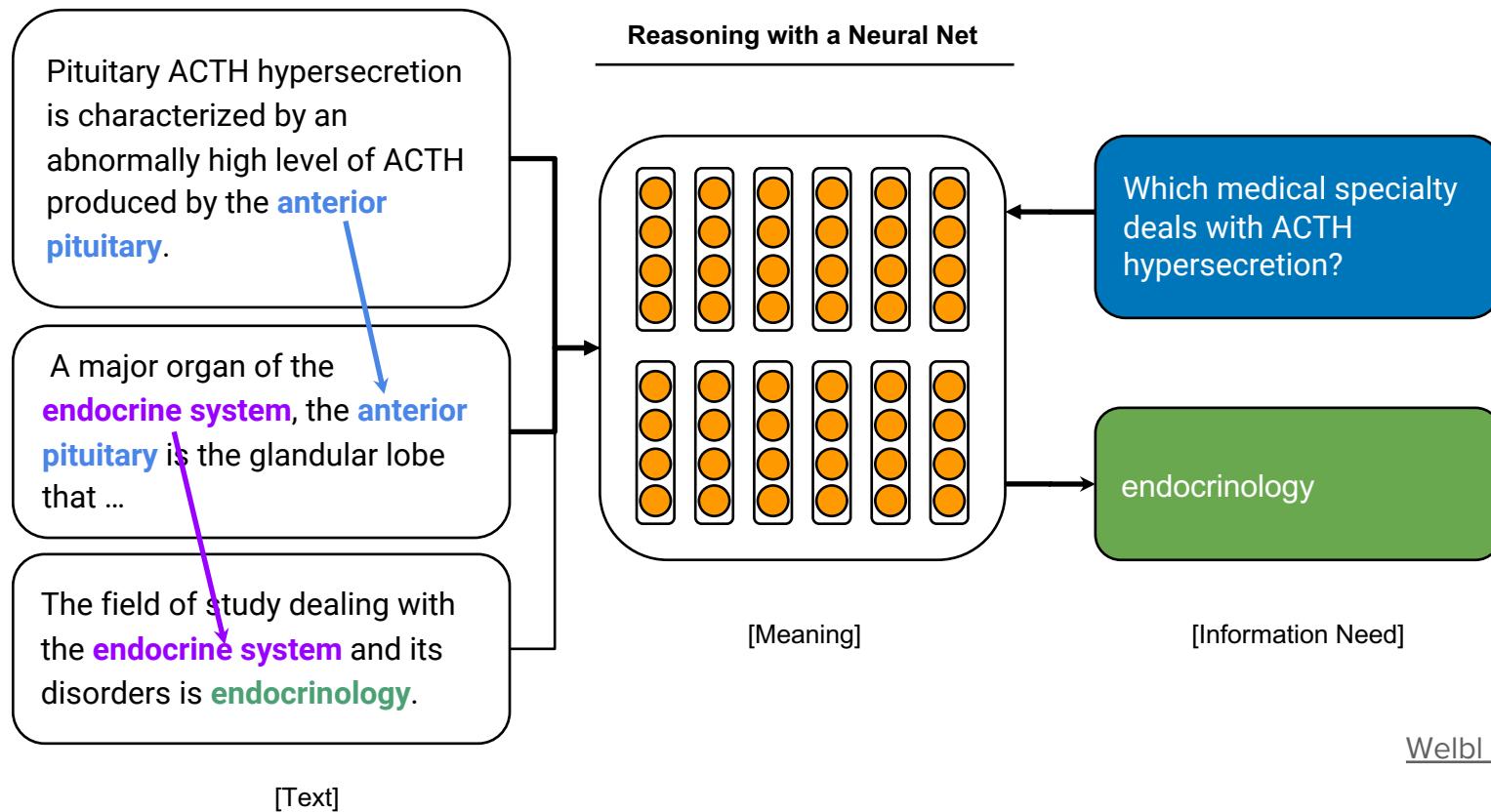
# Challenge 4: Reasoning with Text



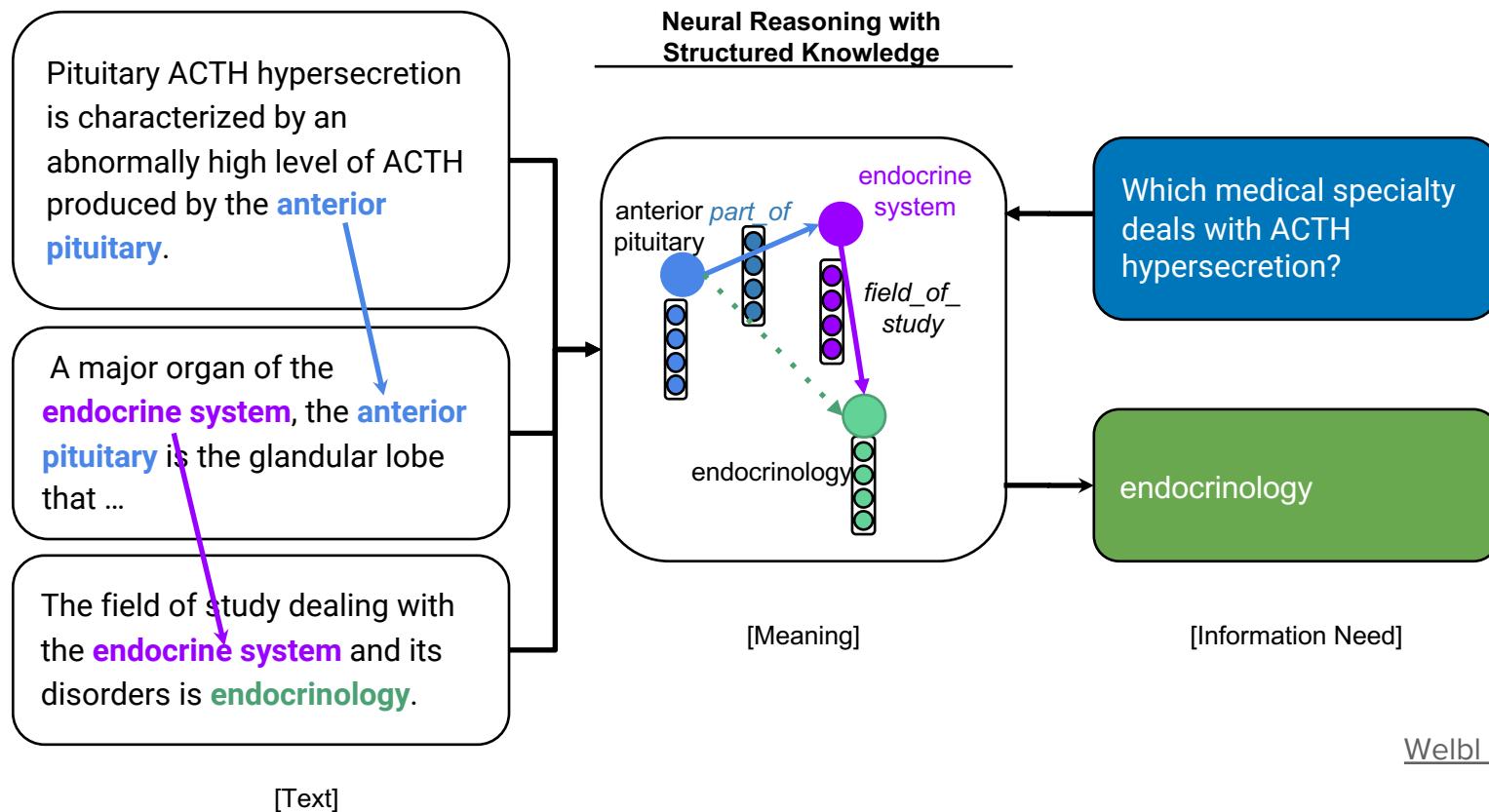
# Challenge 4: Reasoning with Text



# Challenge 4: Reasoning with Text



# Challenge 4: Reasoning with Text



# 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?

**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?

# Dialog

---

How about language with interactions?

# Bots! Bots! Bots!



Hola poncho, what's up!!

Just sitting here refreshing the page. U?

Same. Packing for Barcelona

Fascinating.

yeah well can you tell me the weather there?

It's clear and -3°C in Manhattan, NY rn.

no I meant weather in Barcelona

It's clear and 14°C in Barcelona, Spain.

Should I pack a swimsuit?

Oops, I didn't catch that. For things I can help you with, type "help."

# Terms

- **Utterance:** single sentence or line produced by a human or a dialog agent.
- **Turn:** one utterance in a sequence of consecutive utterances
- **Dialog:**
  - A sequence of turns
  - This can be as few of two turns
- **Context:** Either outside information or previous turns in the dialog
- These all refer to a dialog with two turns:
  - Source/target pair
  - Query/response pair
  - Message/response pair

# Types of Dialog Systems

## Goal-oriented Dialog Agents

- **Goals:**
  - have short conversations
  - getting information from the user to help complete a specific task.
- **Implementation:**
  - Rule-based
  - End-to-end (a bit)
- **Evaluation:** Goal achieved or not

## Chatbots/chit-chat bots

- **Goals:**
  - mimic the unstructured conversations characteristic of human-human interaction.
  - engage user as long as possible
  - Sometimes accomplish an indirect task
- **Implementation**
  - Rule-based
  - Information retrieval
  - End-to-End
- **Evaluation:** User is having a good time?

# Dialog evaluation is hard

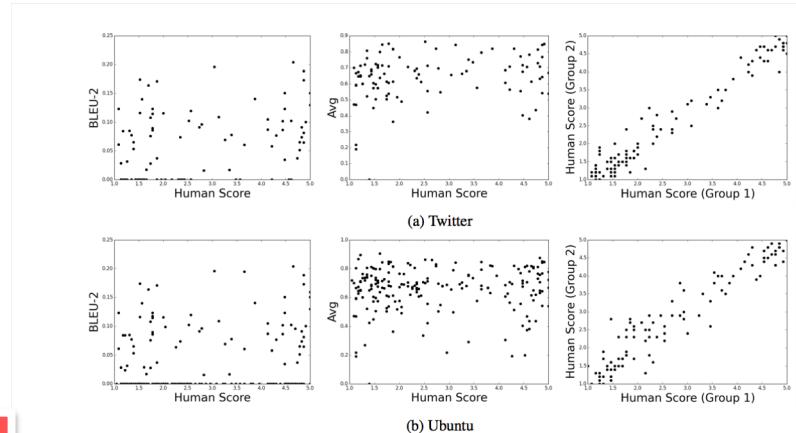
Human evaluations (AMT, etc.):

- **PROS:** test fluency, task completion, actual task
- **CONS:** costly, non-reproducible

Automatic evaluation (BLEU, perplexity, etc.)

- **PROS:** fast, scalable, reproducible,
- **CONS:** non-correlated with actual human eval.

“many metrics commonly used in the literature for evaluating unsupervised dialogue systems do not correlate strongly with human judgement.”



How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation

Chia-Wei Liu<sup>1\*</sup>, Ryan Lowe<sup>1\*</sup>, Iulian V. Serban<sup>2\*</sup>,  
Michael Noseworthy<sup>1\*</sup>, Laurent Charlin<sup>1</sup>, Joelle Pineau<sup>1</sup>

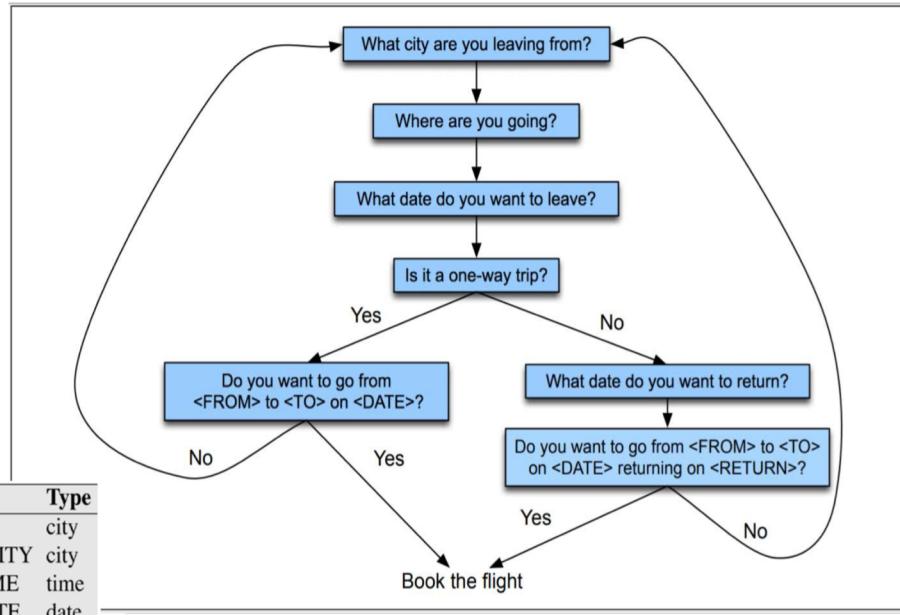
# Dialog / Goal-oriented

---

# Frame-Based Agents for Goal-oriented Dialog

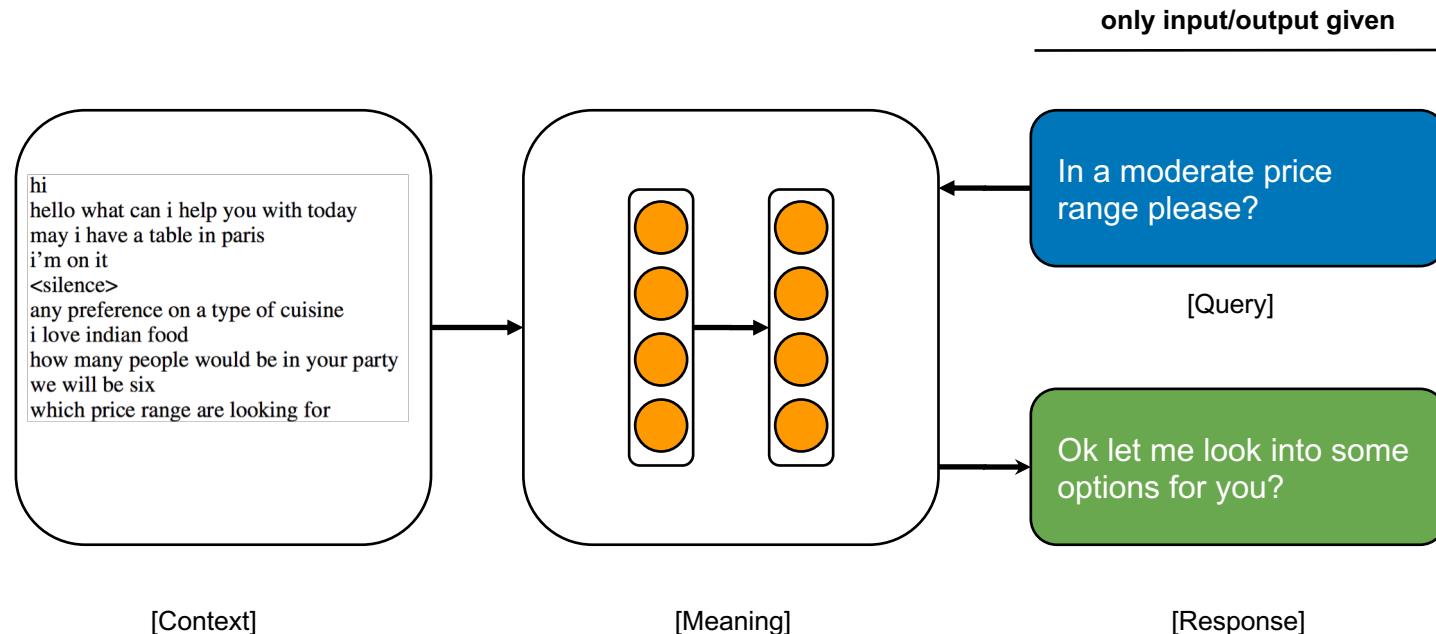
- A **frame** consists of a set of **slots** the dialog agent is trying to fill in
  - i.e. Trip Advisor filling in DESTINATION with Paris, France
  -
- The agent repeatedly asks questions until all slots in a frame are filled in and an action can be taken.
- Questions are chosen through use of an Finite-State Automaton

Slot	Type
ORIGIN CITY	city
DESTINATION CITY	city
DEPARTURE TIME	time
DEPARTURE DATE	date
ARRIVAL TIME	time
ARRIVAL DATE	date



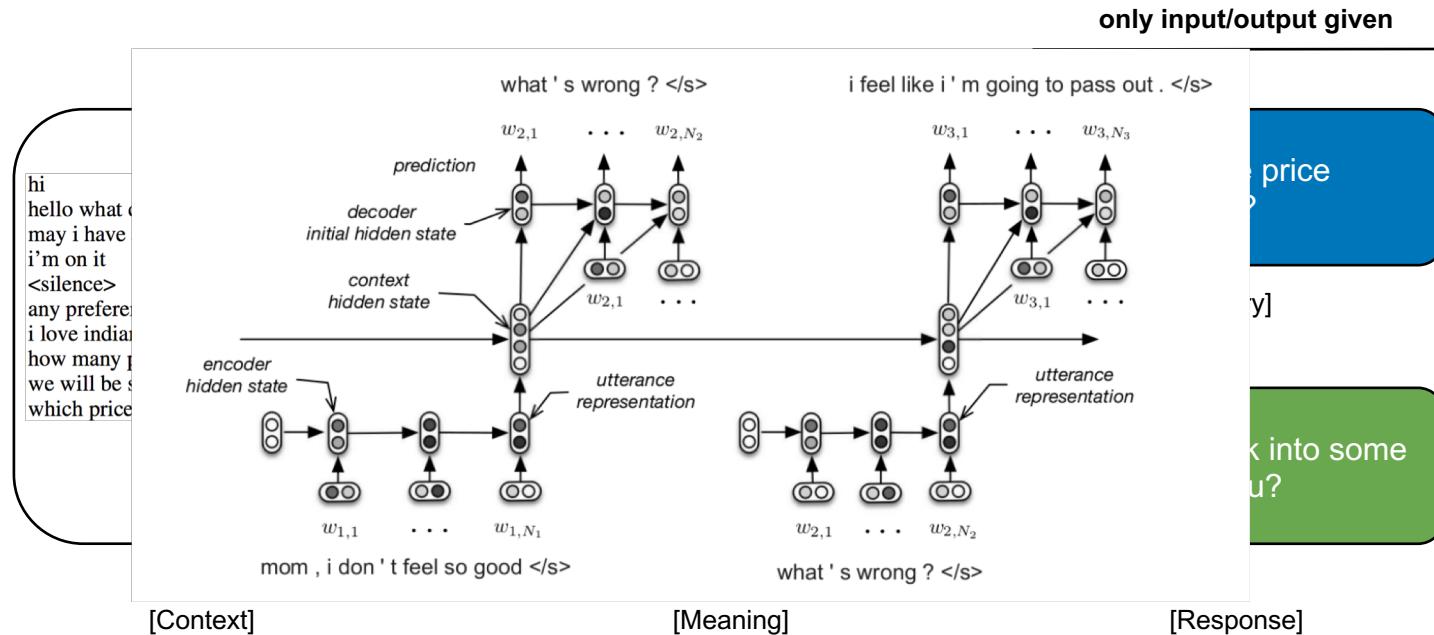
# End-to-end Dialog Systems

Shang *et al.*, 15; Vinyals and Le, 15; Sordoni *et al.*, 15; Serban *et al.*, 15; Dodge *et al.*, 16



# End-to-end Dialog Systems

Shang *et al.*, 15; Vinyals and Le, 15; Sordoni *et al.*, 15; Serban *et al.*, 15; Dodge *et al.*, 16



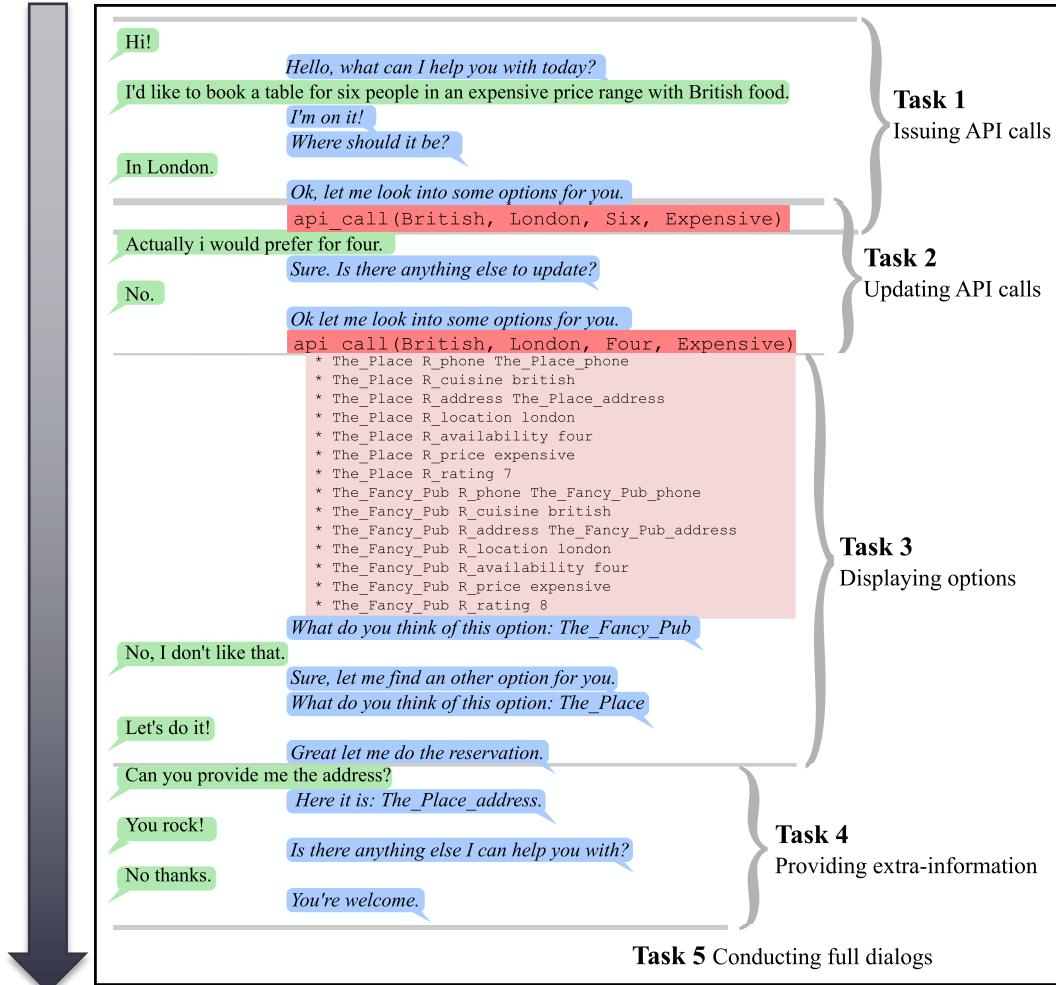
No intermediate task or state that can be used for training and evaluation!

# A dataset for end-to-end goal oriented dialog

- Based on a **simulation like bAbI**:
  - **A KB made of restaurants and their properties**: type of cuisine, location, price, rating, party size, address and phone number.
  - KB can be queried using API calls containing 4fields: location, type of cuisine, price range and party size.
  - Dialogs are generated after creating a user request by sampling an API call.
- Each example is **a dialog between a user and a bot**, as well as API calls and the resulting facts.
- **NL patterns** are used to create user and bot utterances.

# Slicing a booking in five tasks

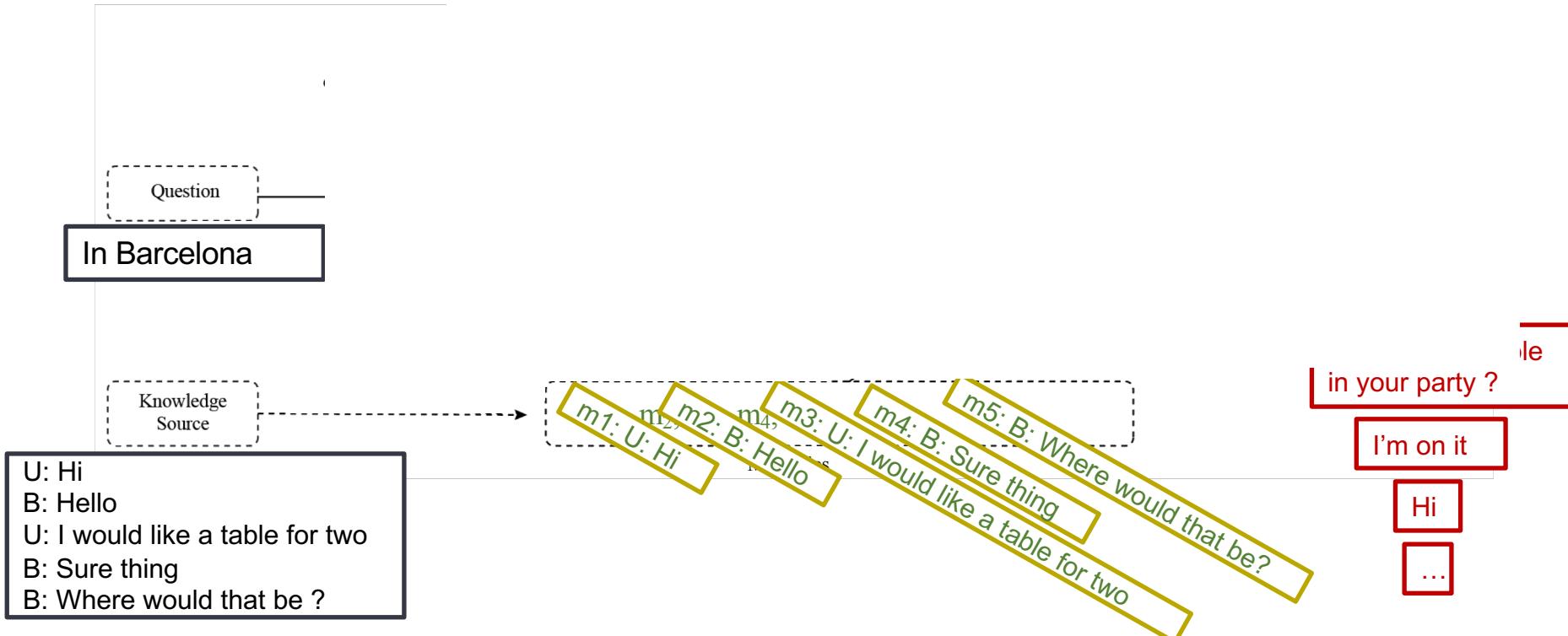
Task completed  
=  
All utterances  
correctly predicted



# Stats

	Tasks	T1	T2	T3	T4	T5	T6	Concierge
DIALOGS <i>Average statistics</i>	Number of utterances:	12	17	43	15	55	54	8
	- user utterances	5	7	7	4	13	6	4
	- bot utterances	7	10	10	4	18	8	4
	- outputs from API calls	0	0	23	7	24	40	0
DATASETS <i>Tasks 1-5 share the same data source</i>	Vocabulary size			3,747		1,229		8,629
	Candidate set size			4,212		2,406		11,482
	Training dialogs			1,000		1,618		3,249
	Validation dialogs			1,000		500		403
	Test dialogs			1,000 <sup>(*)</sup>		1,117		402

# End-to-end Model: Memory Networks



# Dashboard

	IR (TF-IDF)	LSTMs	Embeddings	Memory Networks	Top perf
T1 – API calls	5.6	/	100	99.9	100
T2 – Update API calls	3.4	/	68.4		100
T3 – Show options	8.0	/	64.9	74.9	100
T4 – Extra information	0.5	/	57.2	52.5	100
T5 – Full dialogs	4.6	/	75.4	50.1	100
T6 – DSTC2	1.6	/	22.6	41.1	?
T7 – Concierge	1.1	/	14.6	16.7	?

All datasets agree

Memory Networks can  
not learn to use the KB

# Dashboard

	IR (TF-IDF)	LSTMs	Embeddings	Memory Networks	Top perf
T1 – API calls	0	/	100	99.6	100
T2 – Update API calls	0	/	0	100	100
T3 – Show options	0	/	0	2.0	100
T4 – Extra information	0	/	0	3.0	100
T5 – Full dialogs	0	/	0	49.4	100
T6 – DSTC2	0	/	0	0	?
T7 – Concierge	0.2	/	0.5	1.2	?

Task completion is a  
much harder metric!

# Dashboard

	IR (TF-IDF)	LSTMs	Embeddings	Memory Networks	Top perf
T1 – API calls	5.6	/	100	99.9	100
T2 – Update API calls	3.4	/	68.4	100	100
T3 – Show options	8.0	/	64.9	74.9	100
T4 – Extra information	9.5	/	57.2	59.5	100
T5 – Full dialogs	4.6	/	75.4	96.1	100
T6 – DSTC2	1.6	/	22.6	41.1	?
T7 – Concierge	1.1	/	14.6	16.7	?

What if Out-of-  
Vocabulary entities?

# Dashboard

	IR (TF-IDF)	LSTMs	Embeddings	Memory Networks	Top perf
T1 – API calls	5.8 <b>+0.2</b>	/	60.0 <b>-40.0</b>	<b>72.3 -27.6</b>	100
T2 – Update API calls	3.5 <b>+0.1</b>	/	68.3 <b>-0.1</b>	<b>78.9 -21.1</b>	100
T3 – Show options	8.3 <b>+0.3</b>	/	65.0 <b>+0.1</b>	<b>74.4 -0.5</b>	100
T4 – Extra information	9.8 <b>+0.3</b>	/	57.0 <b>-0.2</b>	<b>57.6 -1.9</b>	100
T5 – Full dialogs	4.6 <b>+0.0</b>	/	48.2 <b>-27.2</b>	<b>65.5 -30.6</b>	100
T6 – DSTC2	1.6	/	22.6	<b>41.1</b>	/
T7 – Concierge	1.1	/	14.6	<b>16.7</b>	/

What if Out-of-Vocabulary entities?

OOV are not obvious for embedding-based models.

# Dashboard

	IR (TF-IDF)	LSTMs	Embeddings	Memory Networks + type features	Top perf
T1 – API calls	5.8 <b>+0.2</b>	/	60.0 <b>-40.0</b>	<b>96.5</b>	100
T2 – Update API calls	3.5 <b>+0.1</b>	/	68.3 <b>-0.1</b>	<b>94.5</b>	100
T3 – Show options	8.3 <b>+0.3</b>	/	65.0 <b>+0.1</b>	<b>75.2</b>	100
T4 – Extra information	9.8 <b>+0.3</b>	/	57.0 <b>-0.2</b>	<b>100</b>	100
T5 – Full dialogs	4.6 <b>+0.0</b>	/	48.2 <b>-27.2</b>	<b>77.7</b>	100
T6 – DSTC2	1.6	/	22.6	<b>41.0</b>	/
T7 – Concierge	1.1	/	14.6	<b>16.7</b>	/

What if Out-of-Vocabulary entities?

Adding type features can help to fix some limitations

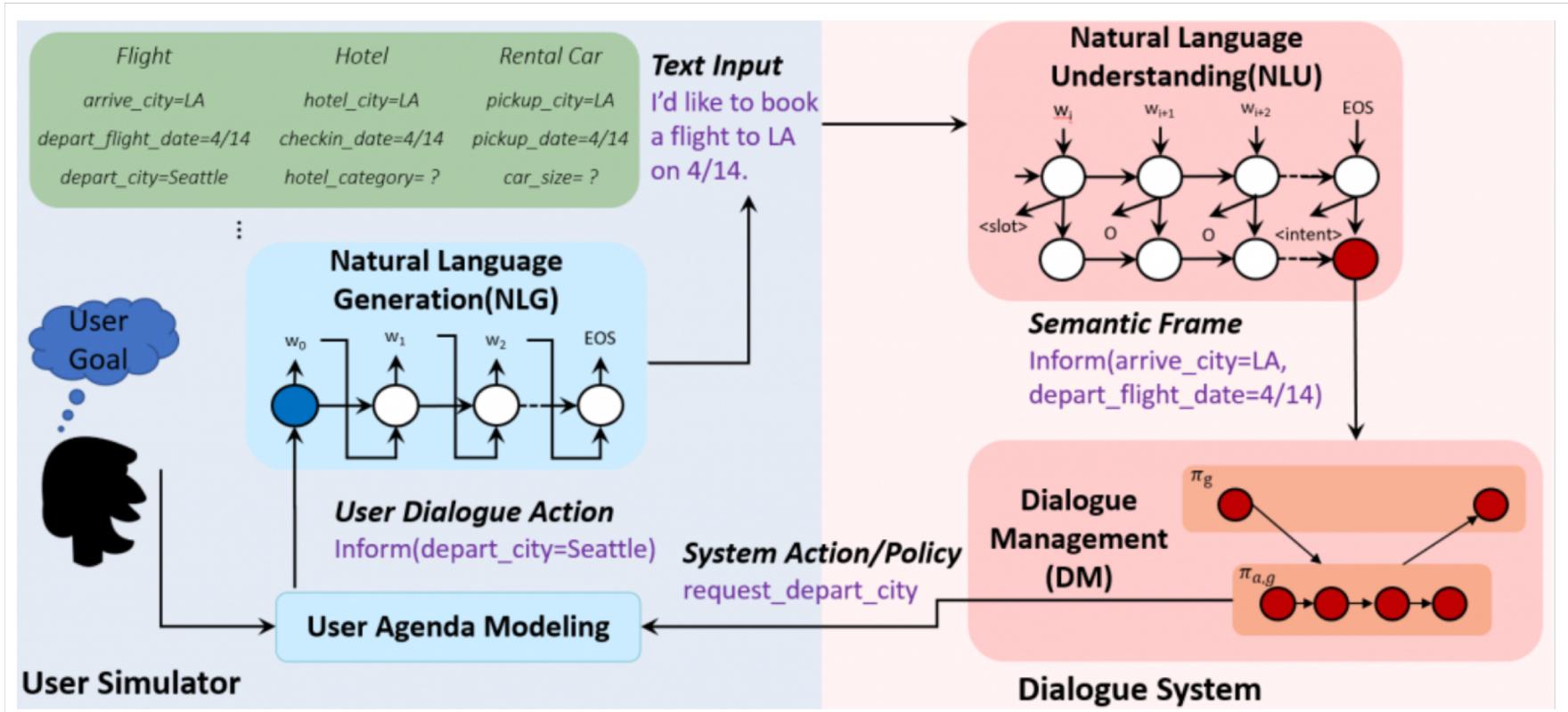
# API Call prediction

Time	Locutor	Dialog History	Hop #1	Hop #2	Hop #3
1	User	hi	.054	.016	.020
2	Bot	hello what can i help you with today	.040	.024	.008
3	User	may i have a table in paris	.099	.137	.141
4	Bot	i'm on it	.048	.028	.004
5	User	<silence>	.181	.099	.037
6	Bot	any preference on a type of cuisine	.056	.090	.014
7	User	i love indian food	.159	.188	.238
8	Bot	how many people would be in your party	.051	.022	.010
9	User	we will be six	.129	.144	.114
10	Bot	which price range are looking for	.039	.028	.006
11	User	in a moderate price range please	.128	.213	.405
12	Bot	ok let me look into some options for you	.016	.011	.003
<b>User input</b>		<silence>			
<b>Correct answer</b>		api_call indian paris six moderate			
<b>Predicted answer</b>		api_call indian paris six moderate	[Correct]		

# On real data?

Time	Locutor	Dialog History	Hop #1	Hop #2
1	User	hey concierge	.189	.095
2	User	could you check if i can get a rservation at <org> <date> for brunch	.209	.178
3	User	<number> people	.197	.142
4	User	<silence>	.187	.167
5	Bot	hi <person> unfortunately <org> is fully booked for <date> and there's <number> people on the waiting list	.225	.410
<b>User input</b>		when's the earliest availability		
<b>Correct answer</b>		i'll check		
<b>Pred. answer #1</b>		i'm on it	[Incorrect]	
<b>Pred. answer #2</b>		i'll find out	[Incorrect]	
<b>Pred. answer #3</b>		i'll take a look	[Incorrect]	
<b>Pred. answer #4</b>		i'll check	<b>[Correct]</b>	
<b>Pred. answer #5</b>		i'll check into it	[Incorrect]	

# Deep RL for Goal-oriented Dialog



# Dialog / Chatbots

---

# Eliza – A rule-base chatbot

Weizenbaum, 1966

- Goal: Mimic the responses of a non-directional psychotherapist
- Approach: Pattern matching and substitution rules and a memory/storage mechanism

YOU: I have trouble sleeping  
ELIZA: Can you elaborate on that ?  
YOU: when I go to bed, sleep does not come  
ELIZA: You say when you go to bed ?  
YOU: yes  
ELIZA: I see.  
YOU: can you help  
ELIZA: Oh, I help ?  
YOU: yes  
ELIZA: I understand.  
YOU: no!  
ELIZA: Are you saying no just to be negative?  
YOU: not really  
ELIZA: Please go on.  
YOU: I'd like help to fall asleep easily  
ELIZA: Can you elaborate on that ?

# Datasets for chit-chat

- [Twitter](#)  
2.6 million query-response pairs.
- [OpenSubtitles](#)  
>2.5 billion sentences  
not speaker aligned
- [SubTle](#)  
6.7 million utterances in 3.35 dialogs  
speaker aligned
- [Ubuntu Dialogue Corpus](#)  
930k dialogs with average 7.71 turns each
- [PersonaChat](#) – humans were given personas and asked to converse  
164,356 utterances over 10,981 dialogs

“A Survey of Available Corpora for Building Data-Driven Dialogue Systems” (Serban et al. 2017)



# ParlAI

: A platform for training and evaluating dialog agents on a variety of openly available datasets.

Its goal is to provide the community:

- a unified framework for training and testing dialog models
- a repository of both learning agents and tasks, use both to iterate research!
- seamless integration of Amazon Mechanical Turk for data collection and human evaluation

Over 20 tasks are supported, including popular datasets such as:

[SQuAD](#), [MCTest](#), [WikiQA](#), [WebQuestions](#), [SimpleQuestions](#), [WikiMovies](#), [QACNN](#) & [QADailyMail](#), [CBT](#), [BookTest](#), [bAbI tasks](#), [bAbI Dialog tasks](#), [Ubuntu Dialog](#), [OpenSubtitles](#), [Cornell Movie](#), [VQA](#), [VisDial](#) & [CLEVR](#). *Check it out: <http://parl.ai>*

+ Kurt Shuster, Emily Dinan, Jack Urbanek

Alexander H. Miller, Will Feng, Adam Fisch, Jiasen Lu, Dhruv Batra, Antoine Bordes, Devi Parikh, Jason Weston

# PersonaChat

Zhang et al., ACL'18

Persona 1	Persona 2
I like to ski	I am an artist
My wife does not like me anymore	I have four children
I have went to Mexico 4 times this year	I recently got a cat
I hate Mexican food	I enjoy walking for exercise
I like to eat cheetos	I love watching Game of Thrones

[PERSON 1:] Hi

[PERSON 2:] Hello ! How are you today ?

[PERSON 1:] I am good thank you , how are you.

[PERSON 2:] Great, thanks ! My children and I were just about to watch Game of Thrones.

[PERSON 1:] Nice ! How old are your children?

[PERSON 2:] I have four that range in age from 10 to 21. You?

[PERSON 1:] I do not have children at the moment.

[PERSON 2:] That just means you get to keep all the popcorn for yourself.

[PERSON 1:] And Cheetos at the moment!

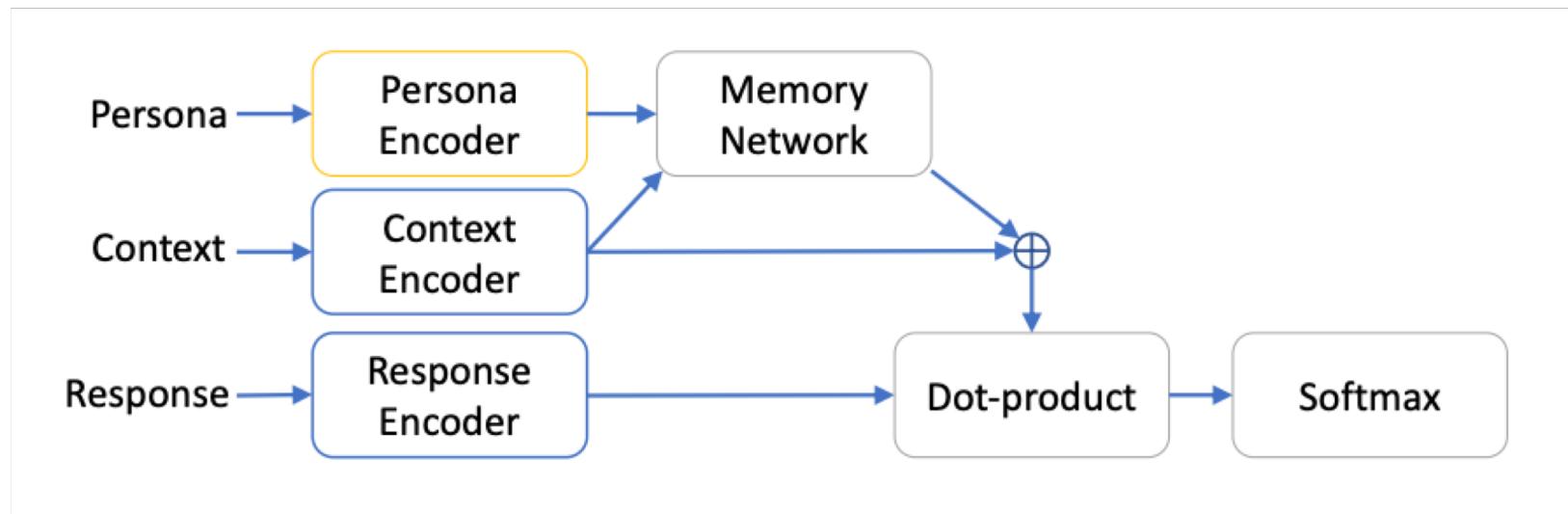
[PERSON 2:] Good choice. Do you watch Game of Thrones?

[PERSON 1:] No, I do not have much time for TV.

[PERSON 2:] I usually spend my time painting: but, I love the show.

# A retrieval model

Mazare et al., EMNLP'18



# A generative model

Zhang et al., ACL'18

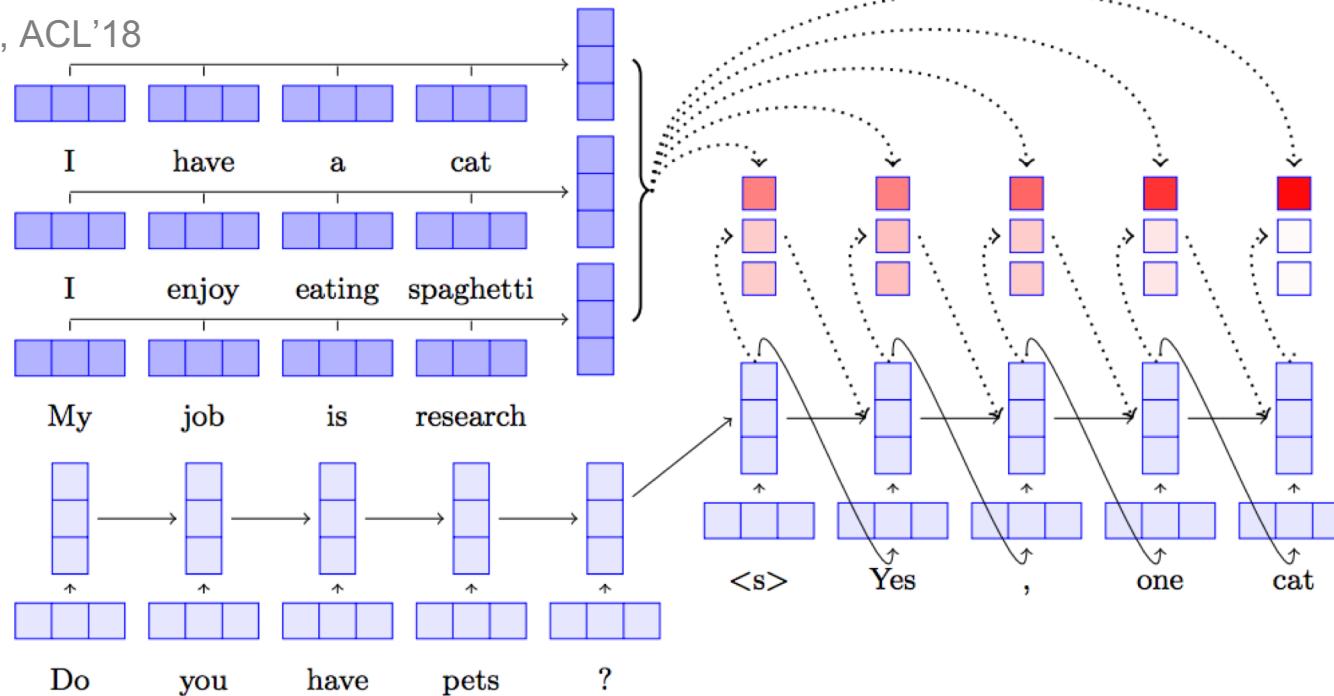


Figure 1: A diagram of the Profile Memory Network for generation. We also implemented a ranking version which has the same architecture except it ranks candidate sentences from the training set instead of generating, representing them using bag-of-word embeddings.

# Automatic evaluation

Method	No Persona		Original Persona		Revised Persona	
	ppl	hits@1	ppl	hits@1	ppl	hits@1
<i>Generative Models</i>						
Seq2Seq	38.08	0.092	40.53	0.084	40.65	0.082
Profile Memory	38.08	0.092	34.54	0.125	38.21	0.108
<i>Ranking Models</i>						
IR baseline	-	0.214	-	0.410	-	0.207
Starspace	-	0.318	-	0.491	-	0.322
Profile Memory	-	0.318	-	0.509	-	0.354
KV Profile Memory	-	0.349	-	0.511	-	0.351

# Human Evaluation

We asked the human to evaluate their partner after the dialogue (we didn't tell them if it's a human or a bot).

*NOTE: we also tried asking humans to evaluate fixed dialogues they didn't take part in, but the scores were squashed towards the mean, Turkers were not engaged. So we abandoned this approach.*

The character below will be assigned to you, please read them carefully:

Welcome! Below is your persona (you can find it on the left side of the chat)

When you are ready to start your conversation, click the "I am ready, continue" button below

i have prior agriculture experience.

i am a good swimmer.

i prefer staying in as opposed to going to the store.

i am a mom.

i enjoy books.

I am ready, continue

**SYSTEM:** Now the conversation is completed!

Please evaluate the other person's **fluency** during this conversation by **entering a score from [1, 2, 3, 4, 5]** below, fluency reflects whether the other people's words are accurate, and whether you can read it quickly and with ease. (1 means "not fluent at all" and 5 means "extremely fluent", e.g., You can enter 3 for an OK fluency)

PERSON\_1: 3

**SYSTEM:** Now please evaluate the other people's **engagingness DISREGARDING the fluency** during this conversation by **entering a score from [1, 2, 3, 4, 5]** below: (1 means "not engaging at all" and 5 means "extremely engaging", e.g., You can enter 3 for an OK dialog)

# Human evaluations

Model	Profile	Method				Persona Detection
		Fluency	Engagingness	Consistency		
Human	Self	4.31(1.07)	4.25(1.06)	4.36(0.92)	0.95(0.22)	
<i>Generative PersonaChat Models</i>						
Seq2Seq	None	3.17(1.10)	3.18(1.41)	2.98(1.45)	0.51(0.50)	
Profile Memory	Self	3.08(1.40)	3.13(1.39)	3.14(1.26)	0.72(0.45)	
<i>Ranking PersonaChat Models</i>						
KV Memory	None	3.81(1.14)	3.88(0.98)	3.36(1.37)	0.59(0.49)	
KV Profile Memory	Self	3.97(0.94)	3.50(1.17)	3.44(1.30)	0.81(0.39)	
Twitter LM	None	3.21(1.54)	1.75(1.04)	1.95(1.22)	0.57(0.50)	
OpenSubtitles 2018 LM	None	2.85(1.46)	2.13(1.07)	2.15(1.08)	0.35(0.48)	
OpenSubtitles 2009 LM	None	2.25(1.37)	2.12(1.33)	1.96(1.22)	0.38(0.49)	
OpenSubtitles 2009 KV Memory	None	2.14(1.20)	2.22(1.22)	2.06(1.29)	0.42(0.49)	

Persona 1	Persona 2
I am afraid of dogs	I am a carpenter
I used to work at a carnival	I used to drink but I stopped five years ago
I like to drink scotch to relax	I have a wife and three kids
I write short stories for a living	I like playing poker
I like dark superhero movies	I dont have many friends

[PERSON 1:] Hi are you relaxing on this fine evening?

[PERSON 2:] I am doing good and you?

[PERSON 1:] I am well, just drinking some scotch and working on my writing

[PERSON 2:] I love to cook and bake but I rarely make time for it

[PERSON 1:] what do you usually bake?

[PERSON 2:] I do not cook, I bake cakes at a local restaurant. how about you?

[PERSON 1:] I used to be a juggler at a carnival but now I am an author

[PERSON 2:] I am a dancer, what do you do?

[PERSON 1:] I just told you. Do you have any pets?

[PERSON 2:] no, I am allergic to pets unfortunately.

[PERSON 1:] I understand. I am not allergic but I am terrified of dogs!

[PERSON 2:] I have a cat. she has been with me through three wives, haha.

[PERSON 1:] Funny that you just told me you have no pets

[PERSON 2:] I do not have any animals.

Person\_1 is a human

Person\_2 is the bot  
(generative)

Table 7: Example dialog between a human (Person 1) and the Seq2Seq model (Person 2).

Persona 1	Persona 2	
I am engaged to be married I used to play sports in high school My favorite sport is soccer I am a college graduate I love watching sports on TV	I am a writer I write romance novels I live on a faraway island I am an introvert I have two cats	Person_1 is a human Person_2 is the bot (retrieval)

[PERSON 1:] I just got done playing soccer. What have you been up to?  
 [PERSON 2:] I do not have pets. I spend my free time writing. what about you?  
 [PERSON 1:] Just graduated from college. Looking for work.  
 [PERSON 2:] yes I graduated, in writing and poetry  
 [PERSON 1:] Have you had anything published?  
 [PERSON 2:] I mainly write fiction novels.  
 [PERSON 1:] Do you like to watch sports?  
 [PERSON 2:] do you like kings of leon my favorite by them is use somebody  
 [PERSON 1:] Are you married? I will be married soon.  
 [PERSON 2:] haha, no time. I have got a novel to finish.  
 [PERSON 1:] What is your novel about?  
 [PERSON 2:] it is a love story. I really enjoy writing. do you write or read?  
 [PERSON 1:] Like to read. Spend most time watching football.  
 [PERSON 2:] yes I do I am and author for a living

Table 9: Example dialog between a human (Person 1) and the Key-Value Profile Memory Network with Self Persona.

# Dialog is about interaction – can we use it?

Hancock et al., 2019

The screenshot shows a messaging interface with a robot icon and a person icon at the top. The conversation is as follows:

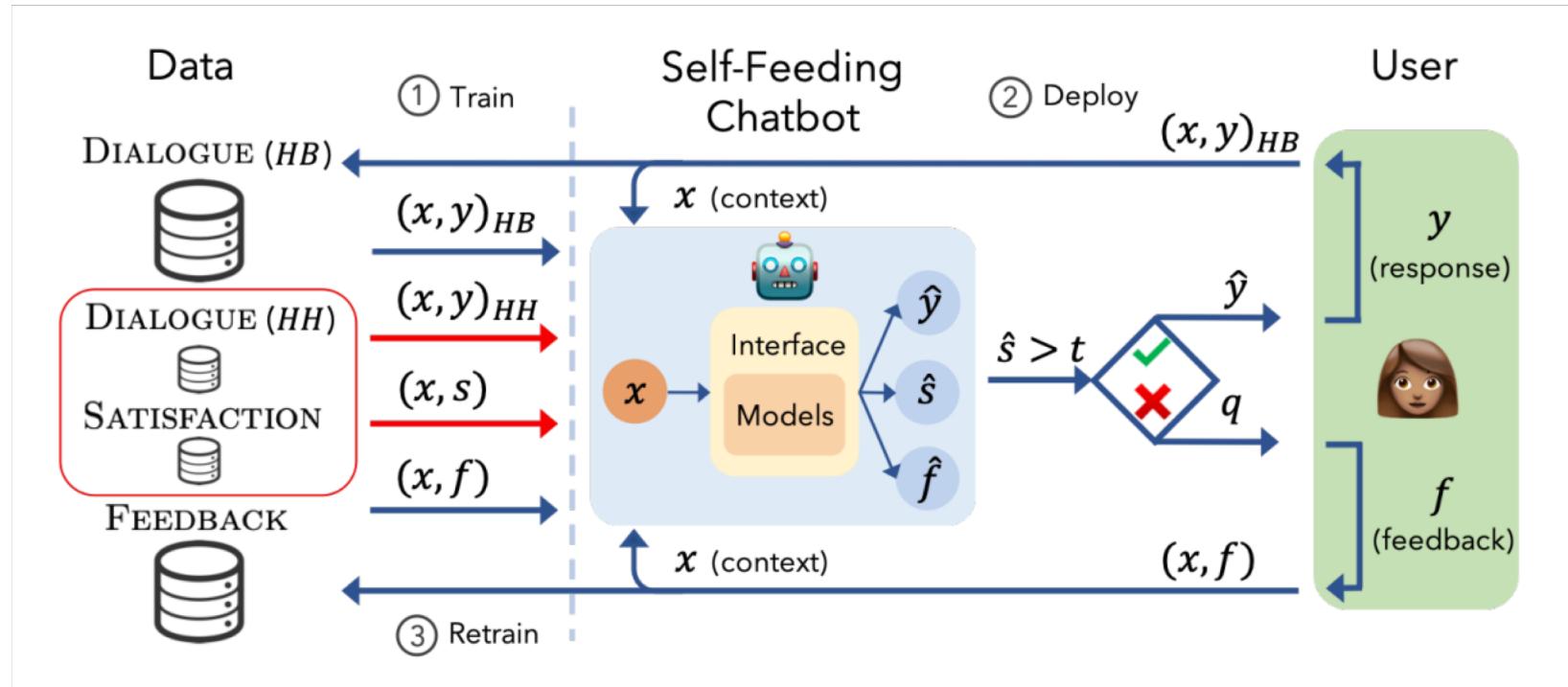
- A** Have you been to France?  
Satisfaction: 0.85
- B** Yes I have! It's beautiful.
- C** Lol. I never eat them!  
Satisfaction: 0.10
- D** What are you talking about?
- E** Oops! I messed up.  
What should I have said?
- F** Maybe ask me what I thought about French food?

Below the interface is a table titled "Extracted Training Examples".

Extracted Training Examples	
DIALOGUE	FEEDBACK
Context <b>A</b> Have you been to France?	Context <b>A</b> Have you been to France? <b>B</b> Yes, I have! It's beautiful.
Response <b>B</b> Yes, I have! It's beautiful.	Feedback <b>F</b> Maybe ask me what I thought about French food?

# A self-feeding chatbot

Hancock et al., 2019



# Open topics

- Active learning from interactions
  - Reinforcement learning from interactions
  - Mixing information-retrieval with generative models
  - Multimodal (conversing about an image)
  - Open-domain: talking about anything, mixing goal-oriented and chit-chat.
- 
- Need interactions at scale!!

# Joint testing on different tasks?

- Dialog systems should be open-domain!
- Dialog is not monolithic:
  - Ask facts?
  - Ask for opinions (recommendations)?
  - Dialog combining facts and opinions?
  - General chit-chat (statements not questions)?
- Combination of all above in one end-to-end model?

# Multi-task dialog dataset(s)

Dodge et al., ICLR'16

## (Dialog 1) QA: facts about movies

Sample input contexts and target replies (in red) from Dialog Task 1:

What movies are about open source? [Revolution OS](#)  
[Ruggiero Raimondi](#) appears in which movies? [Carmen](#)  
What movies did Darren [McGavin](#) star in? [Billy Madison](#), [The Night Stalker](#), [Mrs. Pollifax-Spy](#), [The Challenge](#)  
Can you name a film directed by Stuart Ortiz? [Grave Encounters](#)  
Who directed the film White Elephant? [Pablo Trapero](#)  
What is the genre of the film Dial M for Murder? [Thriller](#), [Crime](#)  
What language is [Whity](#) in? [German](#)

## (Dialog 2) Recs: movie recommendations

Sample input contexts and target replies (in red) from Dialog Task 2:

Schindler's List, The Fugitive, Apocalypse Now, Pulp Fiction, and The Godfather are films I really liked. Can you suggest a film?  
[The Hunt for Red October](#)

Some movies I like are Heat, Kids, Fight Club, Shaun of the Dead, The Avengers, [Skyfall](#), and Jurassic Park. Can you suggest something else I might like? [Ocean's Eleven](#)

## (Dialog 3) QA+Recs: combination dialog

Sample input contexts and target replies (in red) from Dialog Task 3:

I loved [Billy Madison](#), [Blades of Glory](#), [Bio-Dome](#), [Clue](#), and [Happy Gilmore](#). I'm looking for a Music movie. [School of Rock](#)  
What else is that about? [Music](#), [Musical](#), [Jack Black](#), [school](#), [teacher](#), [Richard Linklater](#), [rock](#), [guitar](#)  
I like rock and roll movies more. Do you know anything else?  
[Little Richard](#)

## (Dialog 4) Reddit: real dialog

Sample input contexts and target replies (in red) from Dialog Task 4:

I think the Terminator movies really suck, I mean the first one was [kinda ok](#), but after that they got really cheesy. Even the second one which people somehow think is great. And after that... [forgeddabotit](#).

C'mon the second one was still pretty cool.. [Arnold](#) was still so badass, as was [Sarah Connor's character](#).. and the way they blended real action and effects was perhaps the last of its kind...

# 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, *Kočiský et al. TACL 2018*