Why?

Tutorial on Machine Reading at AthensNLP

Sebastian Riedel *UCL, FAIR // UK*
Johannes Welbl *UCL // UK*
Dirk Weissenborn *Google Brain // Germany*
With help from Antoine Bordes, Angela Fan (FAIR)

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

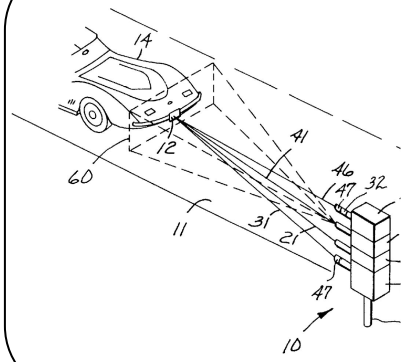
This Tutorial

- Context:
  - What is Machine Reading?
  - Why should we care?
- Methods:
  - What are the central paradigms in Machine Reading?
  - How are they implemented?
- Challenges:
  - Why is Machine Reading hard?
  - What are strengths and weaknesses of current approaches?
- Tools and Resources:
  - what datasets are important?
  - what tools are available?
- I will focus on a **broad high level** overview

What’s *Machine Reading*?

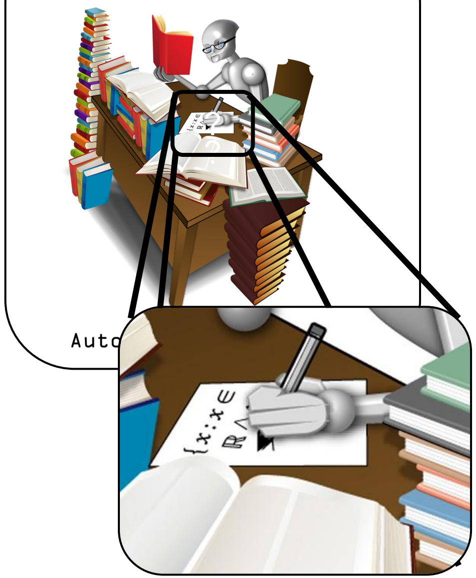
Don’t Anthropomorphize Computers,
They Hate it When You do That.

Something else
entirely!

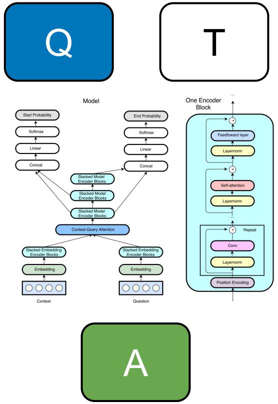


before 2006

Text to Symbolic
Representations



End-to-End
Question Answering

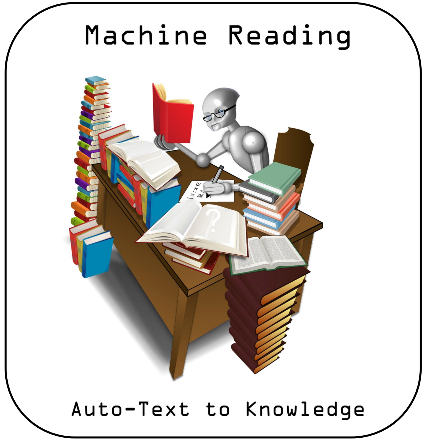


since 2014

Hermann et al., 2014

What’s this Tutorial about?

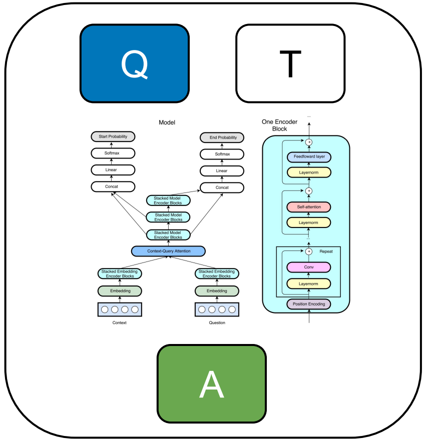
Text to Symbolic
Representations



2006-2014

Etzioni et. all,
DARPA etc.

End-to-End
Question Answering

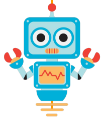


since 2014

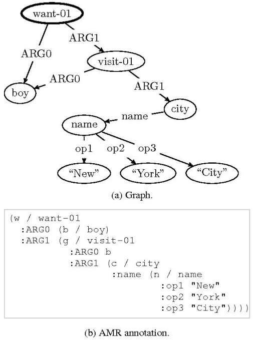
Hermann et al., 2014

Machine Reading

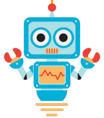
[Text]



converts into



[Meaning]



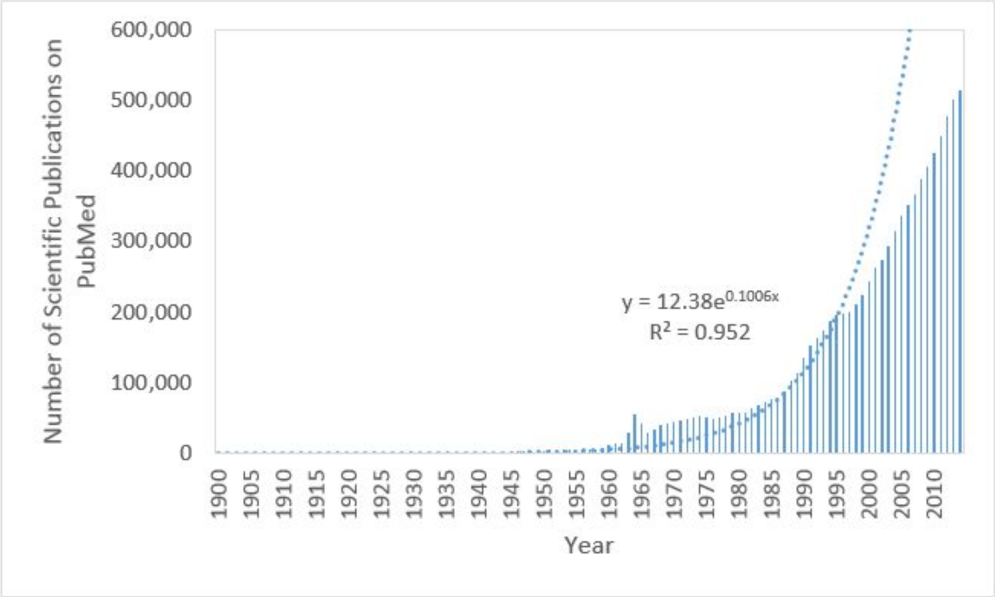
uses for

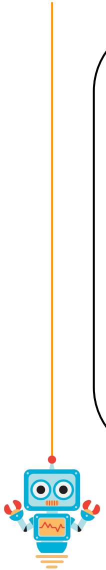
[Information Need]

What do we mean by Machine Reading?

A **machine** converts **text** into a representation of **meaning** that can satisfy (a broad set of) **information needs**

Motivation 1: Information Overload



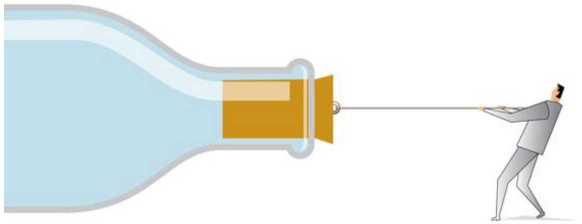


[Information Need]

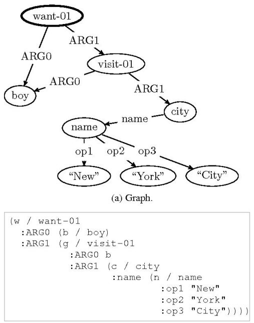
uses for

Motivation 2: The Knowledge Acquisition Bottleneck

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



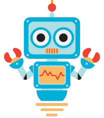
[Knowledge]



(a) Graph.

(b) AMR annotation.

[Meaning]



uses for

[Information Need]

Applications: Question Answering

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

?

What city did Tesla move to in 1880?

Prague

[Text]

[Meaning]

[Information Need]

Applications: Helping Agents to learn Faster

Branavan et al., 2012

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]

?

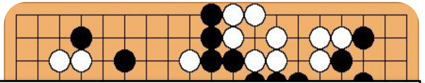
[Meaning]

[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 grou results in an increase of liberties ...

[Text]



Artificial Intelligence / Machine Learning

Instead of practicing, this AI mastered chess by reading about it

Machines that appreciate "brilliant" and "dumb" chess moves could learn to play the game—and do other things—more efficiently.

Applications: Support a Molecular Tumor Board

Poon et. al, 2017

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]

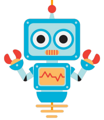
[Meaning]

[Information Need]

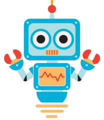
Machine Reading Approaches

?

[Text]



[Meaning]



[Information Need]

converts into

uses for

Semantic Parsing

Ewan forgot the mozarella in his car

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

[Text]

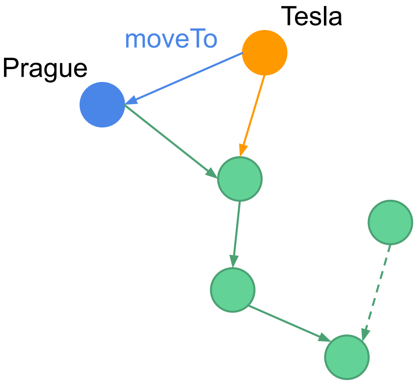
[Meaning]

[Information Need]

Automatic Knowledge Base Construction

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

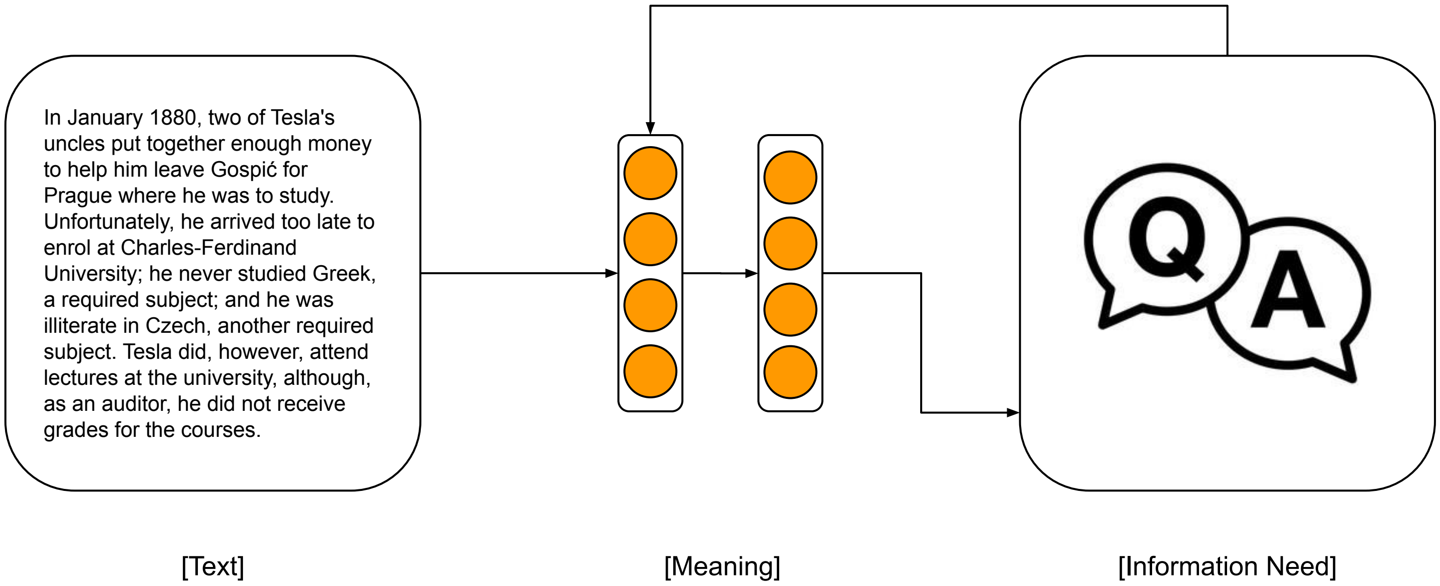
[Text]



[Meaning]

[Information Need]

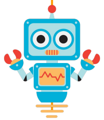
End-to-End Reading Comprehension



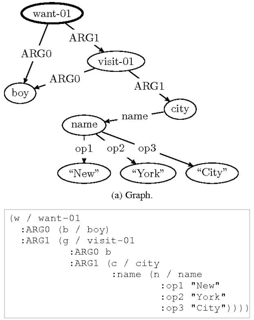
Where do we see you?

use machine reading

[Text]



converts into

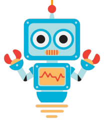


(a) Graph.

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

(b) AMR annotation.

[Meaning]



uses for

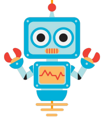
I am a representative member of the Athens NLP summer school community

[Information Need]

Where do we see you?

innovate for machine reading!

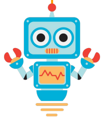
[Text]



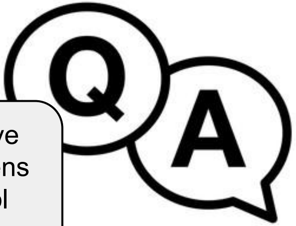
converts into

I am a representative member of the Athens NLP summer school community

[Meaning]



uses for



[Information Need]

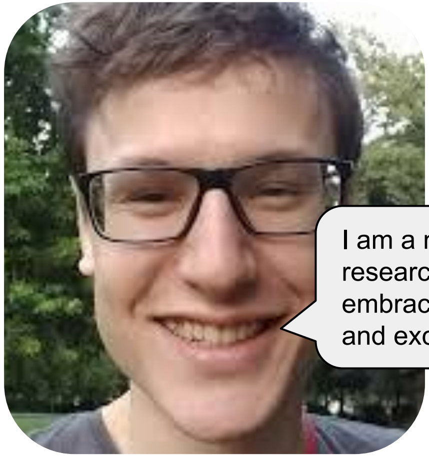
Structure

I am an old-fashioned Machine Reading researcher who just can't move on...

Text to Symbolic Representations

Part 1.1

End-to-End Question Answering



I am a new-school MR researcher who embraces the new and exciting

Part 1.2

Structure

I am an old-fashioned Machine Reading researcher who just can't move on...

Text to Symbolic Representations

Part 1.1

End-to-End Question Answering

I am a new-school MR researcher who embraces the new and exciting

Part 1.2

Structure

Text to Symbolic Representations

Part 1.1

End-to-End Question Answering

Part 1.2

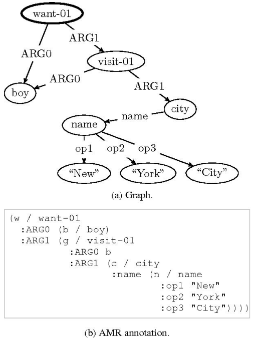
Current Trends Open Problems

Part 2

Text to Symbolic Representations

What do we need from a representation?

[Text]

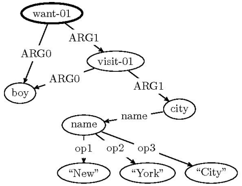


[Meaning]

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

What are the core challenges?

[Text]



(a) Graph.

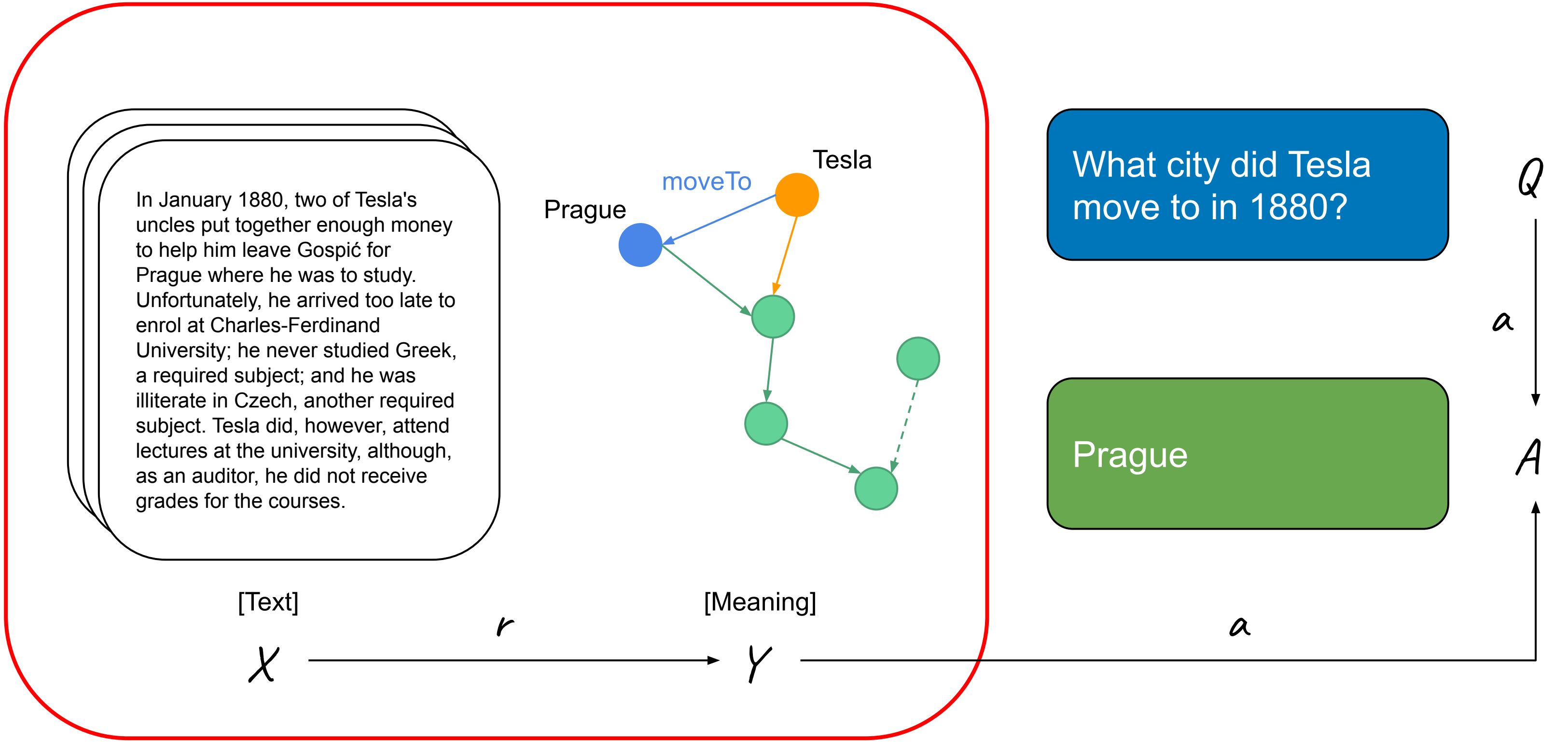
```
(w / want-01
  :ARG0 (b / boy)
  :ARG1 (g / visit-01
          :ARG0 b
          :ARG1 (c / city
                  :name (n / name
                          :op1 "New"
                          :op2 "York"
                          :op3 "City"))))
```

(b) AMR annotation.

[Meaning]

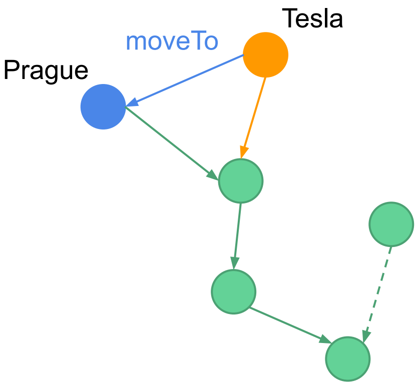
- Ambiguity
- Variation
- Coreference
- Common Sense
- Scale
- ...

Knowledge Graph Construction



Knowledge Graph Construction

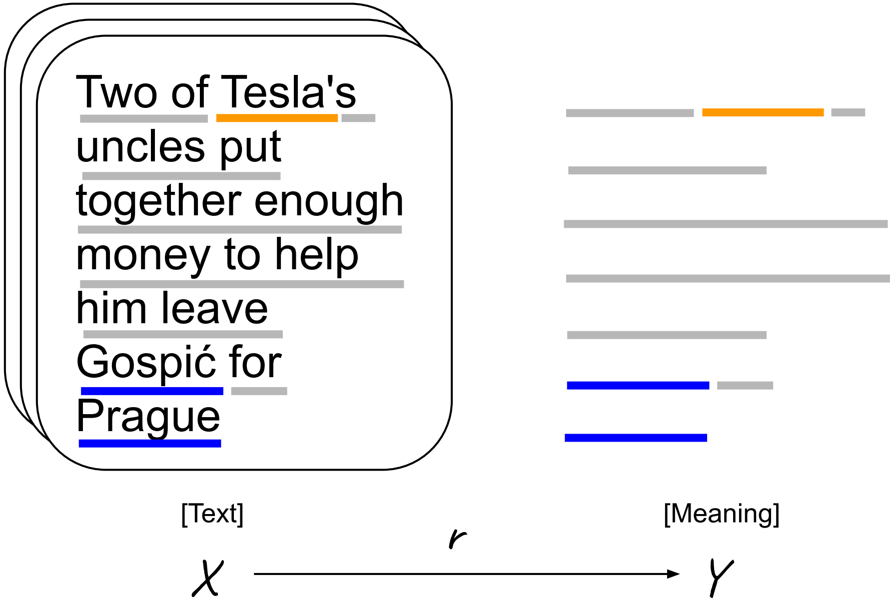
In January 1880, two of Tesla's uncles put together enough money to help him leave Gospić for Prague where he was to study. Unfortunately, he arrived too late to enrol at Charles-Ferdinand University; he never studied Greek, a required subject; and he was illiterate in Czech, another required subject.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



[Text] [Meaning]

$$X \xrightarrow{r} Y$$

Entity Extraction and Typing as Sequence Labelling

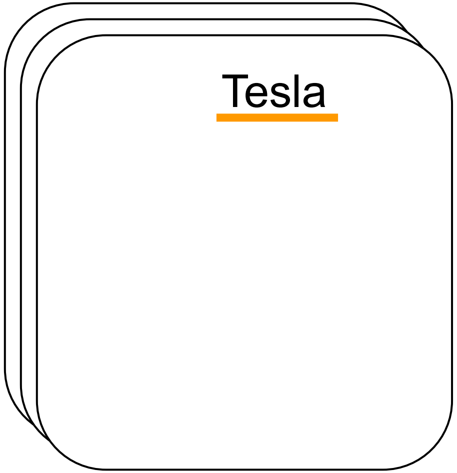


- Linear Chain CRF
- Bi-directional RNNs
- Hybrid RNN & CRFs

Person

Location

Challenge: Ambiguity



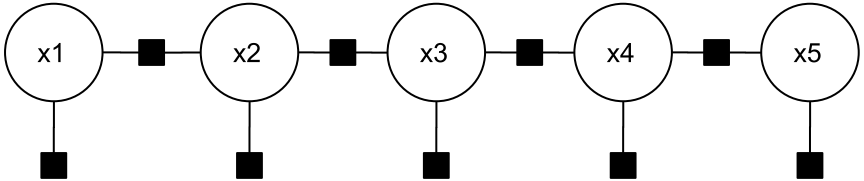
Person?

Brand?

Factor Graph Primer

- We will represent factorization of a probabilistic model using **factor graphs**

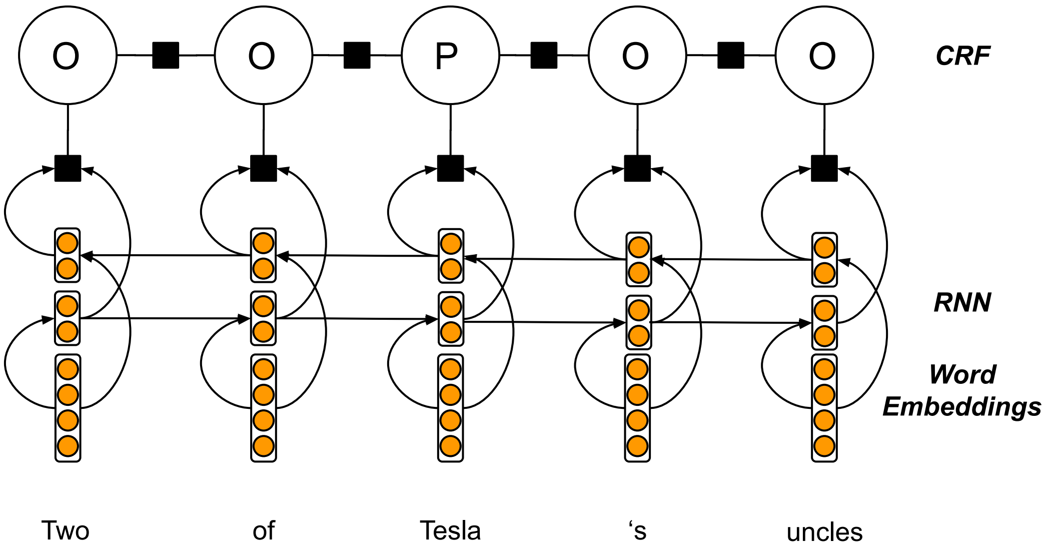
$$p(\mathbf{x})$$



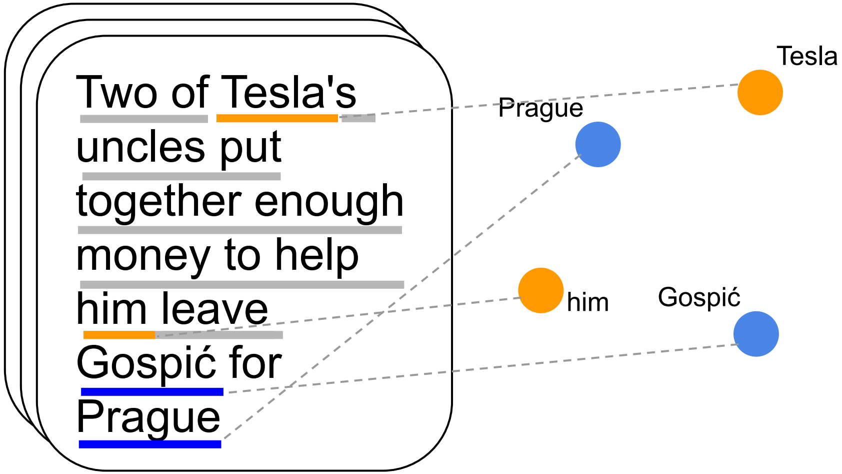
- Used for inference (“most likely assignment, marginal probabilities”)
- Loopy -> Inference Hard

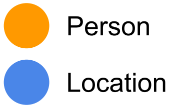
Conditional Random Fields with RNN Potentials

Huang et al., 2015

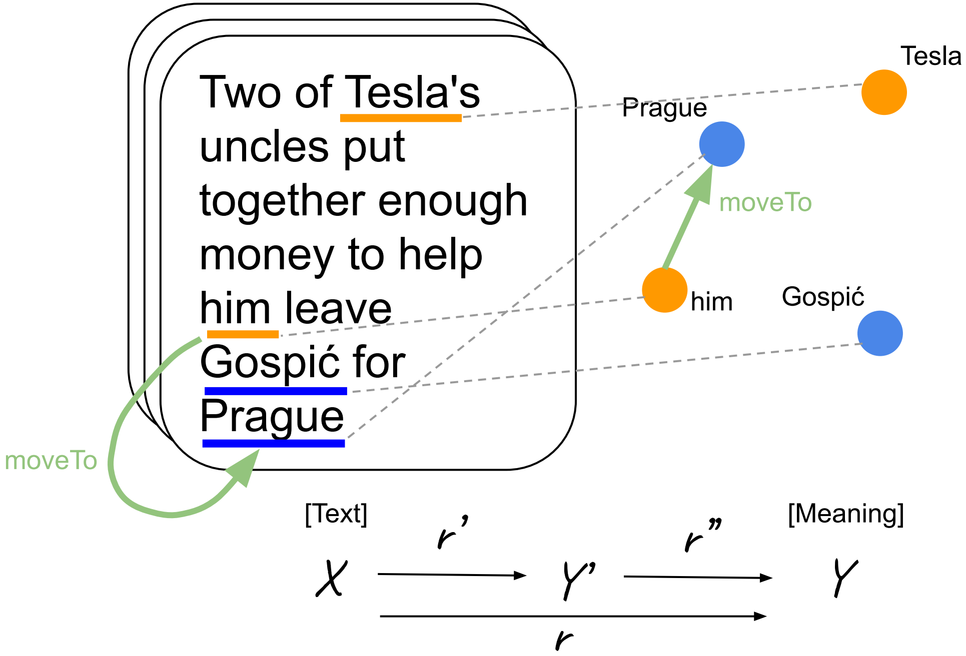


Instantiate Nodes





Relation Extraction



- Neural Classification
- Distant Supervision

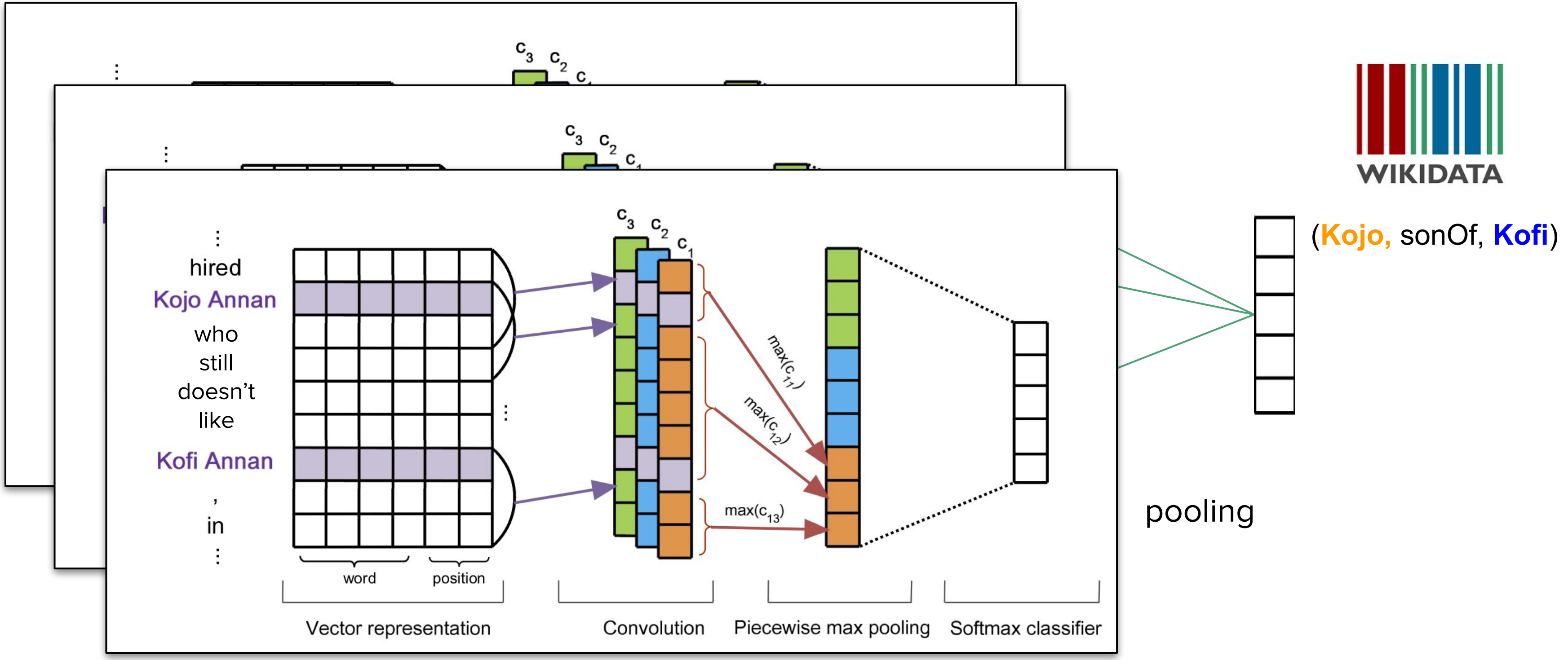
Challenge: Variation

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

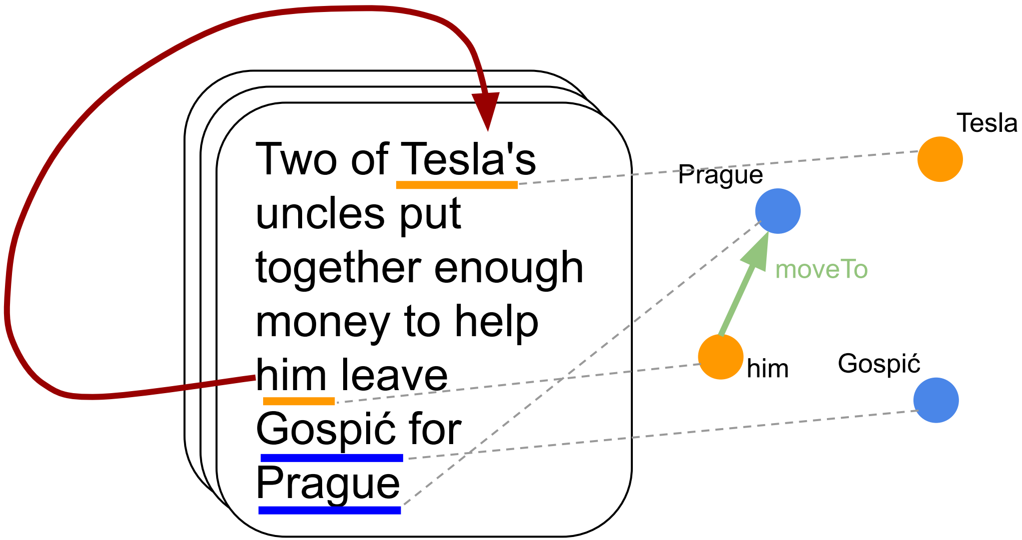
Two of Tesla's uncles put together enough money to help him move to Prague

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

Neural Relation Extraction and Distant Supervision (Zeng et al 2015)



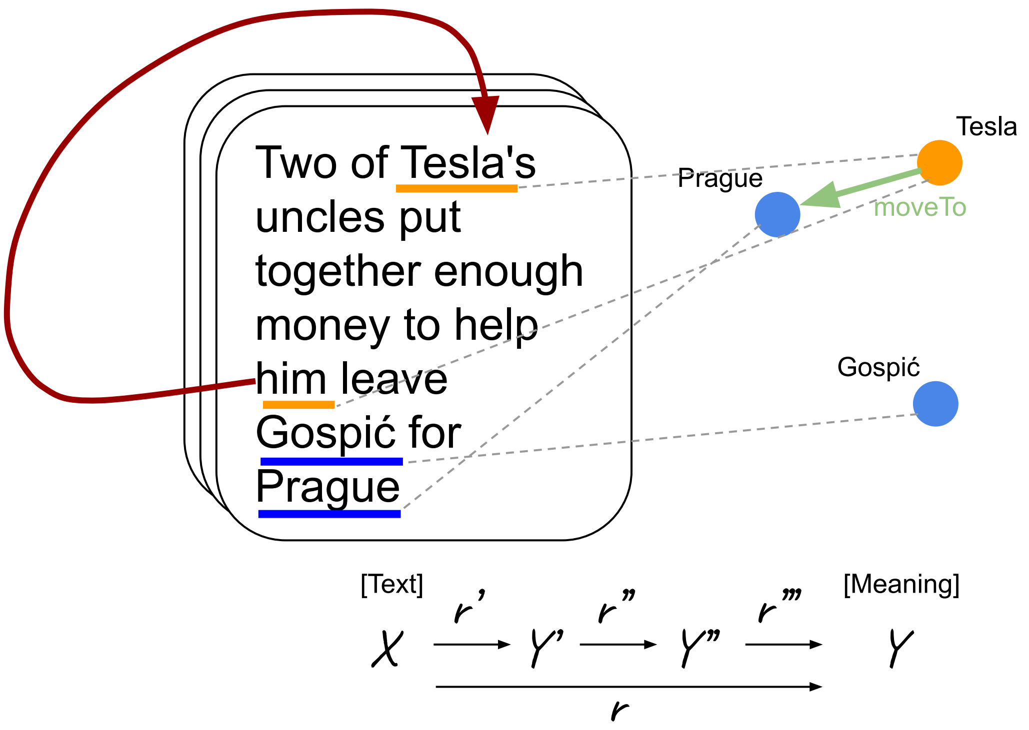
Coreference Resolution



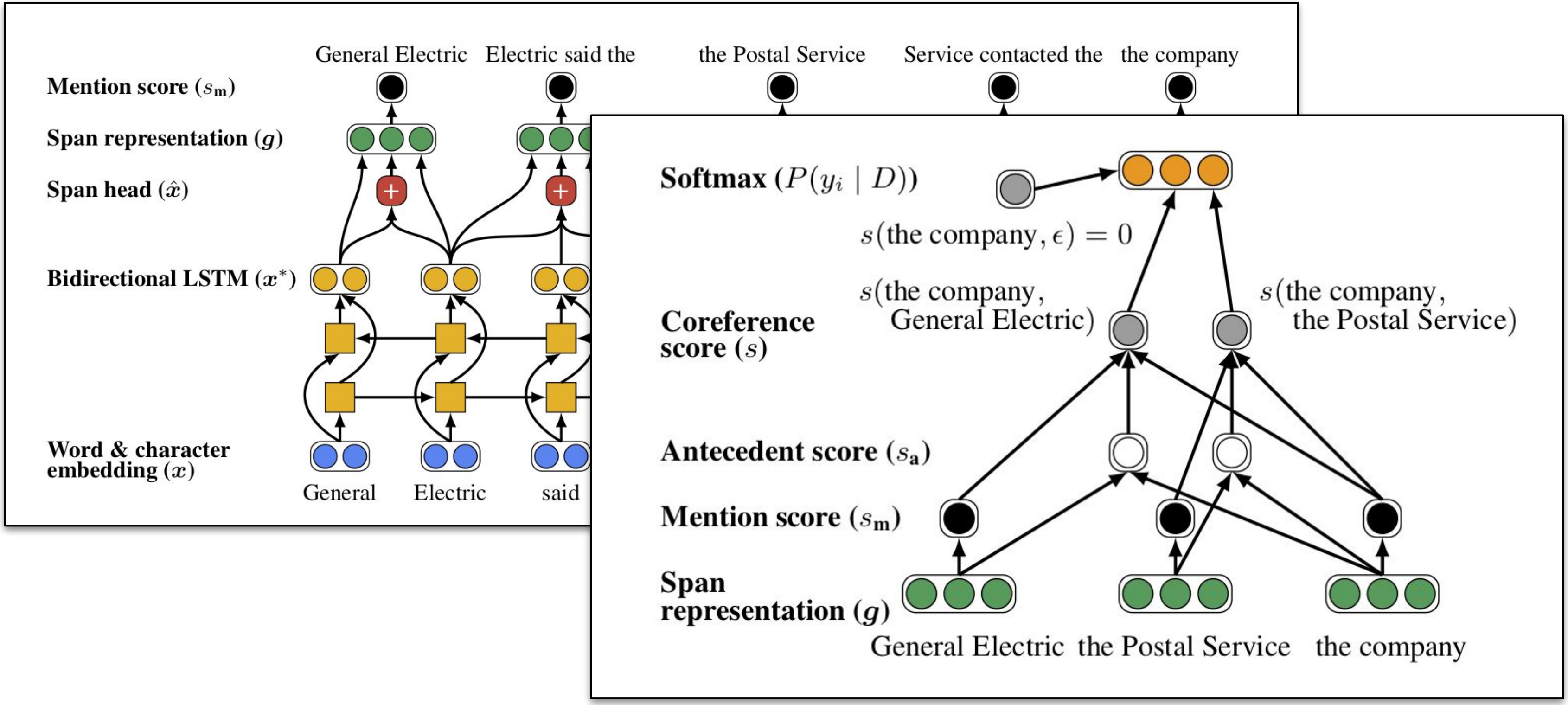
$$X \xrightarrow{r'} Y' \xrightarrow{r''} Y'' \xrightarrow{r'''} Y$$

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

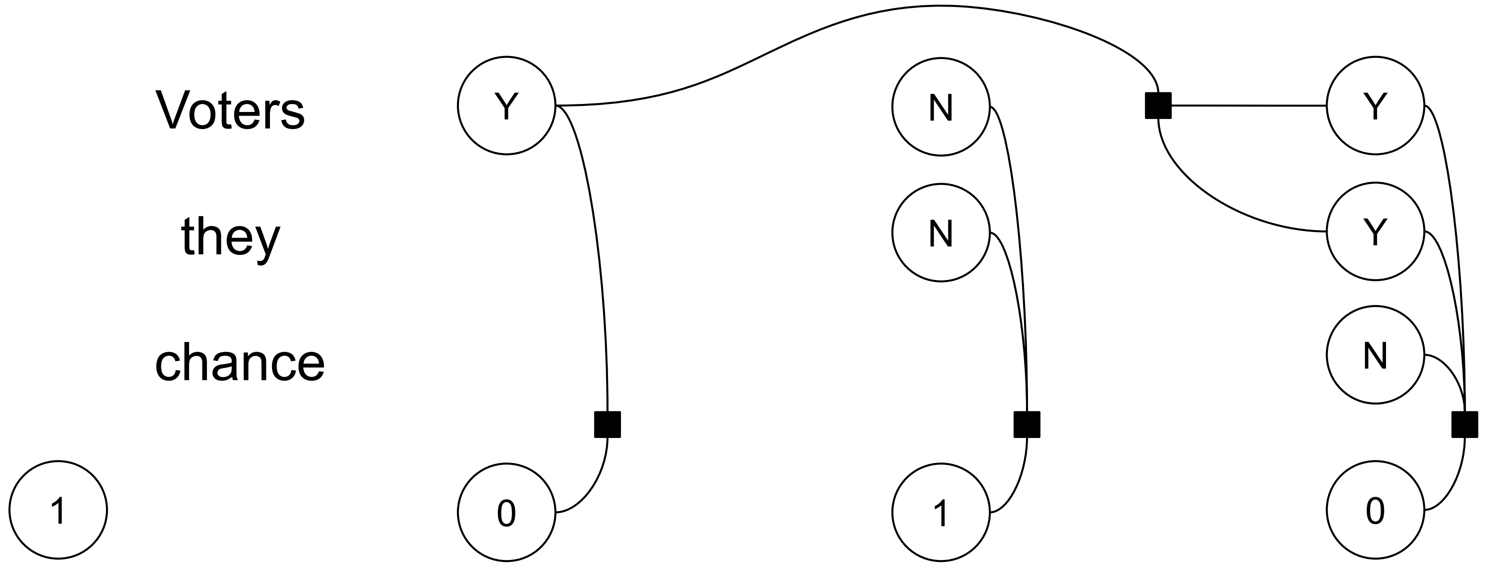
Collapsing Nodes



Lee et al, 2017



Coreference Resolution (Durrett and Klein 2013)



Voters agree when **they** are given a **chance** to decide if **they** ...

1 2 3 4

Challenge: Common Sense

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

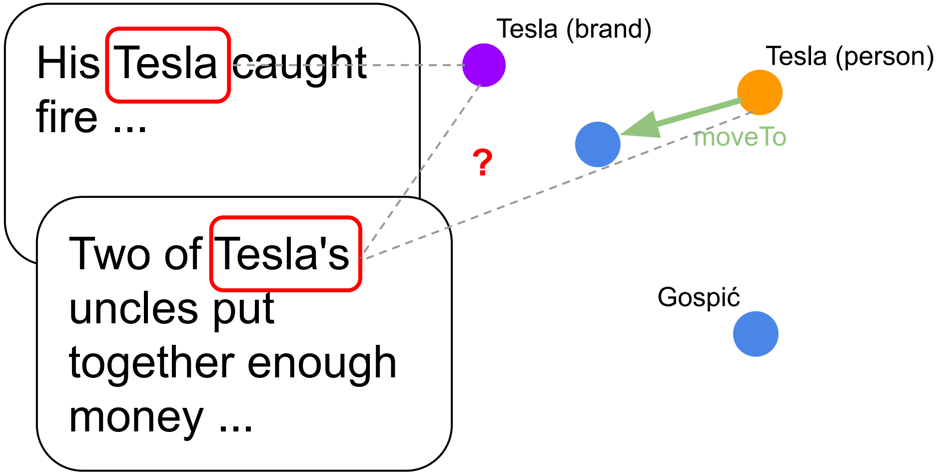
The **trophy** would not fit in the brown **suitcase** because **it** was too ***big***.

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

Surface

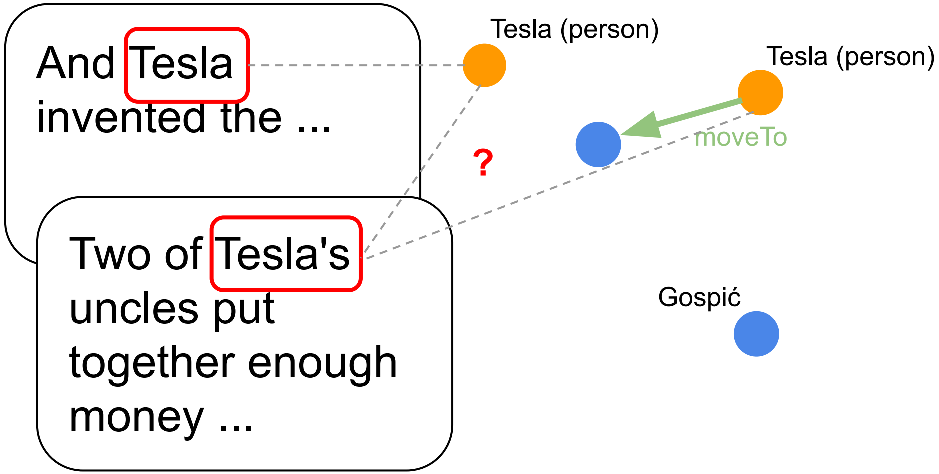
Common Sense

Entity Linking



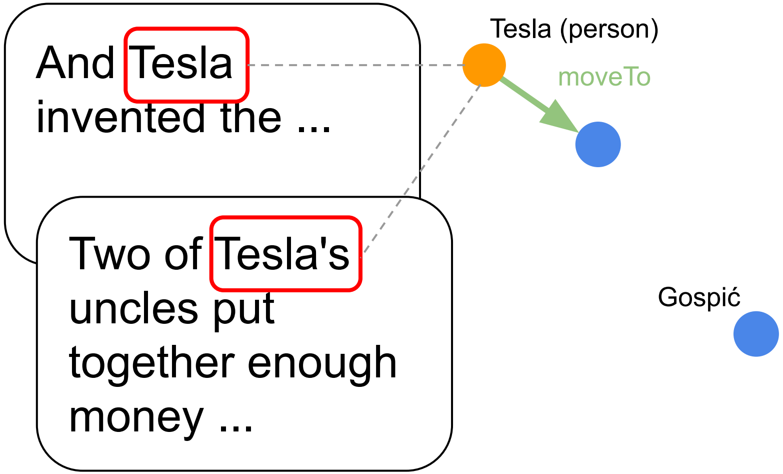
- Reranking ...
- Embeddings ...

Entity Linking



- Reranking ...
- Embeddings ...

Collapsing

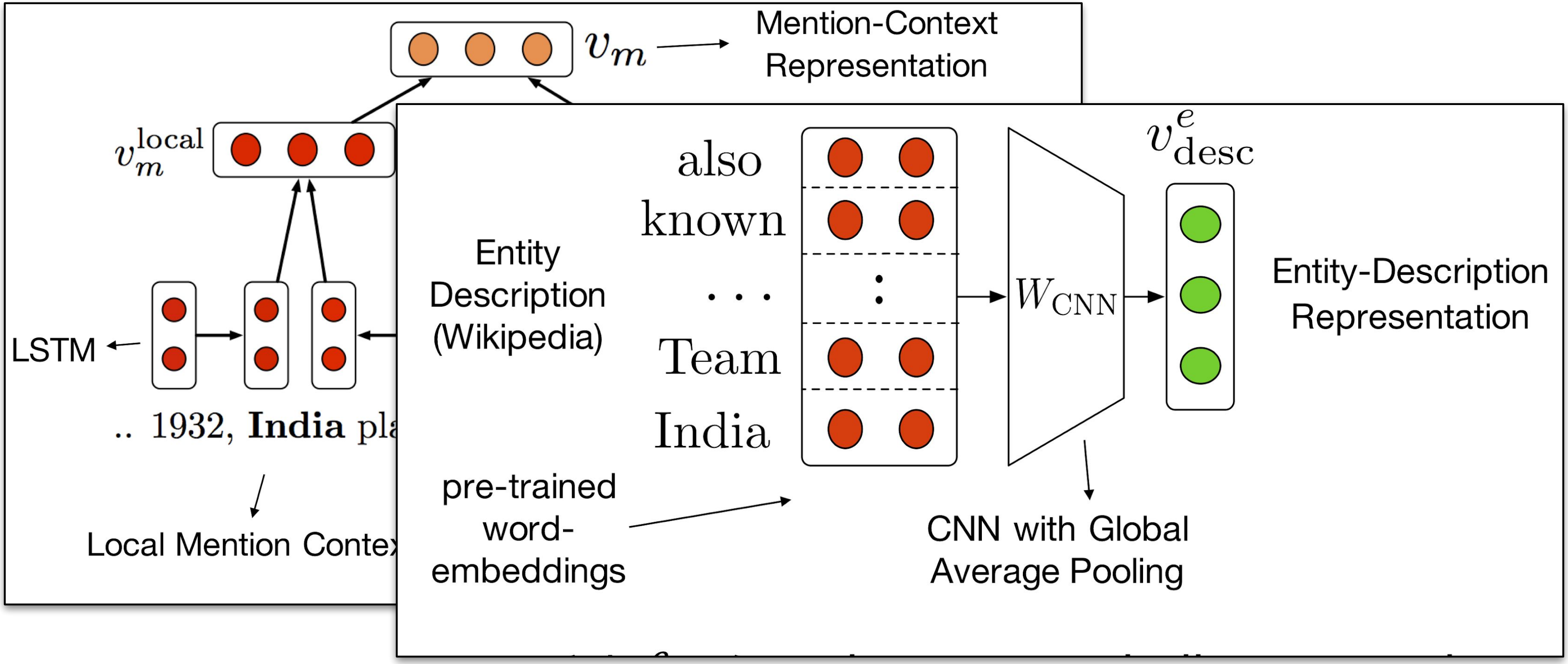


[Text] [Meaning]

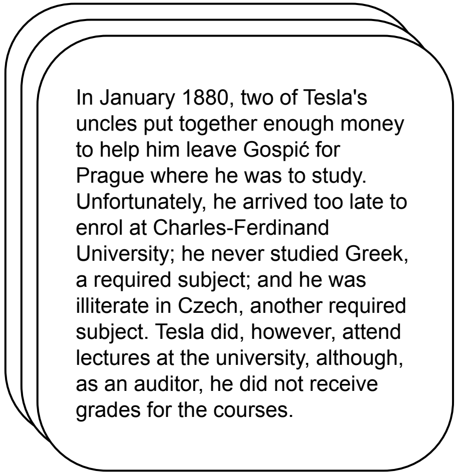
$$X \xrightarrow{r'} Y' \xrightarrow{r''} Y'' \xrightarrow{r'''} Y''' \xrightarrow{r''''} Y$$

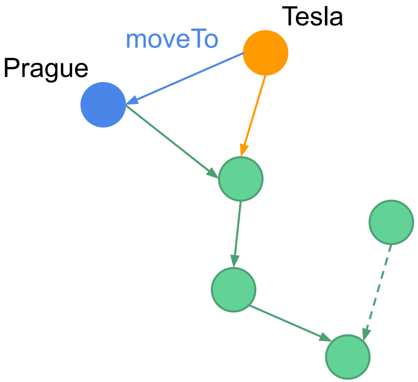
$$X \xrightarrow[r]{\qquad\qquad\qquad} Y$$

Entity Linking (Gupta et al. 2017)



Strengths of Symbolic Knowledge Representations

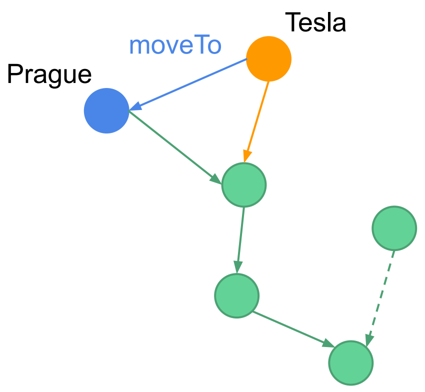

In January 1880, two of Tesla's uncles put together enough money to help him leave Gospić for Prague where he was to study. Unfortunately, he arrived too late to enrol at Charles-Ferdinand University; he never studied Greek, a required subject; and he was illiterate in Czech, another required subject.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



- 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?

Q

a

A

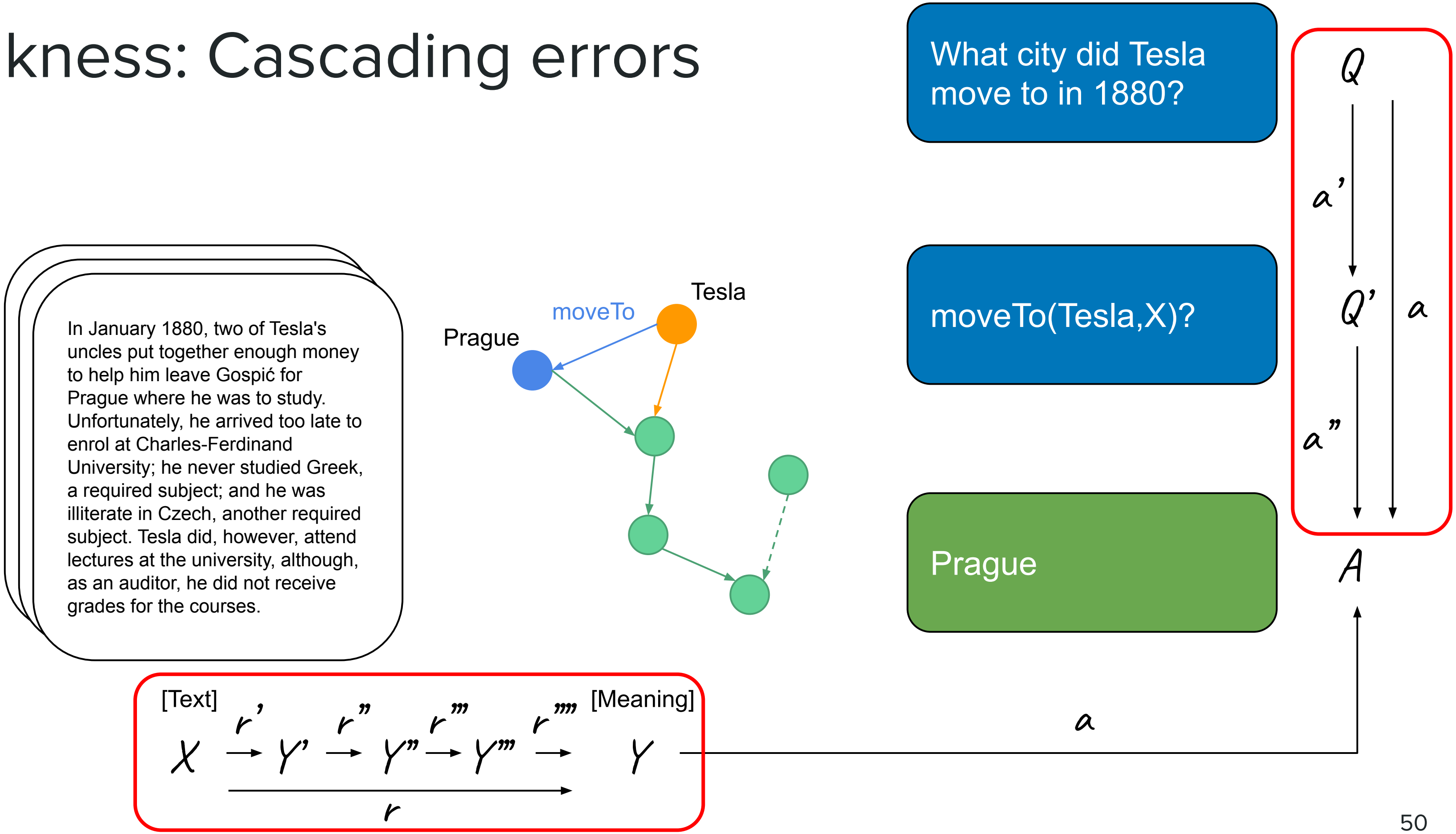
Prague

[Text] [Meaning]

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

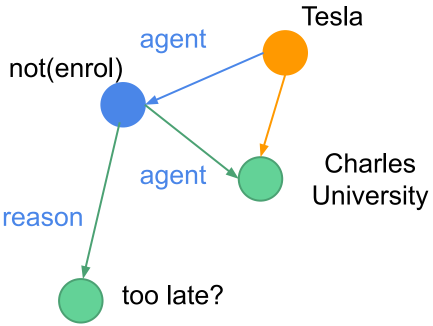
$$X \xrightarrow{r} Y \xrightarrow{a} A$$

Weakness: Cascading errors



Weakness: Engineering Schemas and Formalisms

Unfortunately, he arrived too late to enrol at Charles University



Why did he not enrol?

He arrived too late?

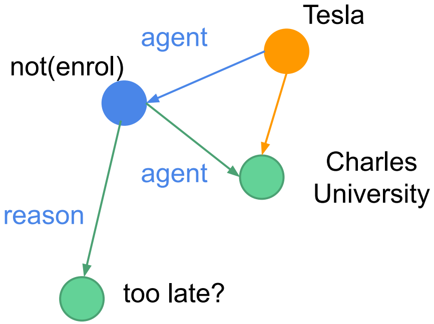
getting this right is hard

Weakness: Annotation

He et al., 2015

much easier

Unfortunately, he arrived too late to enrol at Charles University

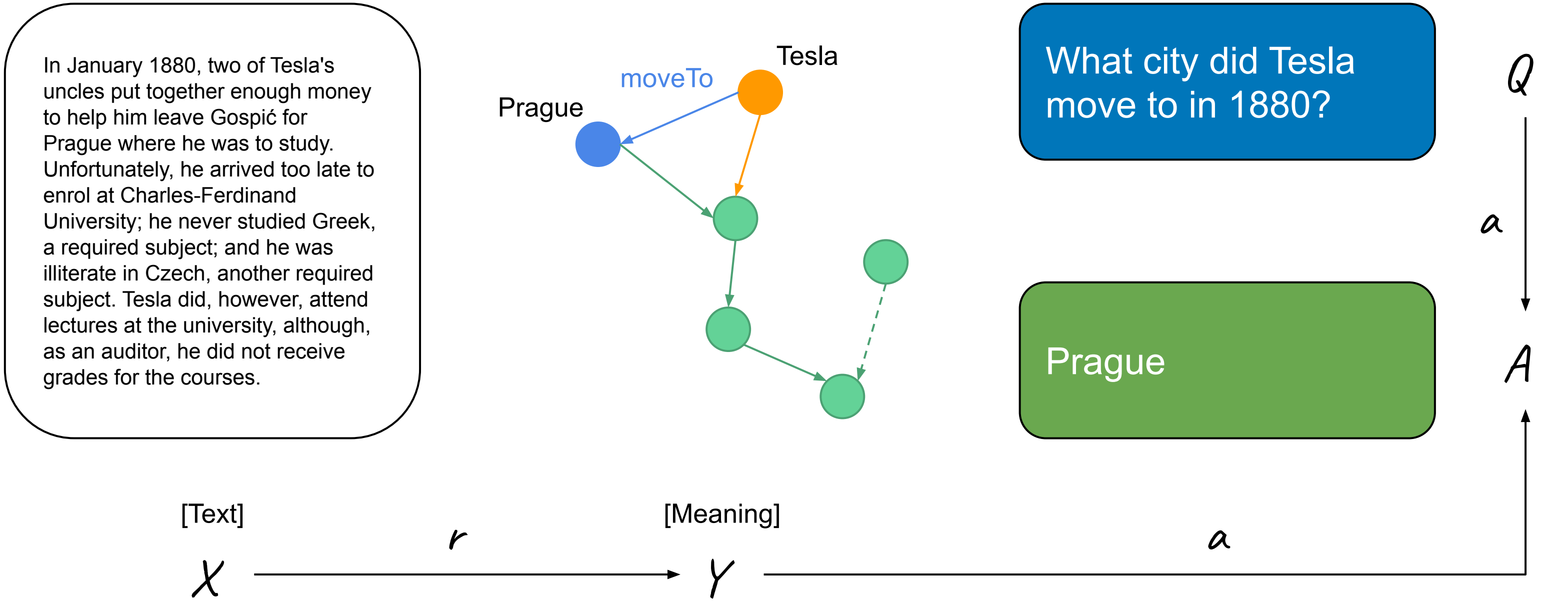


Why did he not enrol?

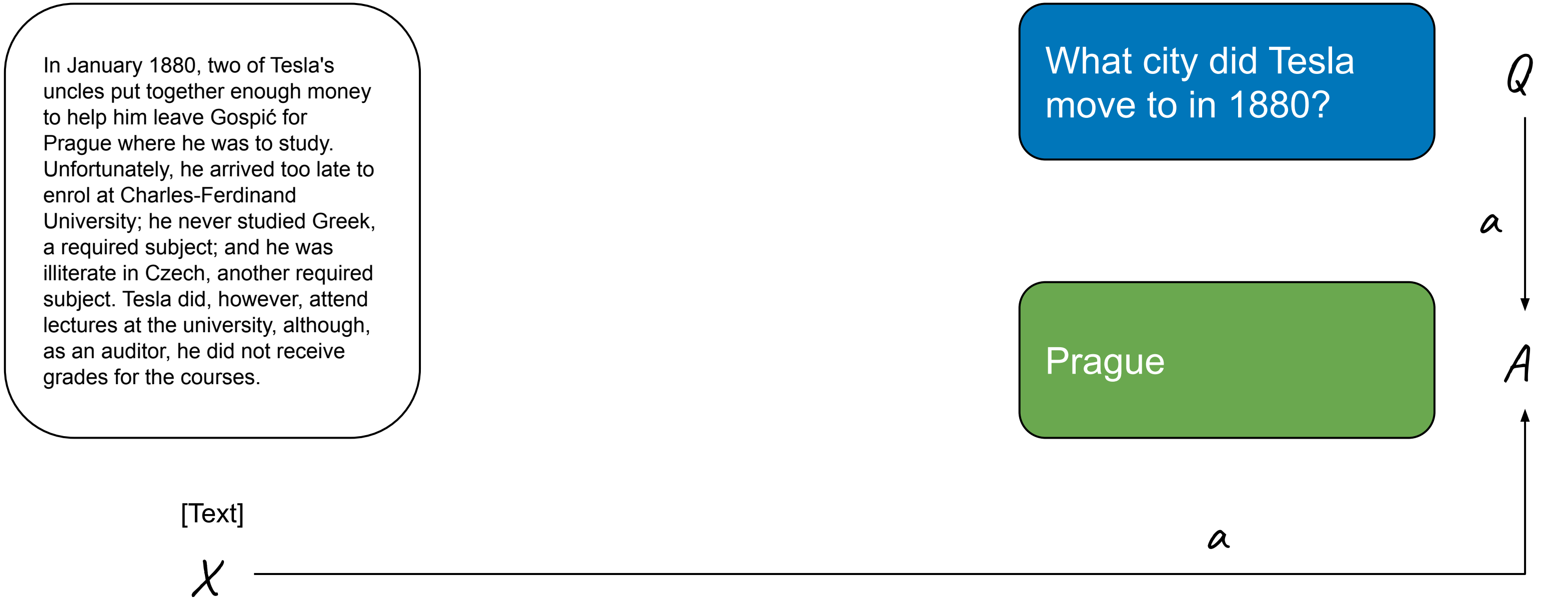
He arrived too late?

Hard to annotate

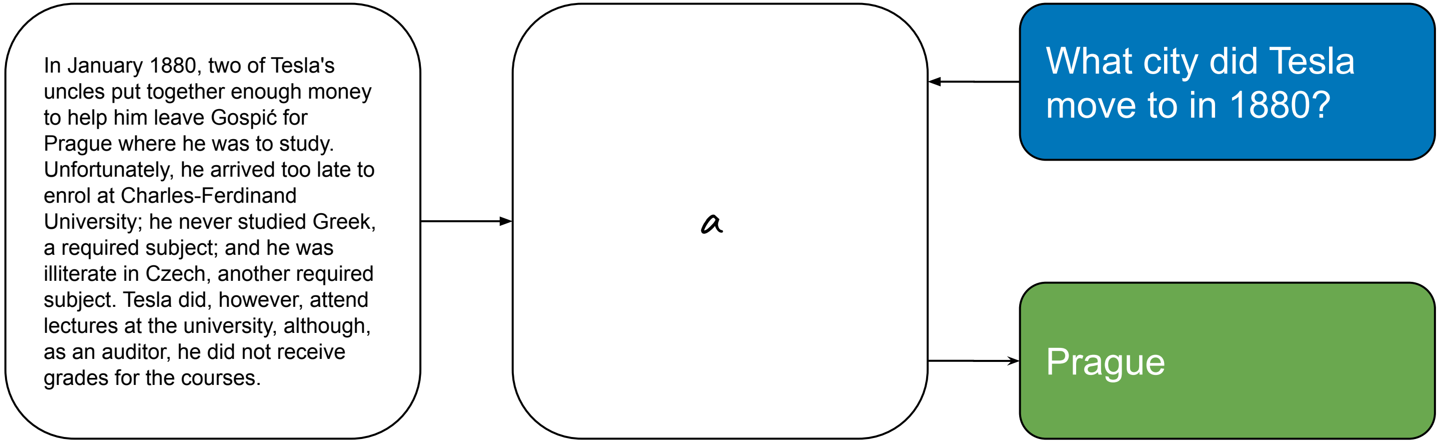
Is there another way?



Omitting Intermediate Meaning Representations



Learn an End-to-End Function



End-to-End Question Answering

Stanford Question Answering Dataset (SQuAD)

Rajpurkar et. al. 2016

Text Passage

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

Question + Answer

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

within a cloud

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

Stanford Question Answering Dataset (SQuAD)

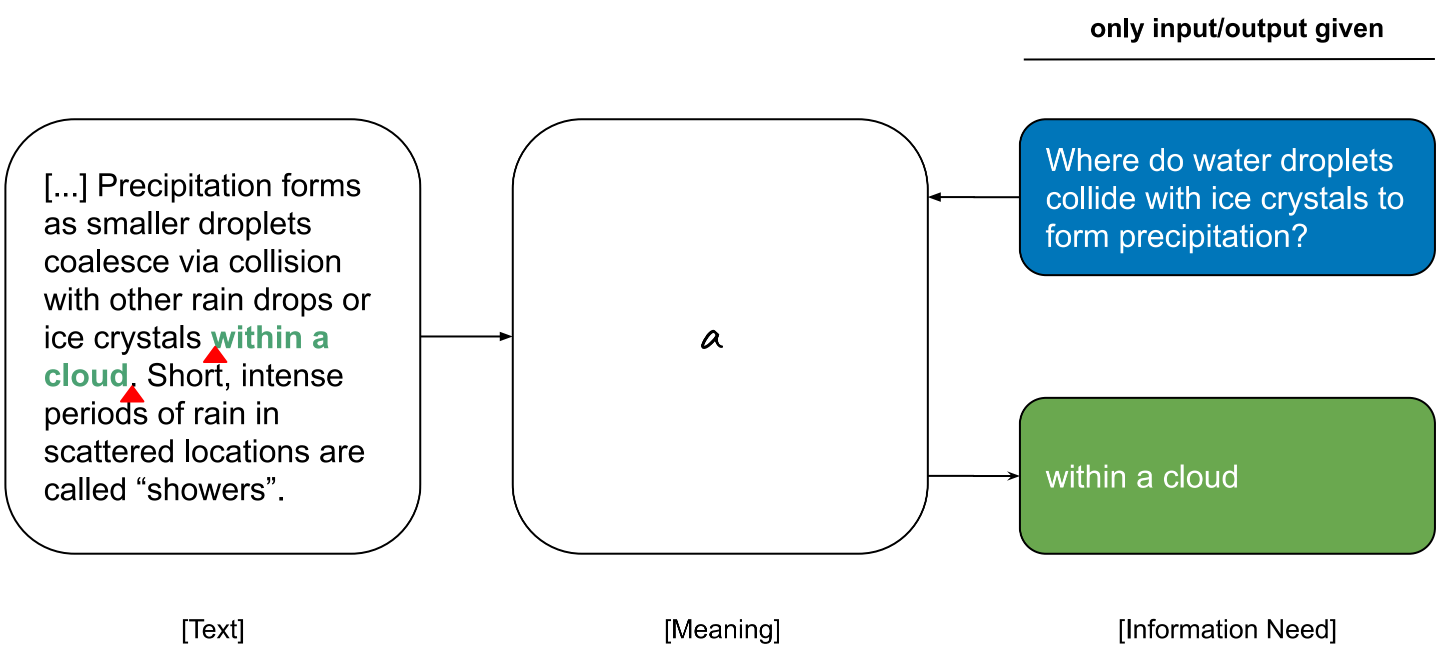
Rajpurkar et. al. 2016

- **Dataset size**: 107,702 samples
- Widely used benchmark dataset
- **Task:** *Extractive* Question Answering
    - Other forms of QA exist, e.g. free-form answer generation, multiple choice

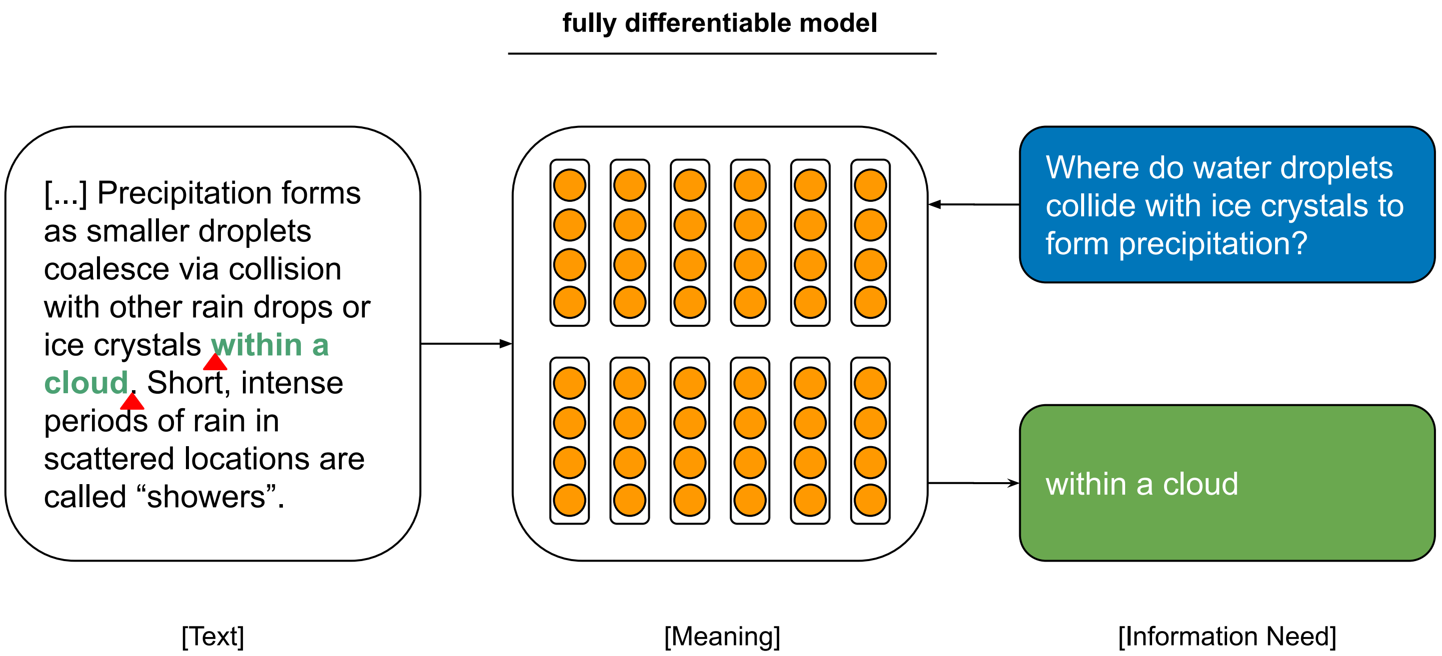
List of Other QA Datasets

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

End-to-end Machine Reading for Question Answering



End-to-end Machine Reading for Question Answering

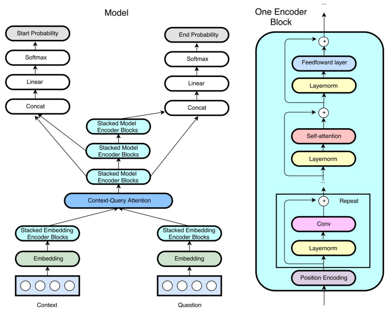


End-to-end Machine Reading for Question Answering

QANet, Yu et. al. 2018

State-of-the-Art Architecture

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



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

within a cloud

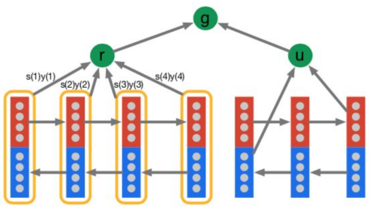
[Text] [Meaning] [Information Need]

End-to-end Machine Reading for Question Answering

Hermann et. al. 2015

Simpler Architecture

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



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

within a cloud

[Text]

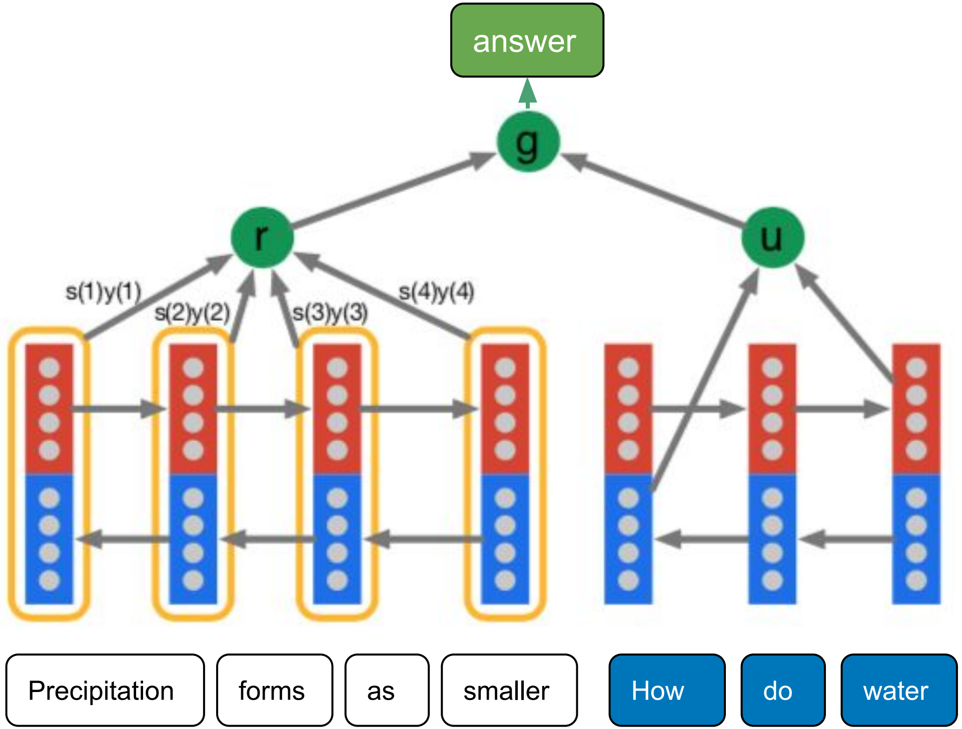
[Meaning]

[Information Need]

The Attentive Reader Model: Overview

Hermann et. al. 2015

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



Modified visualization from Hermann et. al. 2015

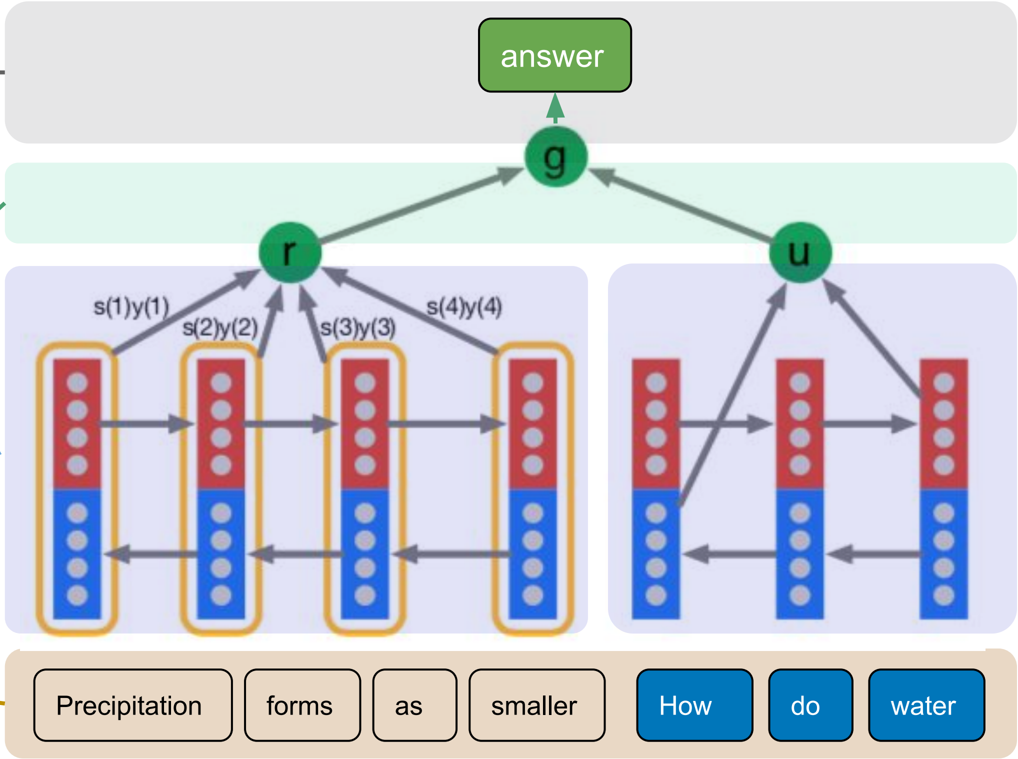
The Attentive Reader Model: Overview

Answer Selection: answer prediction

Sequence Interaction: Matching text with question

Composition: incorporating context around words

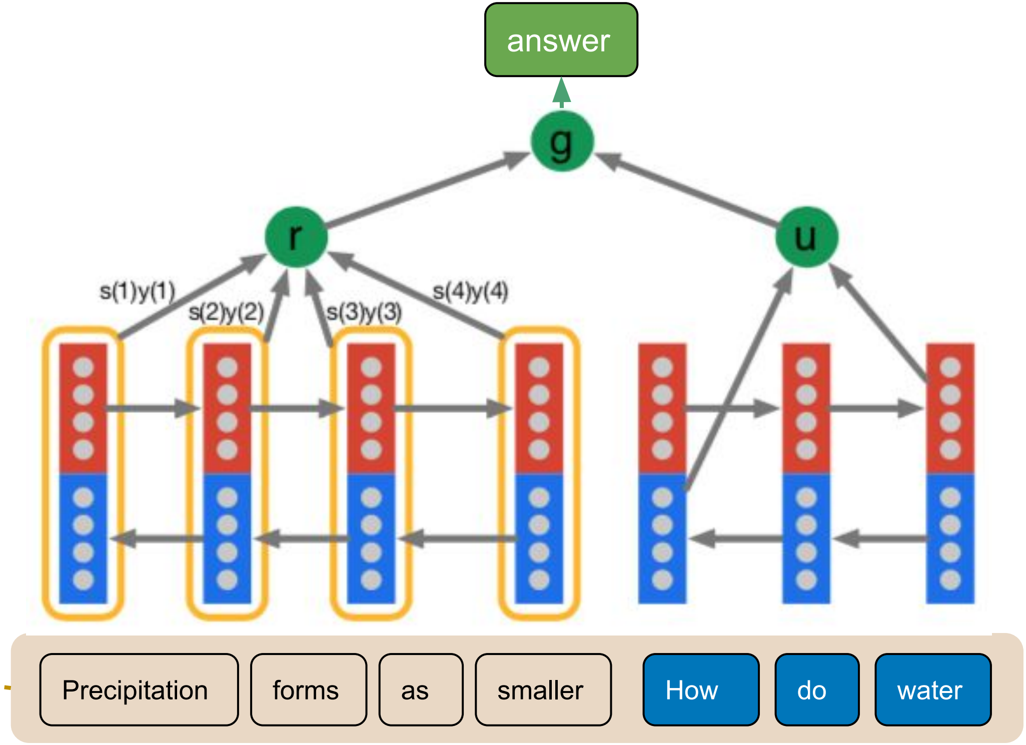
Input: Representing symbols as vectors



Modified visualization from Hermann et. al. 2015

The Attentive Reader Model: Overview

Input: Representing symbols as vectors

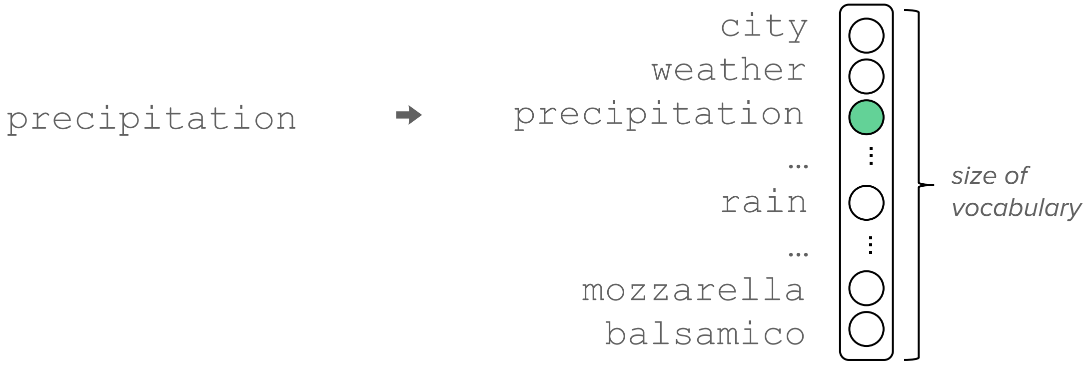


Modified visualization from Hermann et. al. 2015

Representing Symbols as Vectors

- **Problem:** Words / characters are discrete symbols, but neural nets work with vector inputs

- **Naive solution:** construct one-hot vector for each word

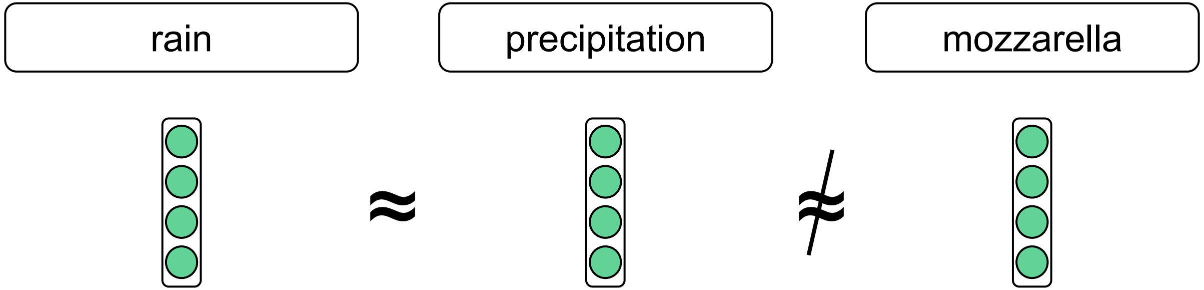


Representing Symbols as Vectors

Problem with naive solution:

- one-hot vectors do not represent relationships between words
    - all one-hot vectors are orthonormal
- high-dimensional, extremely sparse input
- hard to train model which generalizes across similar words
    - e.g. rain vs. precipitation

Ideal Vector Representations for Words



Similar meaning of words → similar vector representations

?

Word Similarity

?

We found a little, hairy wampimuk
sleeping behind the tree.

after Marco Baroni

use context to
infer meaning!

Distributional Hypothesis: “*Words that are used and occur in the same contexts tend to purport similar meanings.*” *(Harris, 1954)*

Short Version:
“You shall know a word by the company it keeps.” (Firth, 1957)

Word Similarity

"You shall know a word by the company it keeps."

➜ Two words are similar if they appear in the same documents.

Term-Document matrix:

	d1	d2	d3	d4	...	dM
city	2	0	0	0	...	1
weather	0	1	0	1	...	1
precipitation	4	2	0	1	...	1
...	...	...	...	...	...	...
rain	1	1	0	1	...	1
mozzarella	0	0	3	0	...	0
balsamico	0	0	1	0	...	0

Vector for "rain" is similar to "precipitation", not to "mozzarella".

Word Similarity

"You shall know a word by the company it keeps."

➜ Two words are similar if they appear in the same documents.

Term-Document matrix:

	d1	d2	d3	d4	...	d*M*
city	2	0	0	0	...	1
weather	0	1	0	1	...	1
precipitation	4	2	0	1	...	1
...	...	...	...	...	...	...
rain	1	1	0	1	...	1
mozzarella	0	0	3	0	...	0
balsamico	0	0	1	0	...	0

Somewhat collinear,
but very sparse

Combatting Sparsity

- **Key Idea:** Approximate Sparse matrix using low-rank matrix factorization

  ➜ Dense Factor matrices for words, and for documents

	d1	**d2**	**d3**	**d4**	**…**	**d*M***
city	2	0	0	0	…	1
weather	0	1	0	1	…	1
precipitation	4	2	0	1	…	1
…	…	…	…	…	…	…
rain	1	1	0	1	…	1
mozzarella	0	0	3	0	…	0
balsamico	0	0	1	0	…	0

$$\approx \quad U \times \Sigma \times V$$

Row vectors: dense representations for each word

Word Embeddings

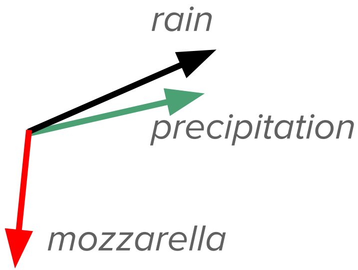
- **word embeddings:**
  dense vector representations for words of low dimensionality (e.g. 300)
- can capture word similarity (to a degree)
- usually pretrained on large text corpus
- e.g. **word2vec** (Mikolov et al., 2013)
- Different approach: character-based word embeddings, *e.g., Kim et al. 2016*

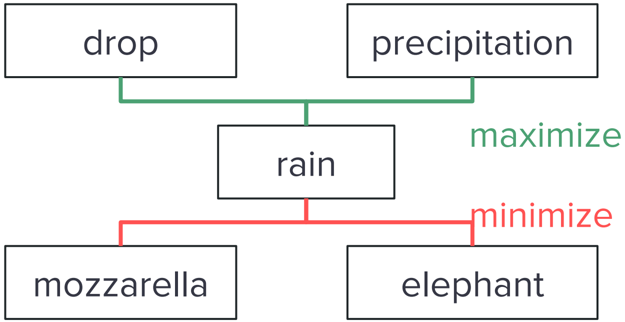
Word2Vec - (SkipGram with Negative Sampling)

1. *Maximize similarity between co-occurring words*
2. *minimize similarity between non co-occurring words*

similarity = collinearity

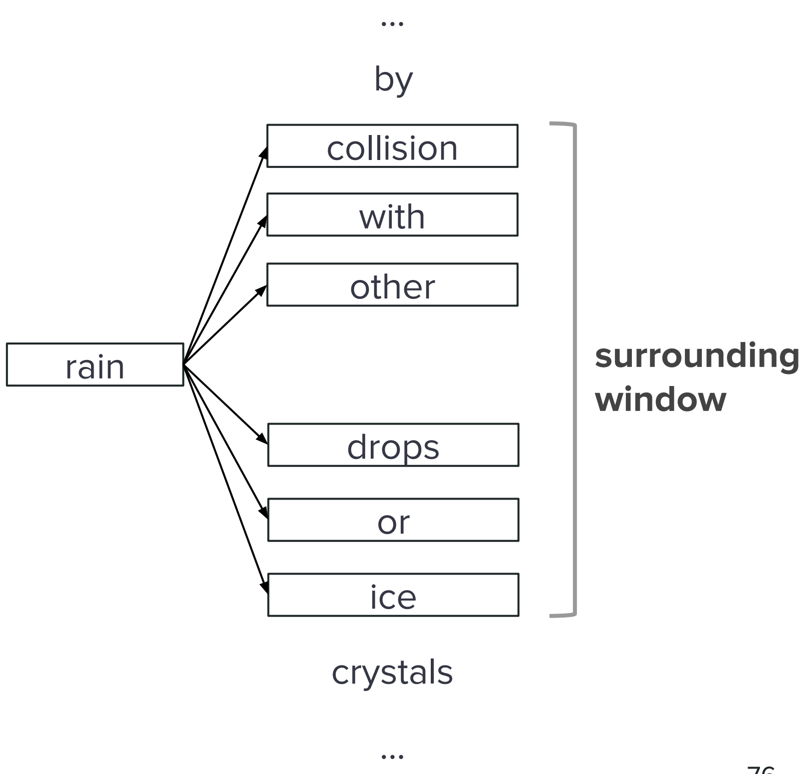
$$s(a,b) = \sigma(\mathbf{v}_a \cdot \mathbf{v}_b)$$



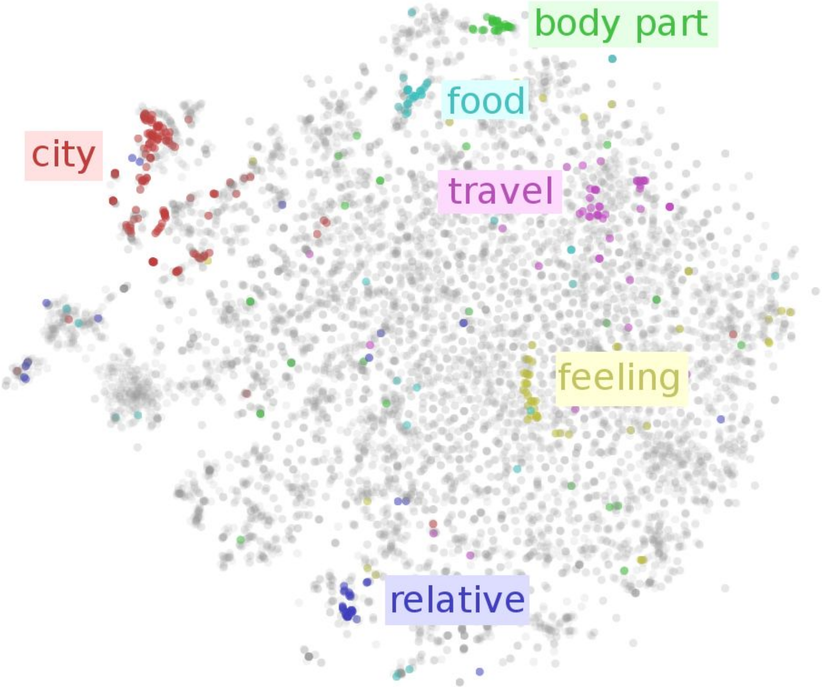


Word2Vec - (SkipGram with Negative Sampling)

- Training: use vectors to predict words in surrounding window
- Implicitly related to factorization of word-context PMI matrix *(Levy and Goldberg, 2014)*

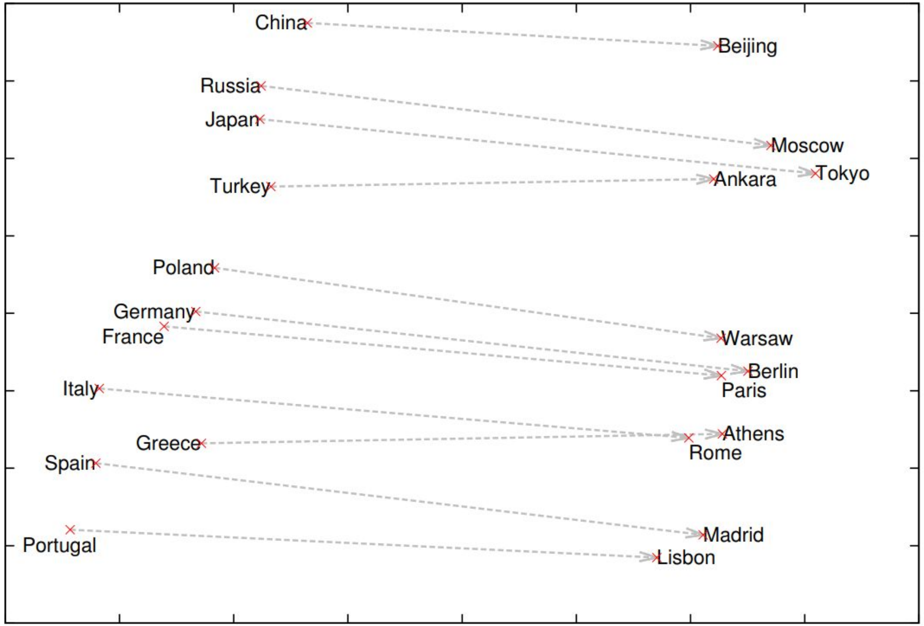


Visualizing Word Embeddings



T-SNE visualization of word embeddings

http://colah.github.io/posts/2015-01-Visualizing-Representations/



PCA Plot of Country Capital

Mikolov et al. (2013)

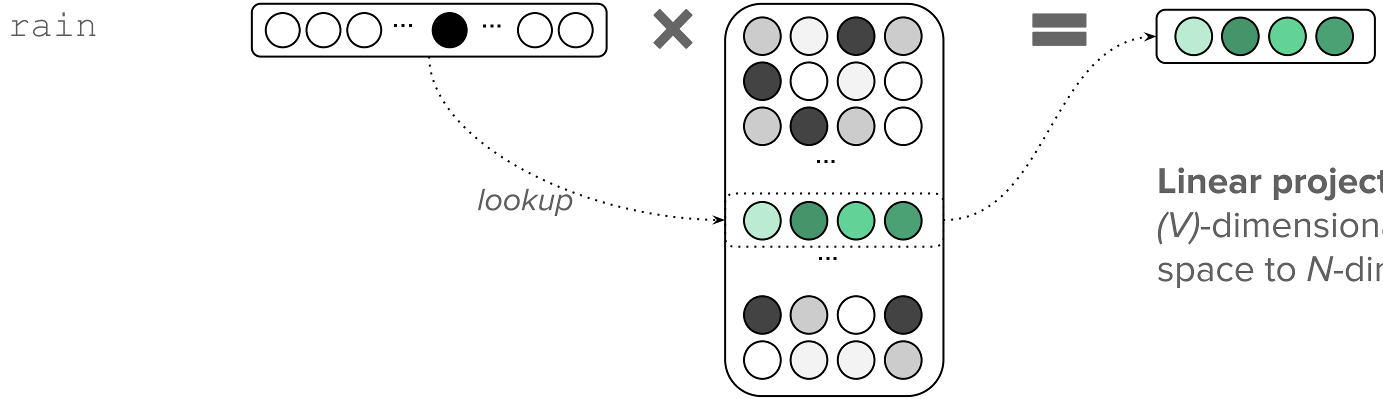
Interpretation as Linear Projection

Symbol
word,
character, …

One-hot vector
High-dimensional *(V)*
Sparse vector

Linear projection

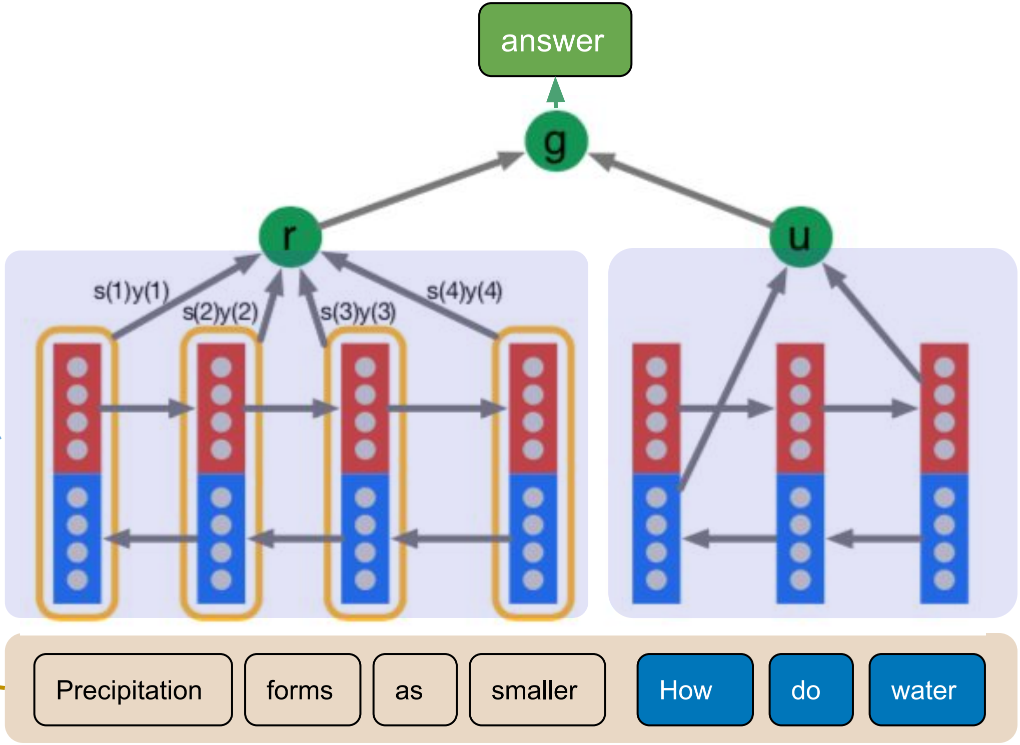
Word embedding
Low-dimensional *(N)*
dense vector



Linear projections from *(V)*-dimensional discrete symbol space to *N*-dimensional

Word Embedding Matrix
[projection matrix]

The Attentive Reader Model: Overview

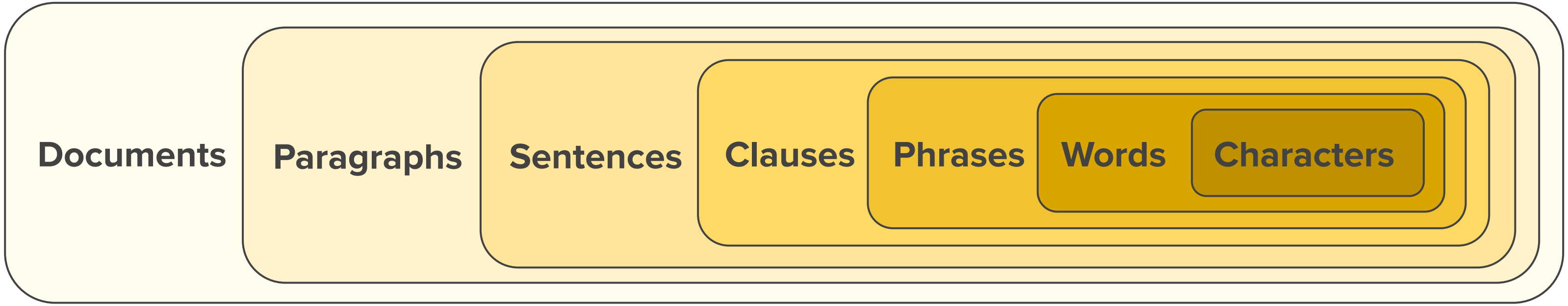


Composition: incorporating context around words

Input: Representing symbols as vectors

Modified visualization from Hermann et. al. 2015

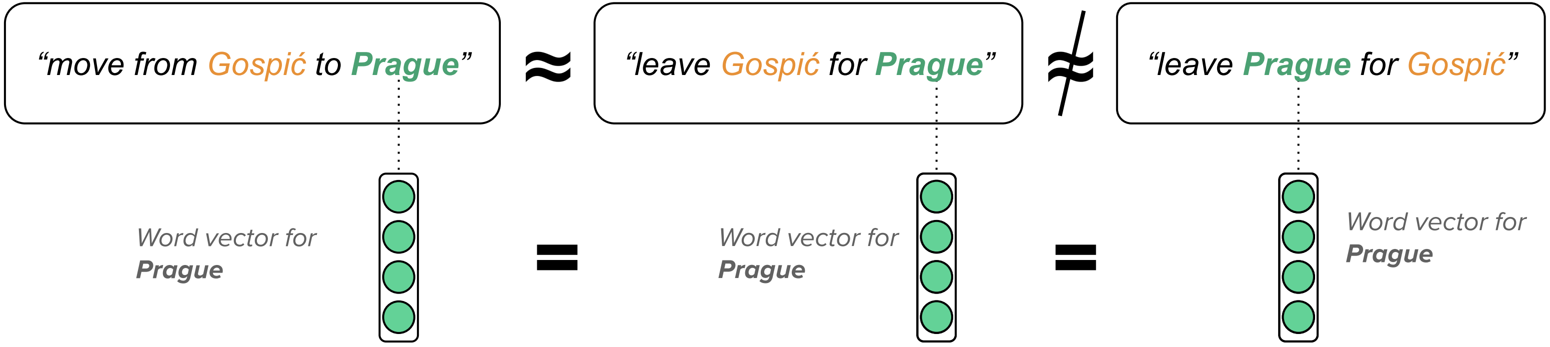
Language is Compositional



Challenges

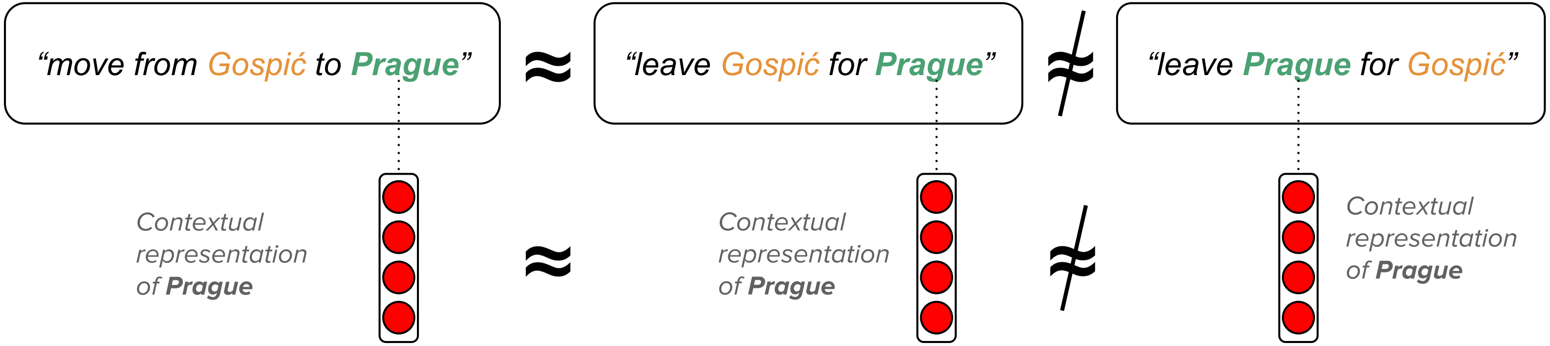
- Inductive bias: which composition function to use?
  - sequence, tree or more general graph structures?
  - Varies for different levels
- capturing long-range dependencies
  - co-reference (tracking entities)
  - effective information flow: ease of learning

Representing Words in Context



- Word representations should vary depending on context

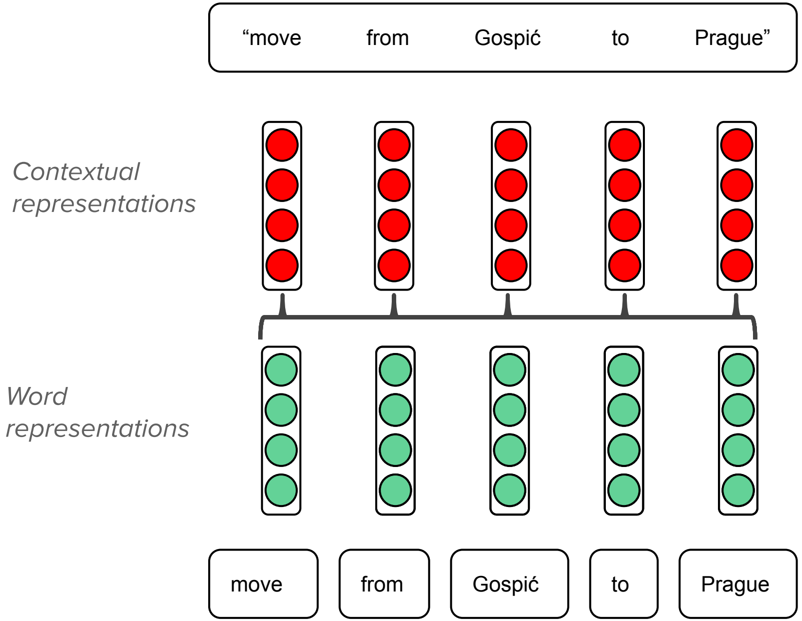
Representing Words in Context



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

Representing Words in Context

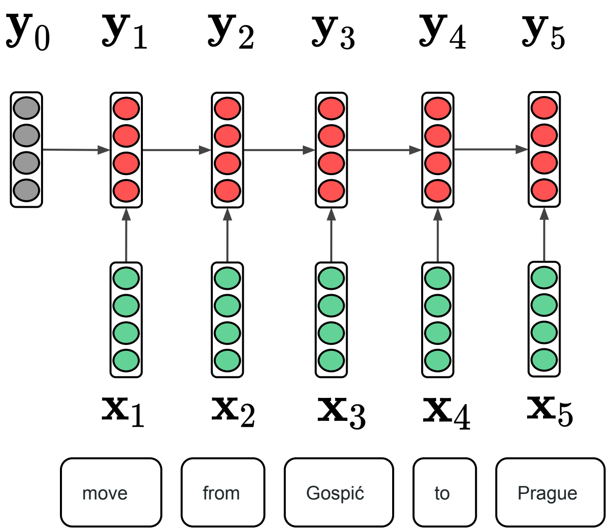
- composition of word vectors into contextualized word representations
- use vector composition function
  - different options



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:

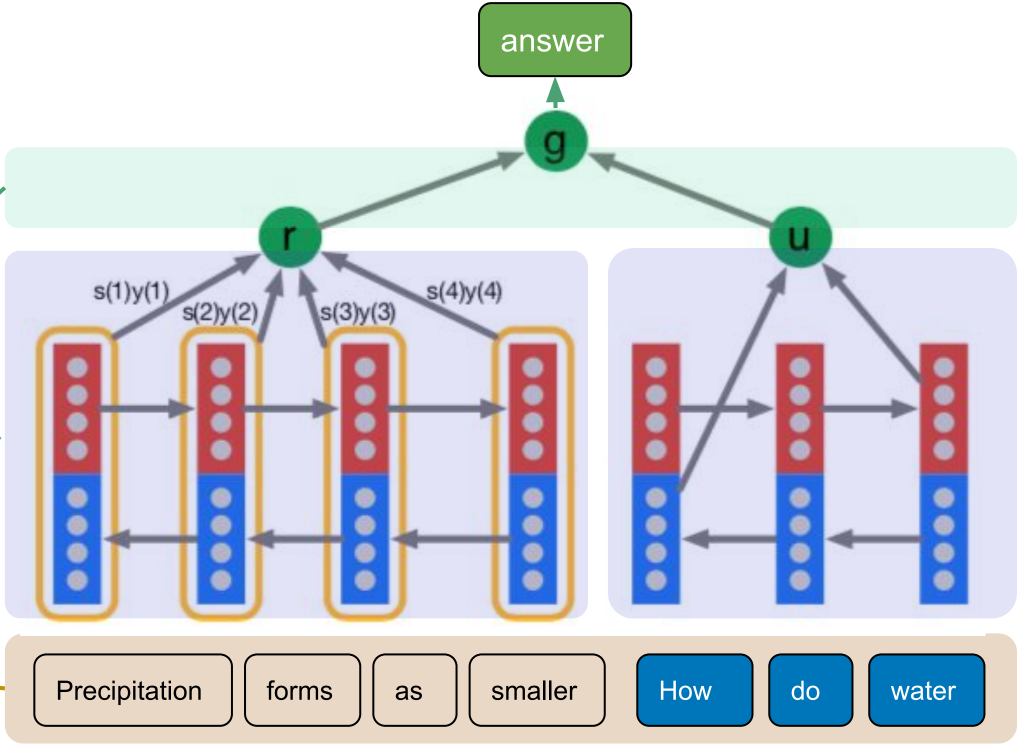
- TreeLSTM (Tai et al. 2015)
- RNN Grammars (Dyer et al. 2016)
- Bias towards syntactic hierarchy

The Attentive Reader Model: Overview

Sequence Interaction: Matching text with question

Composition: incorporating context around words

Input: Representing symbols as vectors



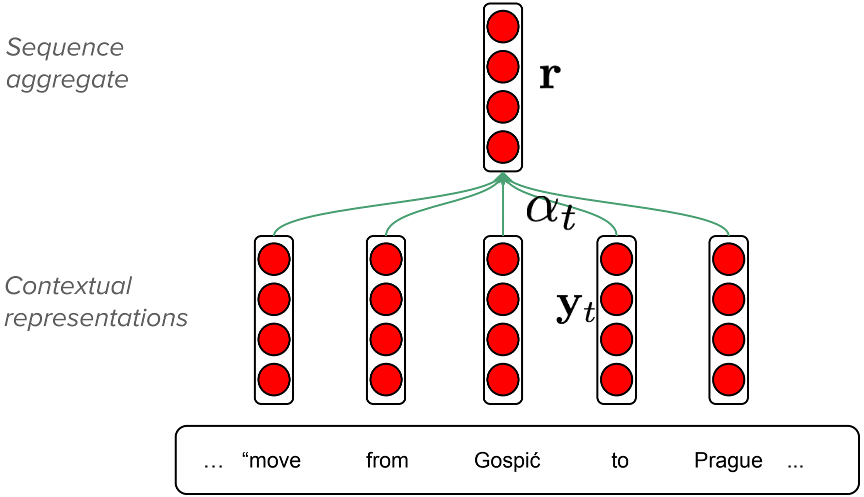
Modified visualization from Hermann et. al. 2015

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: 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



$$\alpha_t = f(\mathbf{y}_t, \mathbf{q})$$

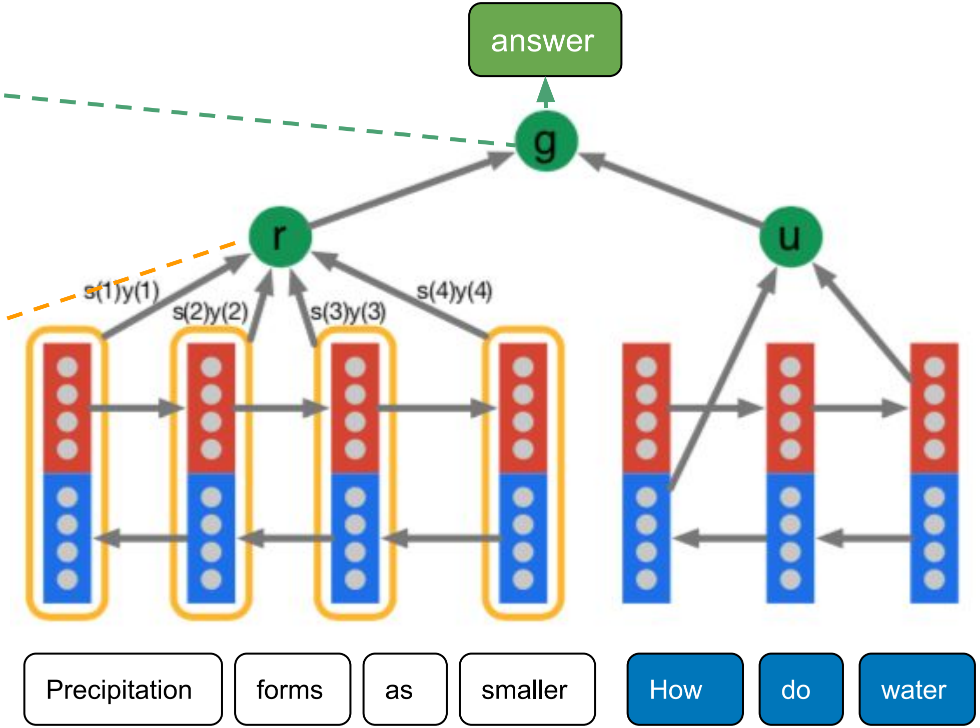
$$\mathbf{r} = \sum_{t=1}^{T} \alpha_t \mathbf{y}_t$$

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

Modelling Sequence Interactions

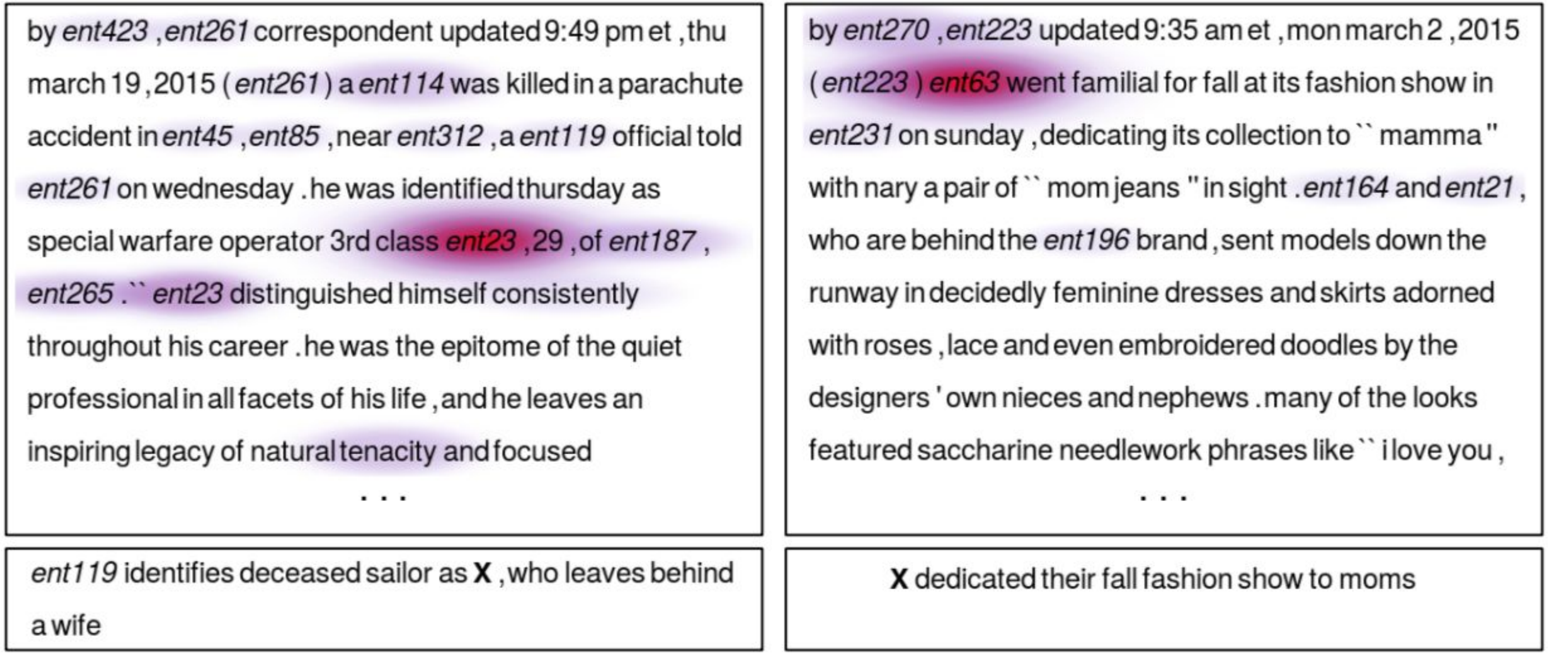
Combination of question and text representation

attention-weighted sum of contextualised word representations



Modified visualization from Hermann et. al. 2015

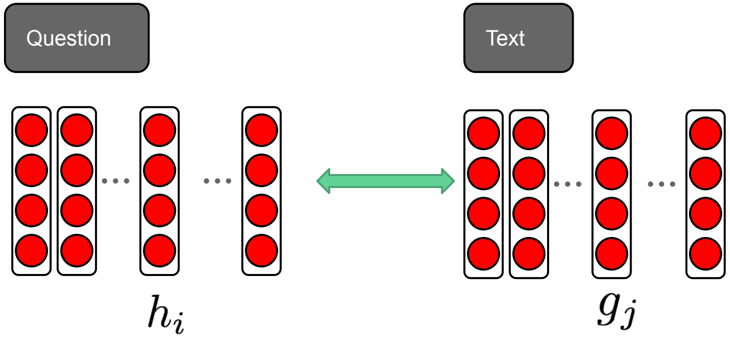
Example: Learned Attention Patterns



Intuition: Relevancy Masks

Visualization from Hermann et. al. 2015

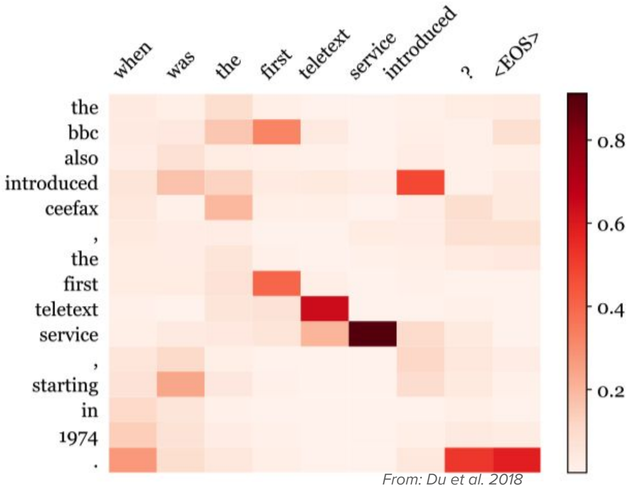
Modeling Sequence Interaction



"Naive" approach:

- **Goal in QA**: match question with text
- conditioning sequence representations **on one another**
  ➜ e.g., compute token-token attention masks from latent states
- Interpretation: per-word relevancy mask, (soft-)alignment

Modeling Sequence Interaction - Attention



From: Du et al. 2018

Word-to-word attention masks

e.g. $a_{ij} \propto \text{Bilinear}(h_i, g_j)$

- **Goal in QA**: match question with text
- conditioning sequence representations **on one another**
  ➔ e.g., compute token-token attention masks from latent states
- Interpretation: per-word relevancy mask, (soft-)alignment

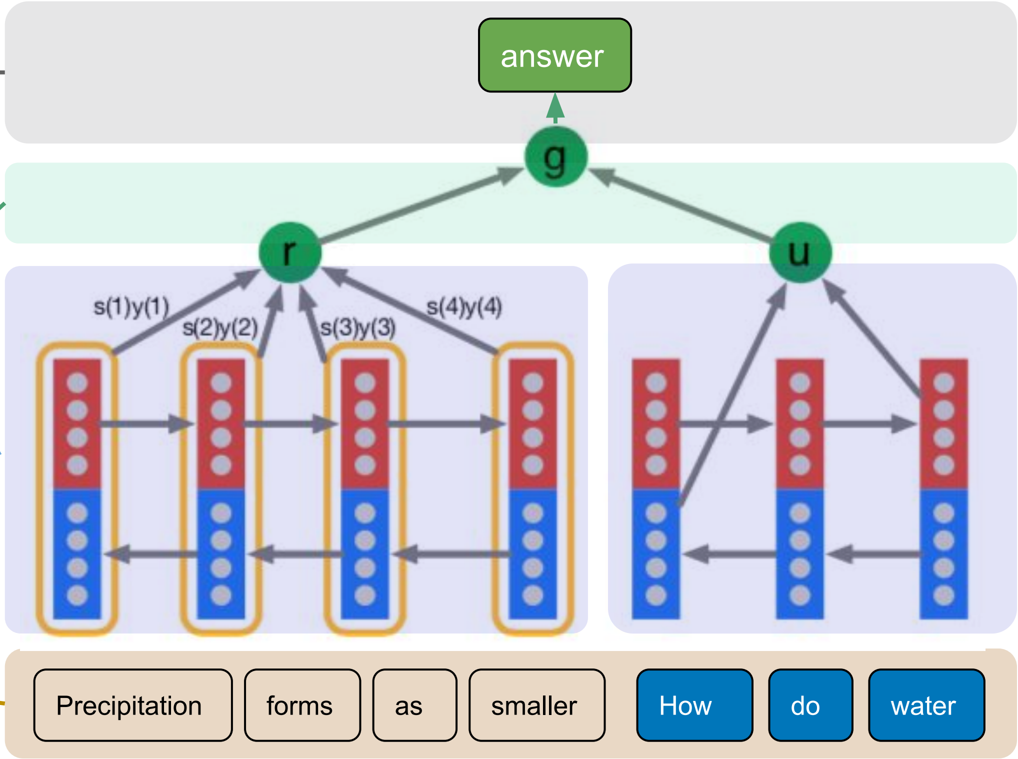
The Attentive Reader Model: Overview

Answer Selection: answer prediction

Sequence Interaction: Matching text with question

Composition: incorporating context around words

Input: Representing symbols as vectors



Modified visualization from Hermann et. al. 2015

Answer Prediction

- Linear projection
- **Probability distribution over different answer options**
  - spans in text -- distribution over positions for beginning and end
  - multiple choice: candidates
- **Training:** cross-entropy loss

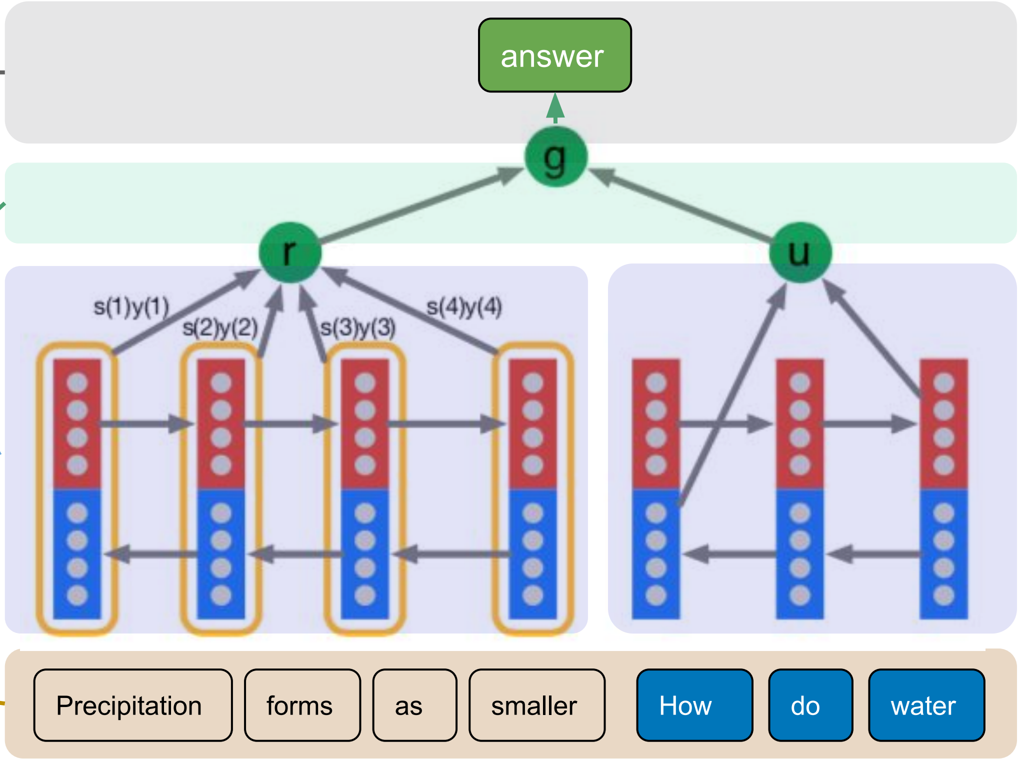
The Attentive Reader Model: Overview

Answer Selection: answer prediction

Sequence Interaction: Matching text with question

Composition: incorporating context around words

Input: Representing symbols as vectors



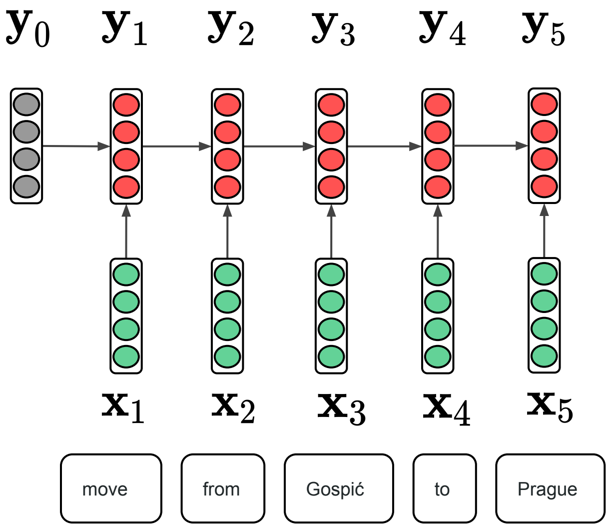
Modified visualization from Hermann et. al. 2015

Other Types of Composition Functions

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



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

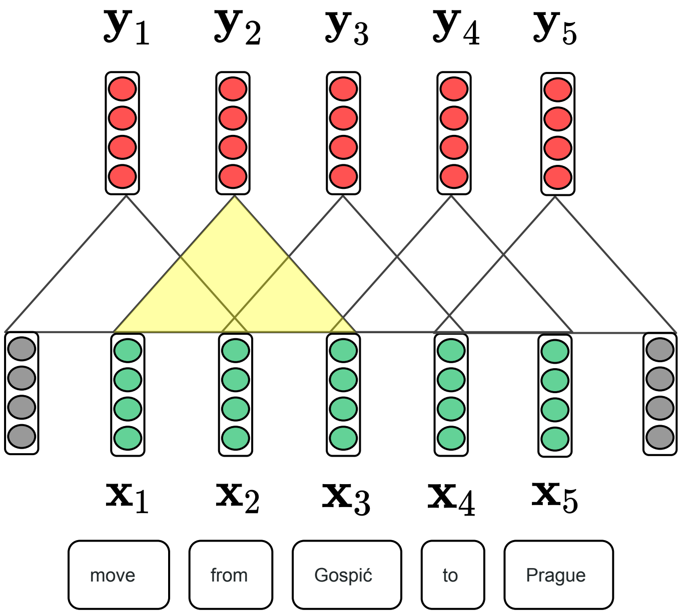
Tree-variants:

- TreeLSTM (Tai et al. 2015)
- RNN Grammars (Dyer et al. 2016)
- Bias towards syntactic hierarchy

Convolutional Layer

- **Idea**: text as collection of N-Grams
- **Inductive bias:** Locality
  - Only symbols within context window have impact on the current hidden state
  - typically: pooling across sequence
- **Advantages**
  - Parallelizable and fast
- **Disadvantages**
  - Limited context window
  - remedy: stacking many layers + dilation

- *Good for:* Character-based word representations, phrases, multi-word representations



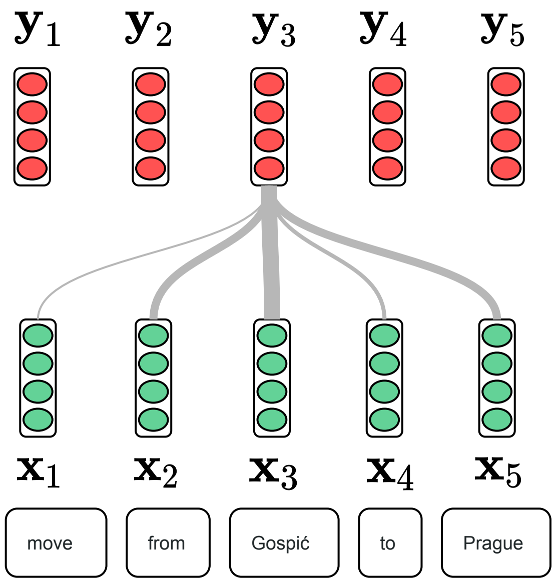
$$\mathbf{y}_t = f(\mathbf{x}_{t-k}, \ldots, \mathbf{x}_t, \ldots \mathbf{x}_{t+k})$$

See e.g.: Kim et al. 2016

Self-Attention Layer

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

- *Good for:* phrases, sentences, paragraphs



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

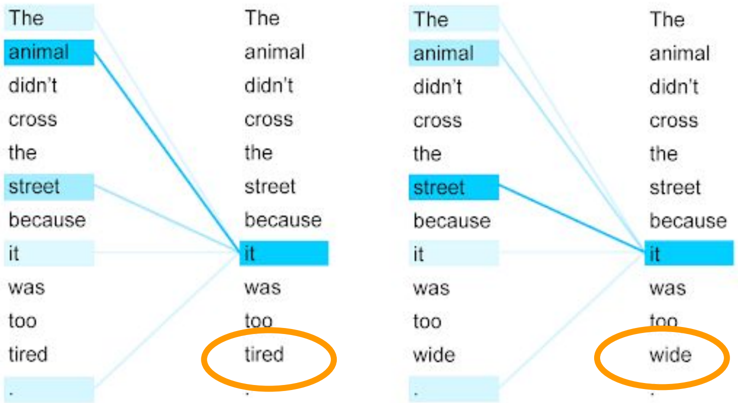
$$\tilde{\mathbf{x}}_t^k = \sum_{j=1}^{T} \alpha_{j,t}^k \mathbf{x}_j \qquad k = 1, \ldots, K$$

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

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

Self-Attention Layer

- **graph with weighted edges** of *K* types
- Can capture:
  - coreference chains
  - syntactic dependency structure in text
  - see for instance: Vaswani et al. 2017; Yang & Zhao et al. 2018



Transformer Self-Attention Coreference Visualization

https://ai.googleblog.com/2017/08/transformer-novel-neural-network.html

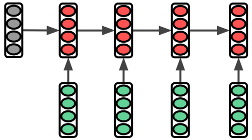
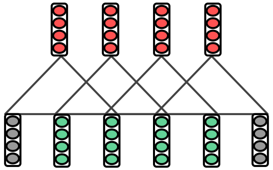
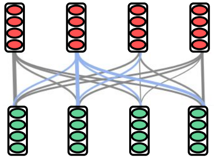
Self-Attention Layer

used in many SoTA MRC models, e.g.

- Language Modelling, Natural Language Inference: Cheng et al. 2016 (*intra-attention*)
- QA: Wang et al. 2017 (*self-matching*), Yu et al. 2018 (*self-attention)*
- Via Transformers: pretty much everywhere today!

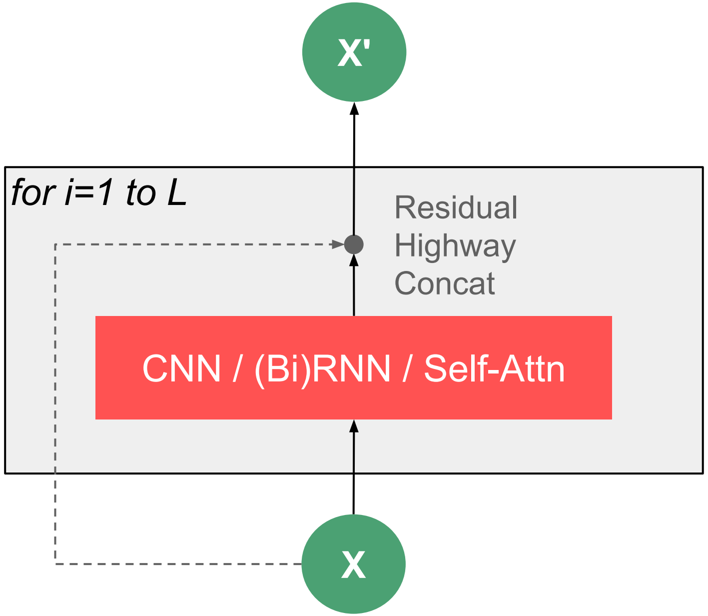
Compositional Sequence Encoders - Overview

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

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

Deep Compositional Sequence Encoders

- pure RNN based models usually not deep (typically L < 3)
  - Depth in RNNs comes naturally by processing sequentially
- CNN based models are quite deep
  - E.g. QANet: 42 CNN + 21 SelfAttn
  - use residual/highway layers or concatenation to avoid vanishing gradient
- Self-Attn. is usually applied after layers of CNN or RNN
  - exception: Transformer (Vaswani et al. 2017)

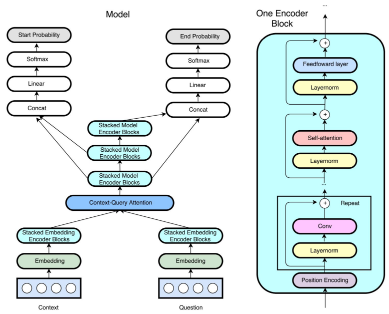


End-to-end Machine Reading for Question Answering

QANet, Yu et. al. 2018

State-of-the-Art Architecture

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



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

within a cloud

[Text]

[Meaning]

[Information Need]

QANet - A (non BERT) State-of-the-Art Architecture

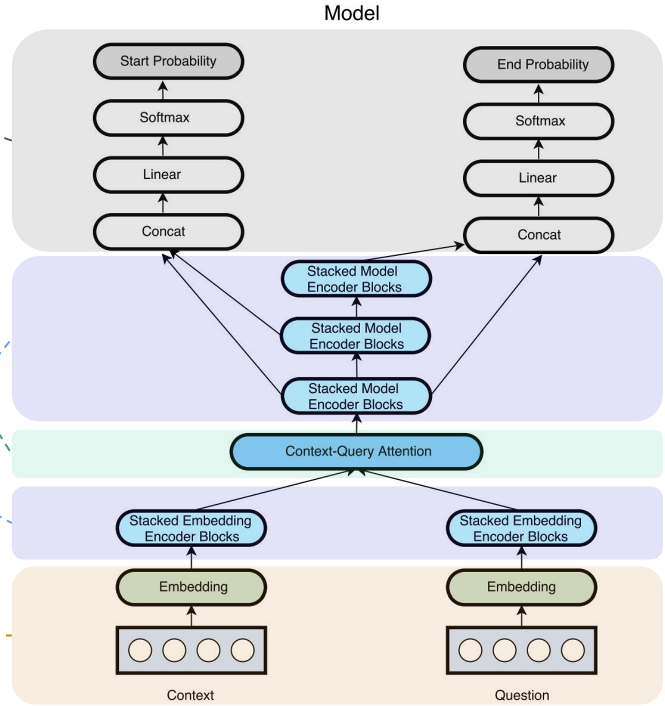
QANet, Yu et. al. 2018

Span Scoring:
answer prediction

Sequence Interaction:
Matching text with question

Composition: incorporating context around words

Input: Representing symbols as vectors



QANet - A State-of-the-Art Architecture

QANet, Yu et. al. 2018

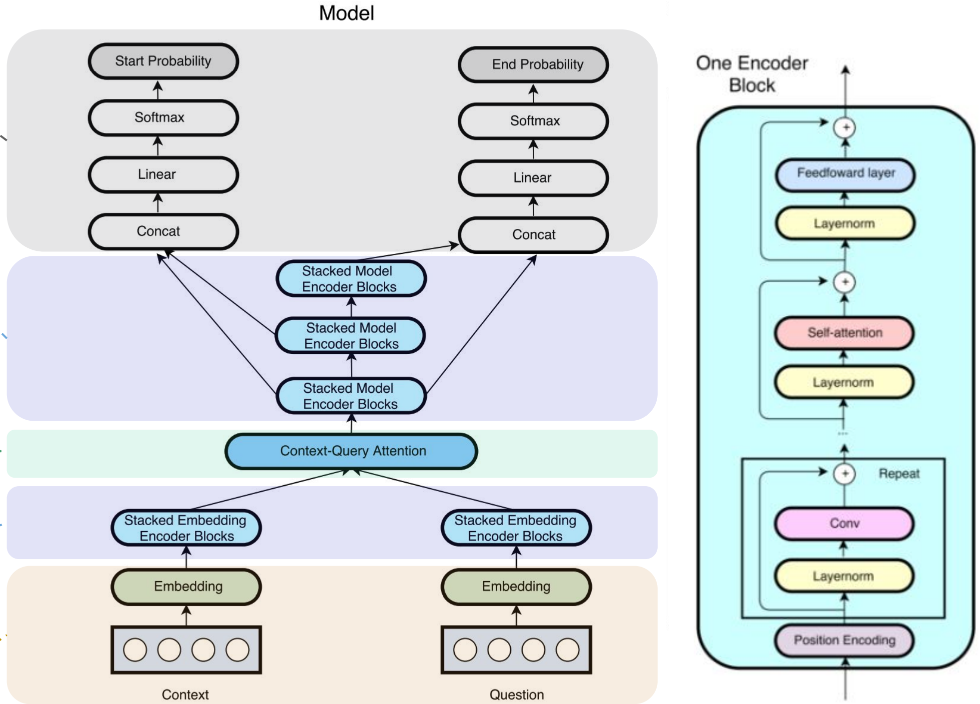
Span Scoring: linear projection, score for start and end position

Composition 2:
(2 * Conv + 1 * Self Attn) * 7 Blocks
= 21 Layers * 3 = **63** Layers

Sequence Interaction:
Bidirectional Attention

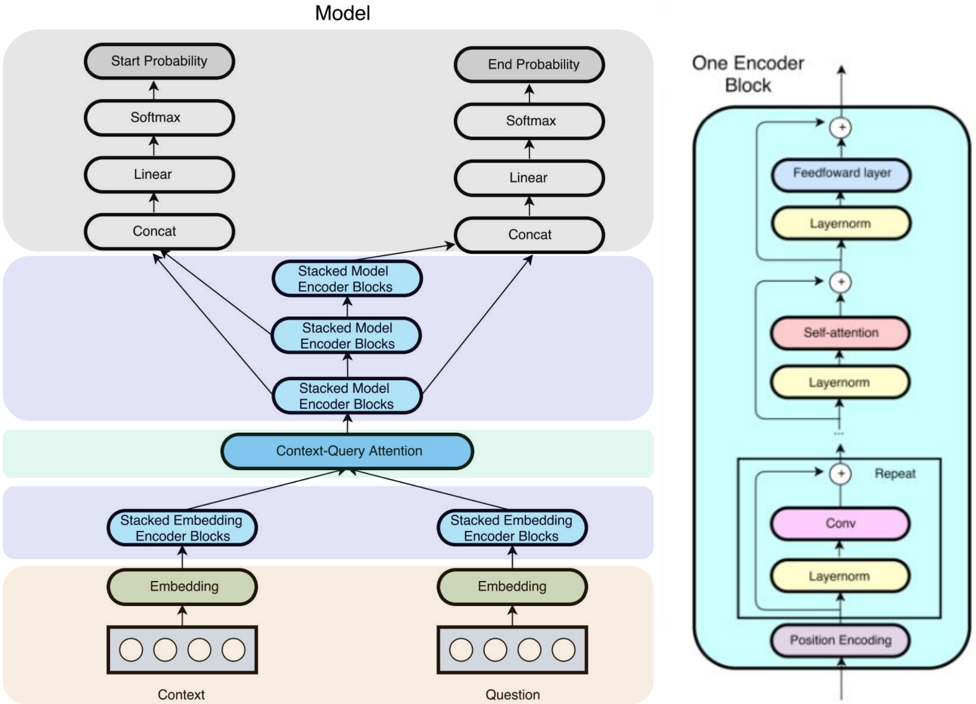
Composition 1:
(4 * Conv + 1 * Self Attn) = **5** Layers

Input: Representing symbols as vectors



QANet - A (Non-BERT) State-of-the-Art Architecture

- extremely deep
  - **68** compositional, residual layers
- but no RNNs
  - parallelizable and fast
- ~~Currently best~~ model on SQuAD
  - Self-attention
  - Data augmentation
  - Parallelizable ➜ faster training / tuning



Transformer-based State-of-the-Art Architecture

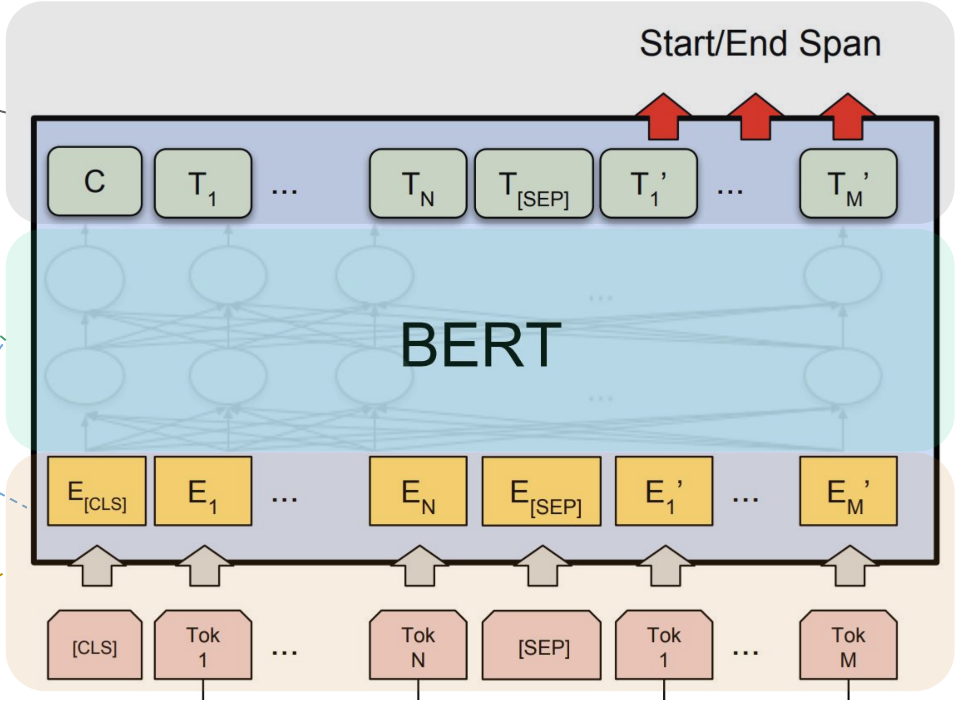
Devlin et al, 2019

Span Scoring: answer prediction

Sequence Interaction: Matching text with question

Composition: incorporating context around words

Input: Representing symbols as vectors



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.

Conclusion

👍 We gathered all ingredients to build state-of-the-art supervised MRQA systems!

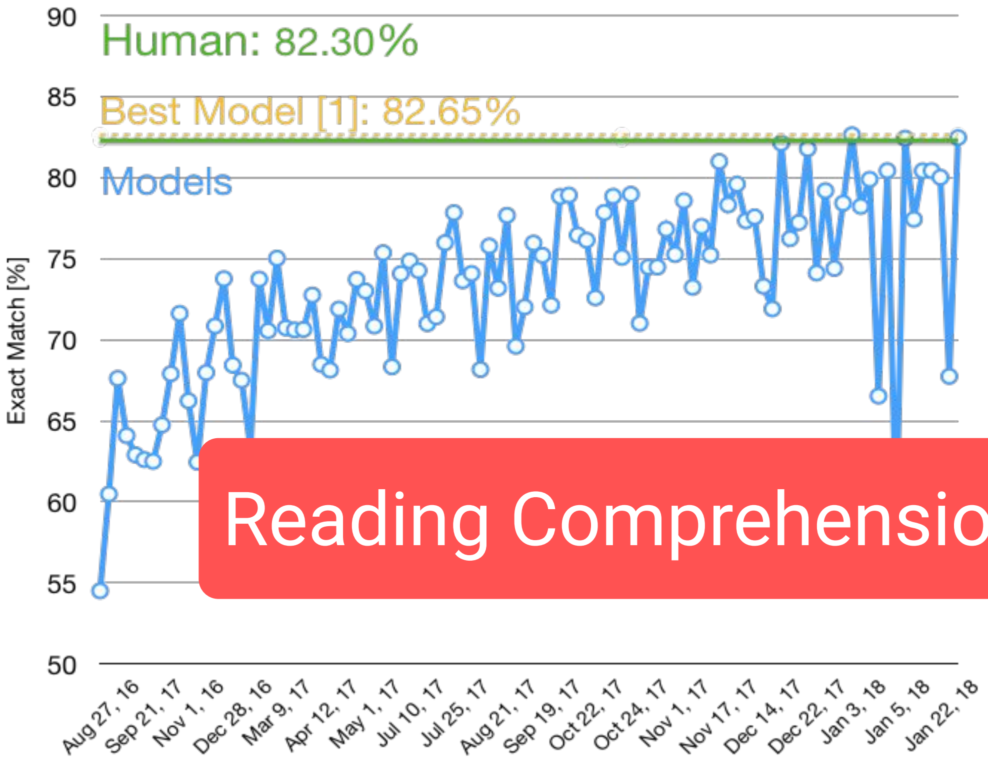
- We know about:
  - Representing words with and without context
  - Modeling compositionality
  - Modeling sequence interaction (question-paragraph)
  - Answer questions by pointing to the start and end of the answer-span

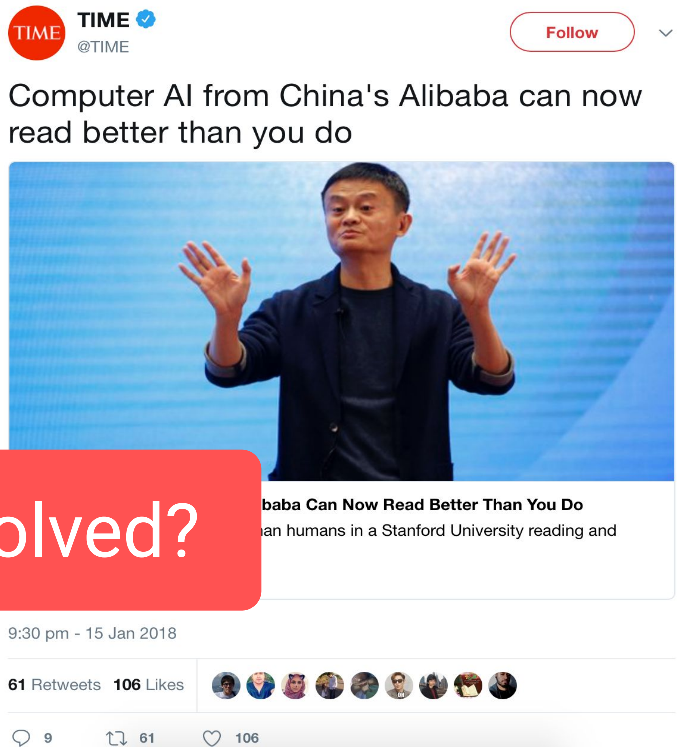
- architectures work well in practice

  … as long as we stay in-domain and questions are simple

Trends & Open Problems

Progression of SQuAD Model Performance





Reading Comprehension Solved?

QA System Demo

Machine Reading Demo

Beat-the-AI

Where RC models work well today

- question is answerable
- relevant paragraph / text is given
- 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

Upon closer look...

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.

Upon closer look...

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.

Upon closer look...

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.

Upon closer look...

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.

Upon closer look...

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. 2017)

Is all this model complexity necessary?

- single-layer BiLSTM with a simple word-in-question feature still very competitive on SQuAD (Weissenborn et al., 2017)

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?

Trends & Open Problems

Directions for Improving Robustness

Solvability

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

What was the name of the 1937 treaty?

[UNANSWERABLE]

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

From: Rajpurkar et al. 2018

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	**78.6**	**85.8**	**65.1**	**67.6**	**63.4**	**66.3**
Human	82.3	91.2	86.3	89.0	86.9	89.5
Human–Machine Gap	3.7	5.4	**21.2**	**21.4**	**23.5**	**23.2**

From: Rajpurkar et al. 2018

Adversarial Attacks: Oversensitivity

- Models are too sensitive (Snowflakes!)
- They change their predictions when the input “changes a little”
  - Appending Sentences (Jia et al. 2017)
  - Erasing words (Li et al. 2017)
  - Character flips (Ebrahimi et al. 2018)
  - Paraphrases (Iyyer et al. 2018, Ribeiro et al. 2018)

Adversarial Attacks: Undersensitivity

- Models are not sensitive enough (Feng et al., Ribeiro et al., Welbl et al, ongoing)
- They don't change their predictions when the input "changes a lot"

Question (original / adversarial)	Predicted Answer	Confidence
Who was the duke in the battle of Hastings? (original)	William the Conqueror	75.9%
Who was the duke in the expedition of Roger? (adv.)	William the Conqueror	99.8%
Who patronized the monks in Italy? (original)	Robert Guiscard	99.6%
Who patronized the monks in Grantmesnil? (adv.)	Robert Guiscard	99.8%

	Person		*Date*		*Numerical*	
	EM	F_1	EM	F_1	EM	F_1
BERT BASE	55.92	63.07	48.92	58.16	38.66	47.95
+ Robust Training	**59.06**	**66.55**	**58.44**	**65.61**	**48.72**	**58.89**

Model Diagnostics: Right for the Wrong Reason?

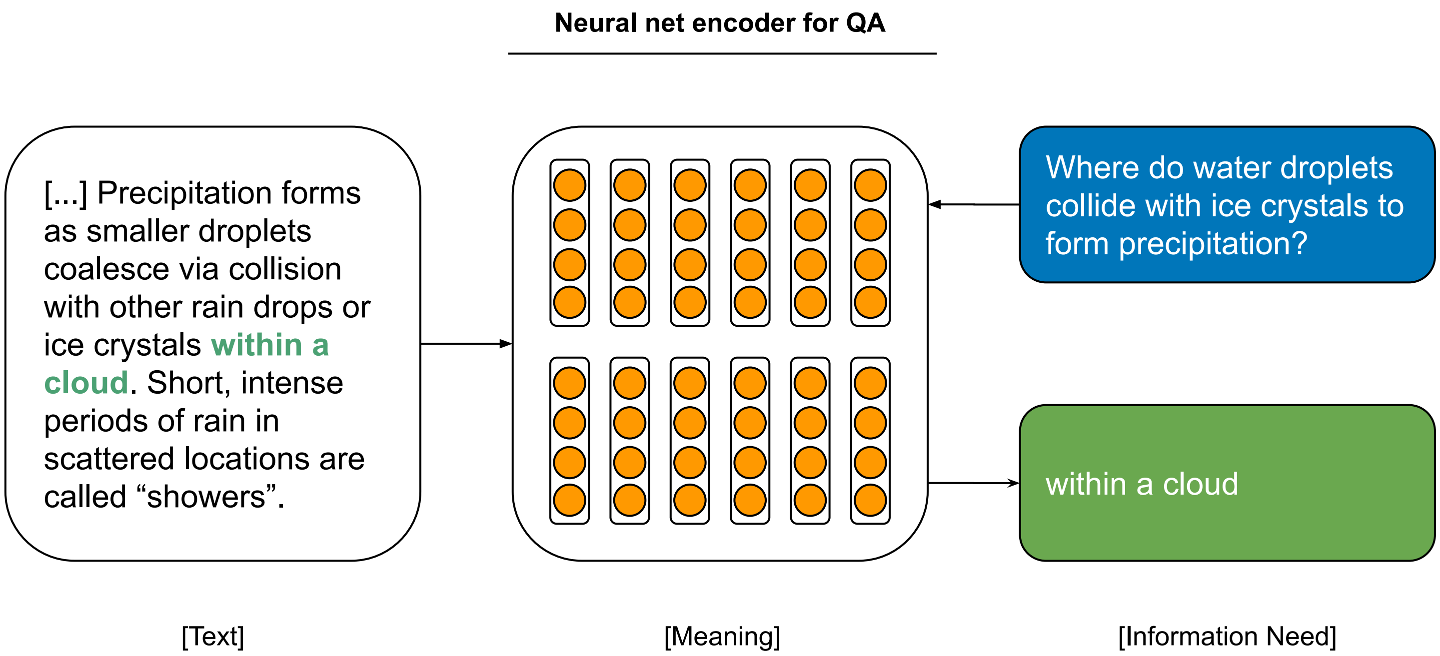
- What do models rely on to form predictions?
  - Analysing sensitivity to input: Ribeiro et al. (2016), Alvarez-Melis and Jaakkola (2017)
- Example: Anchors (Ribeiro et al. 2018)
  - Finding a minimal set of sufficient conditions to make a prediction

Anchor

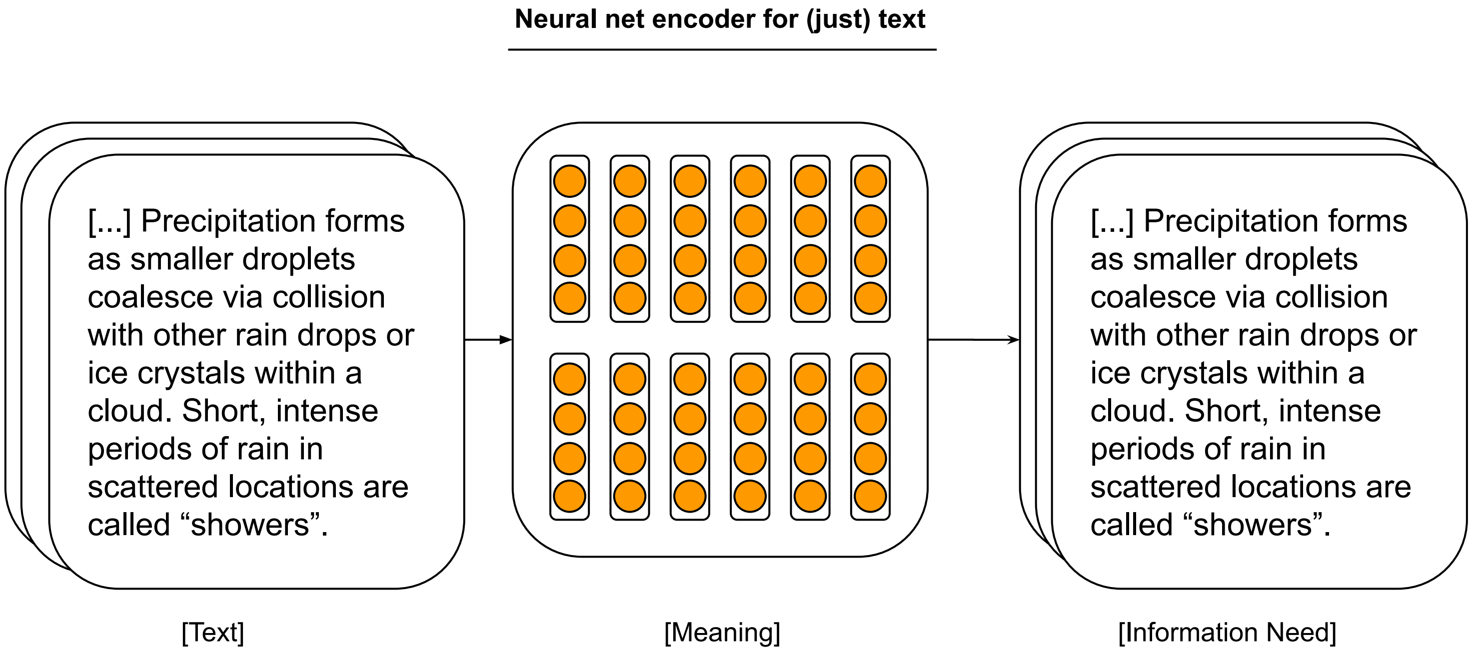
What is the mustache made of?	banana
What is the ground made of ?	banana
What is the bed made of ?	banana
What is this mustache ?	banana
What is the man made of?	banana
What is the picture of ?	banana

How **many** bananas are in the picture?	2
How **many** are in the picture?	2
many animals the picture ?	2
How **many** people are in the picture ?	2
How **many** zebras are in the picture ?	2
How **many** planes are on the picture ?	2

Pretraining Representations



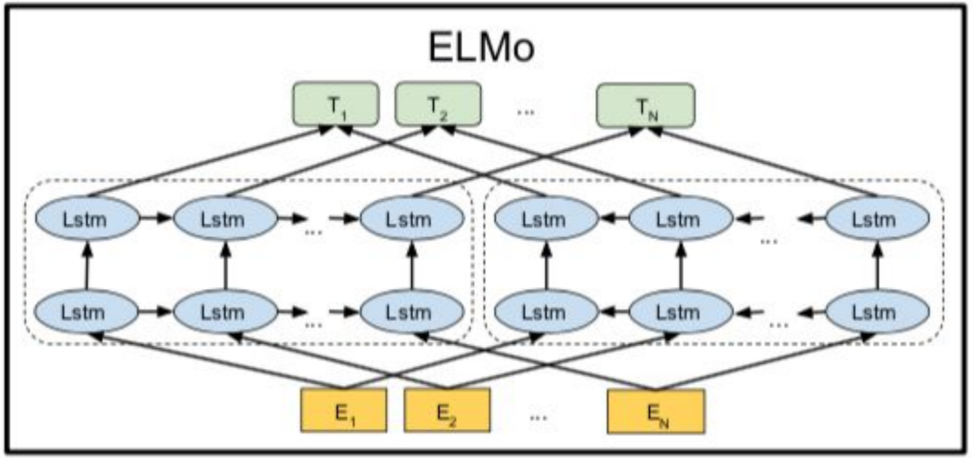
Pretraining Representations



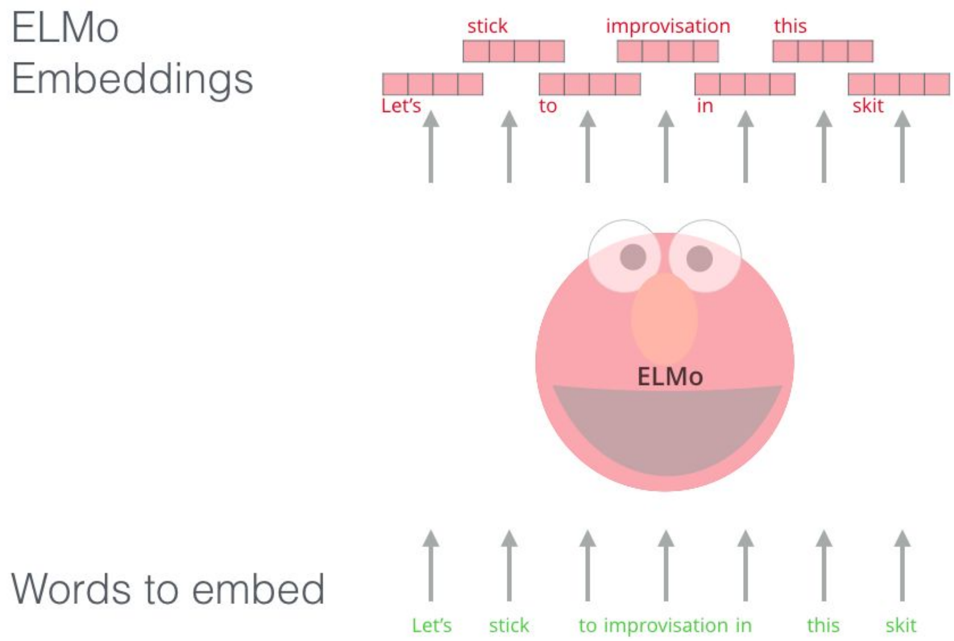
ELMo: **E**mbeddings from **L**anguage **Mo**dels

Peters et al., NAACL'18

- Train a BiLSTM for Bidirectional language modeling on a large dataset
- Run the sentence to encode through both forward and backward LSTMs
- Combine forward and backward representations into final contextual embeddings

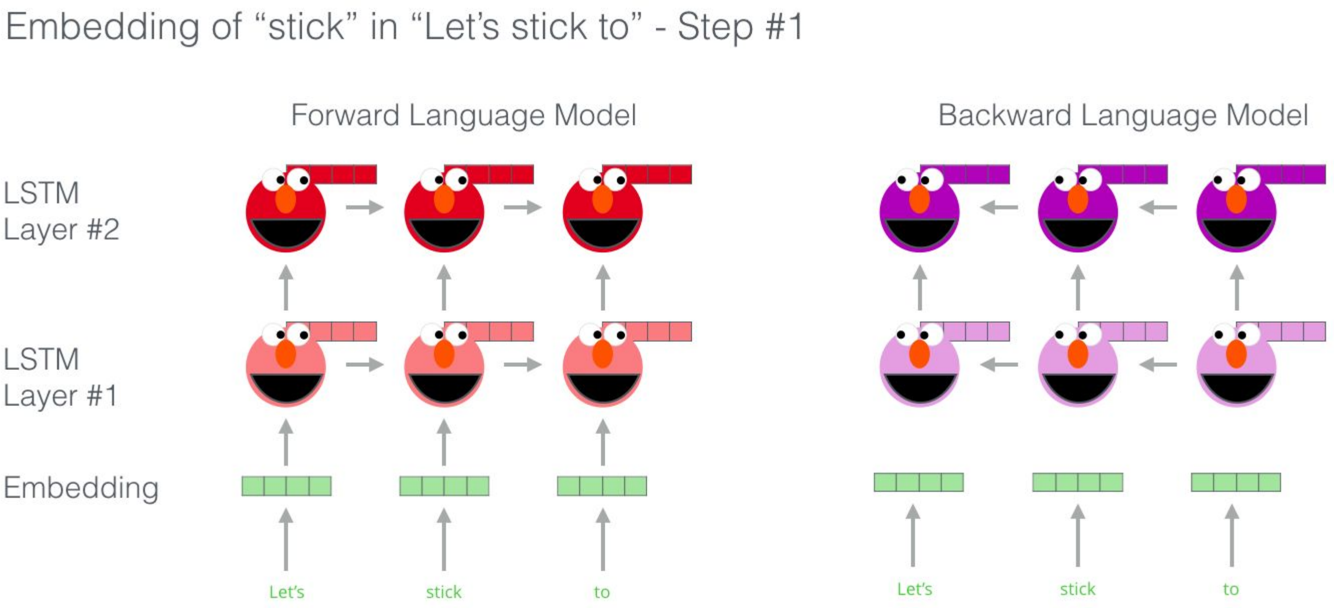


ELMo: **E**mbeddings from **L**anguage **Mo**dels



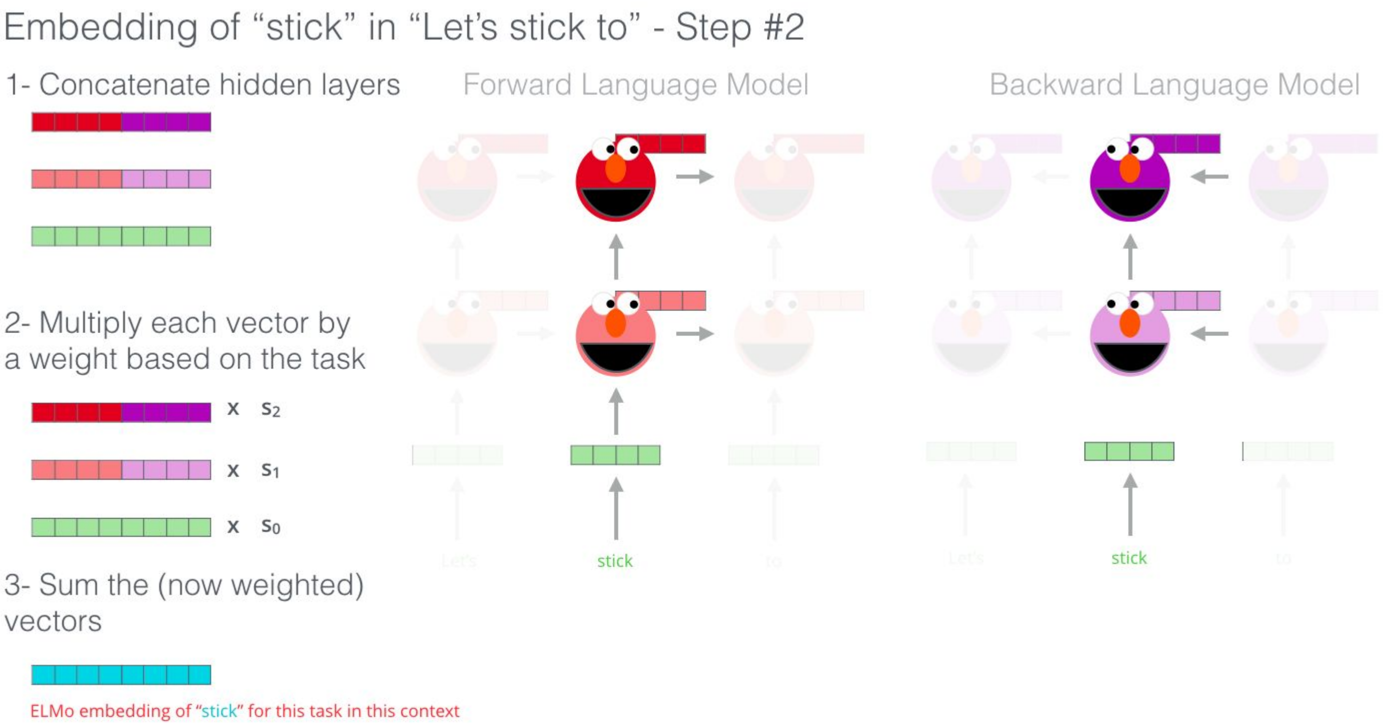
Figures from http://jalammar.github.io/illustrated-bert/

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



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

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



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

ELMo performance

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

What is ELMo learning ?

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

	Source	Nearest Neighbors
GloVe	play	playing, game, games, played, players, plays, player, Play, football, multiplayer
biLM	Chico Ruiz made a spectacular <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 .

Deals with variation and ambiguity

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!

How is this different from pretrained word embeddings?

Pretrained <u>Word</u> Embeddings (word2vec)

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

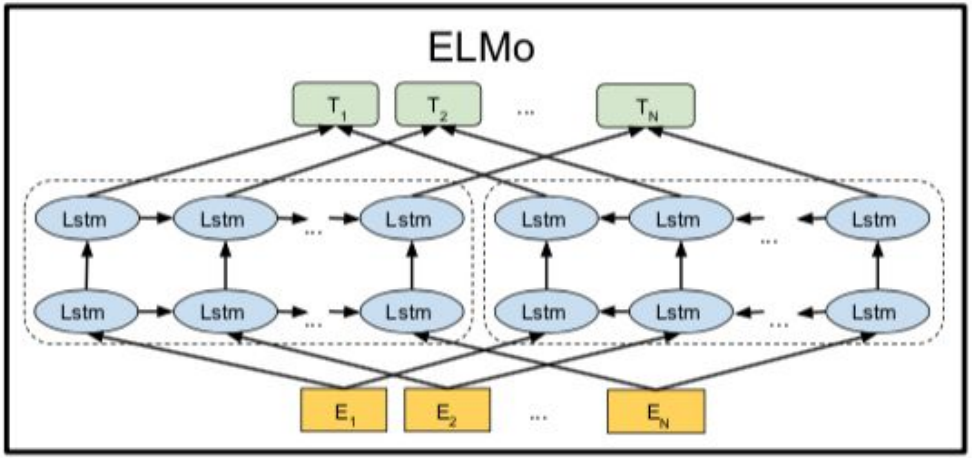
Pretrained Contextualized Embeddings (e.g. ELMo)

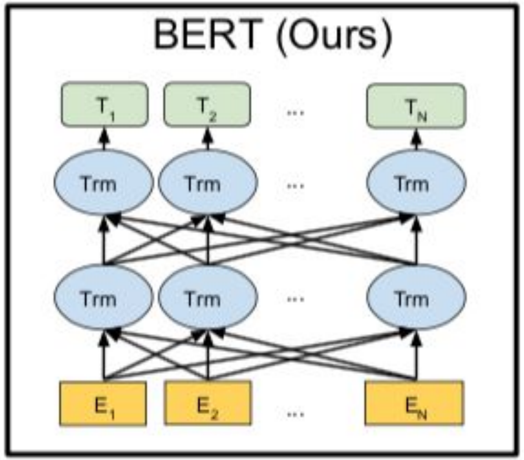
- Predicting whole text (using LSTM, or Self-Att.)
- Full dependence on other context

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

Devlin et al., NAACL'19

Uses Transformer instead of left-right decoder layers





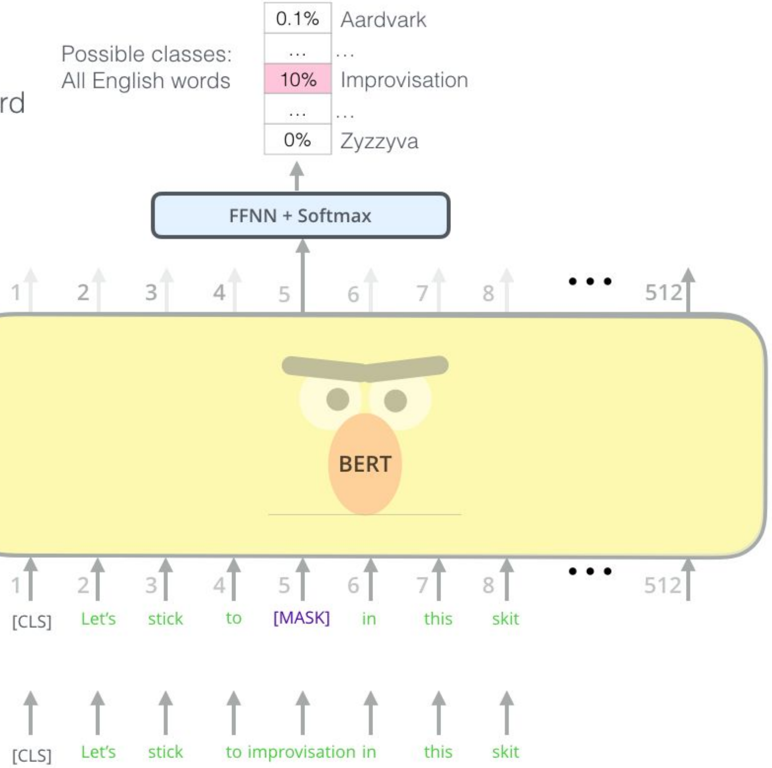
Innovation with multiple pretraining tasks

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

Use the output of the masked word's position to predict the masked word



Randomly mask 15% of tokens

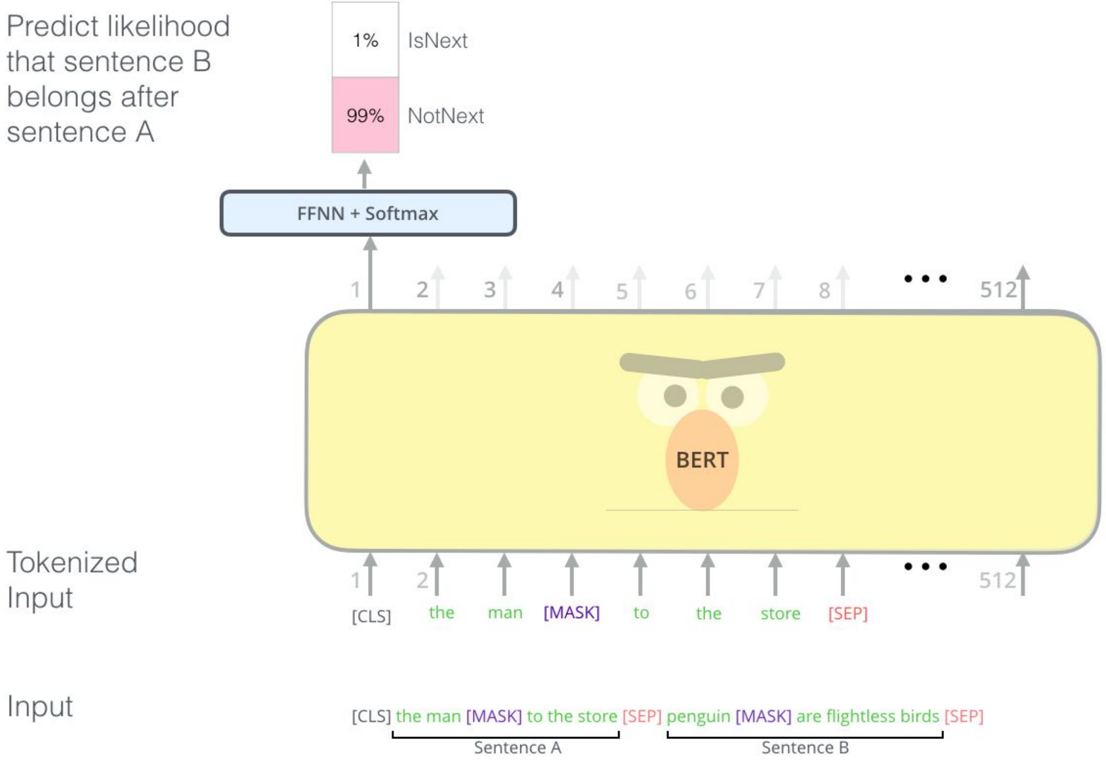
Input

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

BERT – Pretraining 2: next sentence prediction

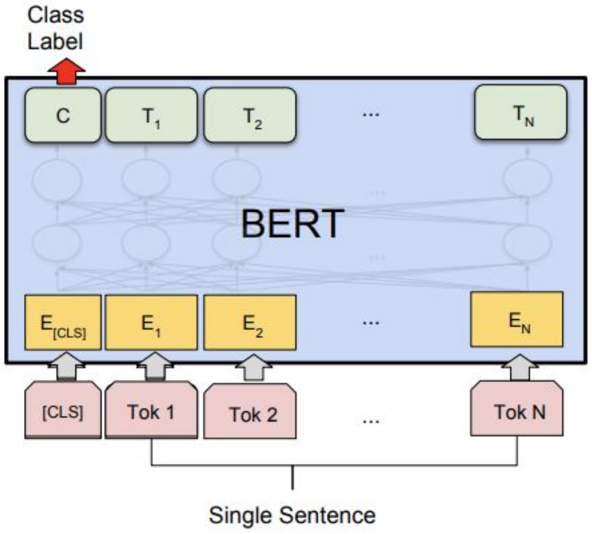
- Given two sentences, does the first follow the second?
- Teaches BERT about relationship between two sentences
- 50% of the time the actual next sentence, 50% random

BERT – Pretraining 2: next sentence prediction

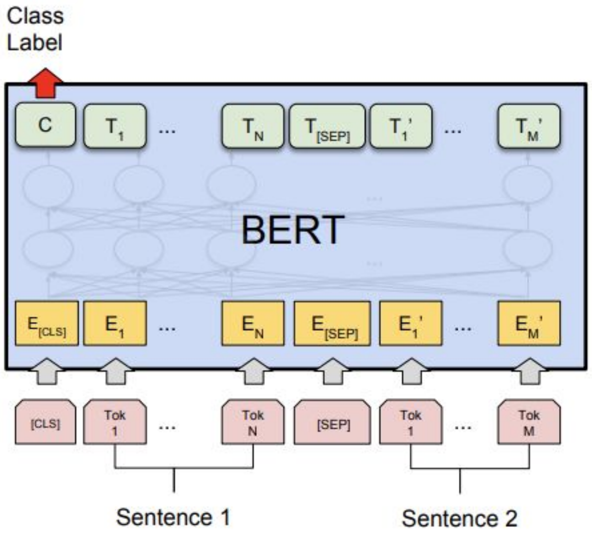


Figures from http://jalammar.github.io/illustrated-bert/

BERT – Fine-tuning for Classification



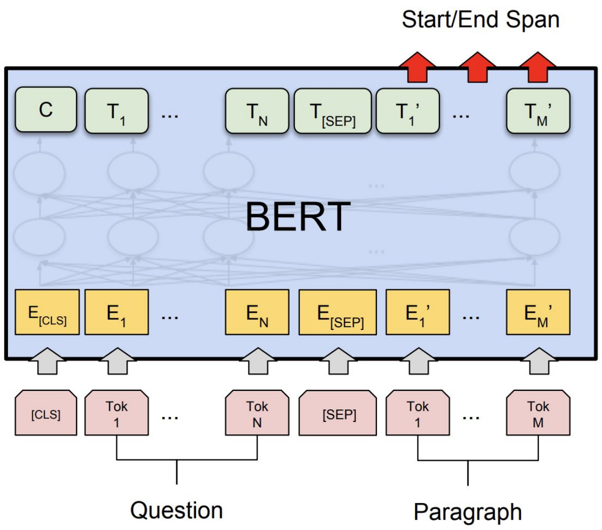
Single sentence classification
Sentiment analysis, spam detection, etc.



Pair of sentences classification
Entailment, paraphrase detection, etc.

Figures from Devlin et al. 18'

BERT – Fine-tuning for Machine Reading



(c) Question Answering Tasks:
SQuAD v1.1

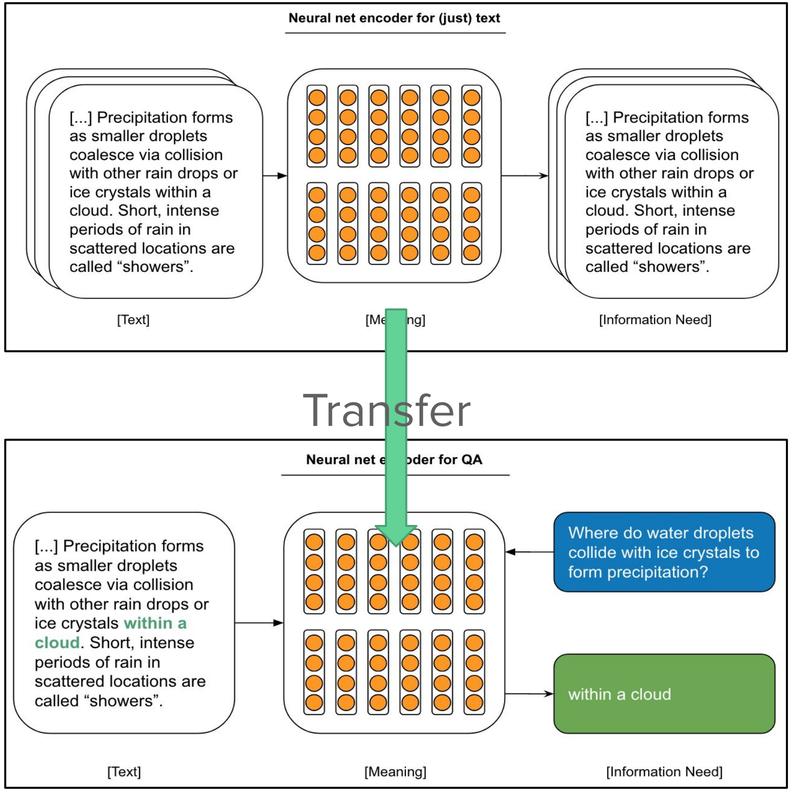
System	Dev EM	Dev F1	Test EM	Test F1
Leaderboard (Oct 8th, 2018)				
Human	-	-	82.3	91.2
#1 Ensemble - nlnet	-	-	86.0	91.7
#2 Ensemble - QANet	-	-	84.5	90.5
#1 Single - nlnet	-	-	83.5	90.1
#2 Single - QANet	-	-	82.5	89.3
Published				
BiDAF+ELMo (Single)	-	85.8	-	-
R.M. Reader (Single)	78.9	86.3	79.5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

Figures from Devlin et al. 18'

Lifting over Pretrained Representations

Pretrained Language Model

Document QA



Pretrained Sequence Encoders

… improve NLU tasks significantly!

- ELMo, *Peters et al. 2018. NAACL (Best Paper)*
  - pre-trained bi-directional LSTM language model
  - SQuAD (+4%), SRL (+3%), SNLI (+1.5%)
- Transformer LM, *Radford et al. 2018. arXiv.*
  - pre-trained language model based on pure self-attention (Vaswani et al., 2017)
- ULMFit, *Howard & Ruder 2018. ACL.*
  - pre-trained language model, fine-tuning on classification tasks
- CoVE, *McCann et al. 2017. NIPS.*
  - pre-trained LSTM encoder from Machine Translation
- Conneau et al. 2017
  - Pre-trained representations from Natural Language Inference

Other tasks?

Summary: Directions for Improving Model Robustness

- Task Refinement: being more precise in what to learn
- Diagnostics: shedding insight into model failure modes
- Adversarial training / regularization
- Better prior models for contextualised representations

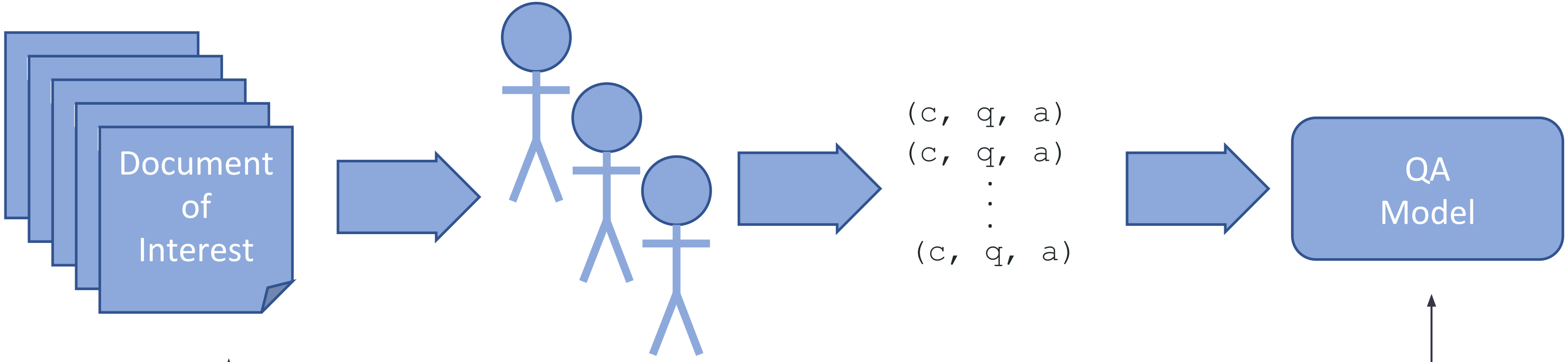
Trends & Open Problems

Other Challenges

Open Challenges I: Limited Supervision

- strong results with large annotated training sets
- How about smaller datasets?
  - Ideally: shift from 100K to 1K training points
  - less costly, large-scale annotation
- Approaches:
  - domain adaptation, e.g. Wiese et al. (2017)
  - Synthetic data generation, e.g. Dhingra et al. (2018)
  - transfer learning, e.g. Mihaylov et al. (2017)
  - (un?-)supervised pretraining, e.g. ELMo, Peters et al. (2018)
  - Unsupervised QA (Lewis et al 2019)

Unsupervised QA (Lewis et al, 2019)



New Domain?

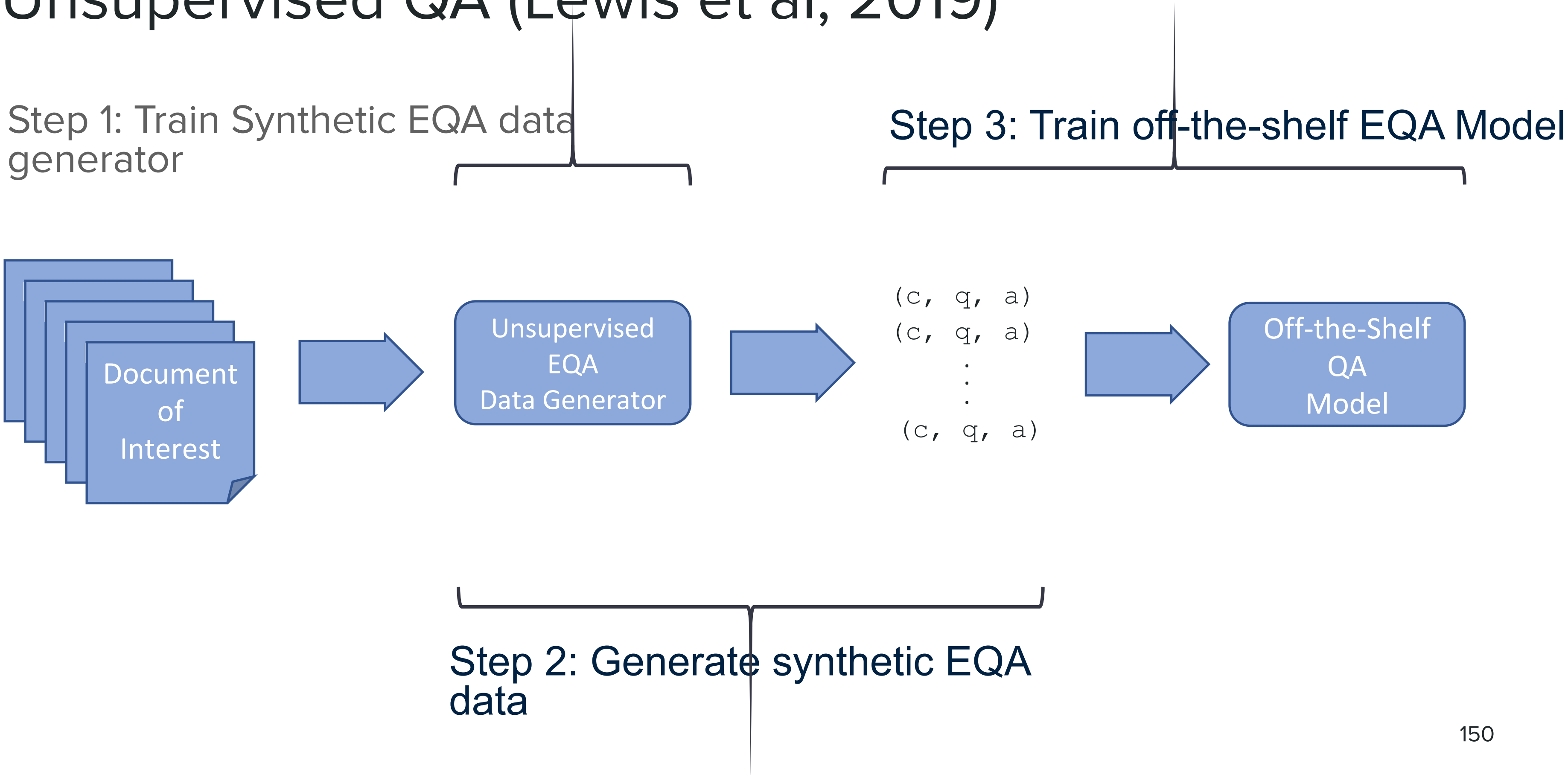
New Language?

Expensive, Difficult

New Domain?

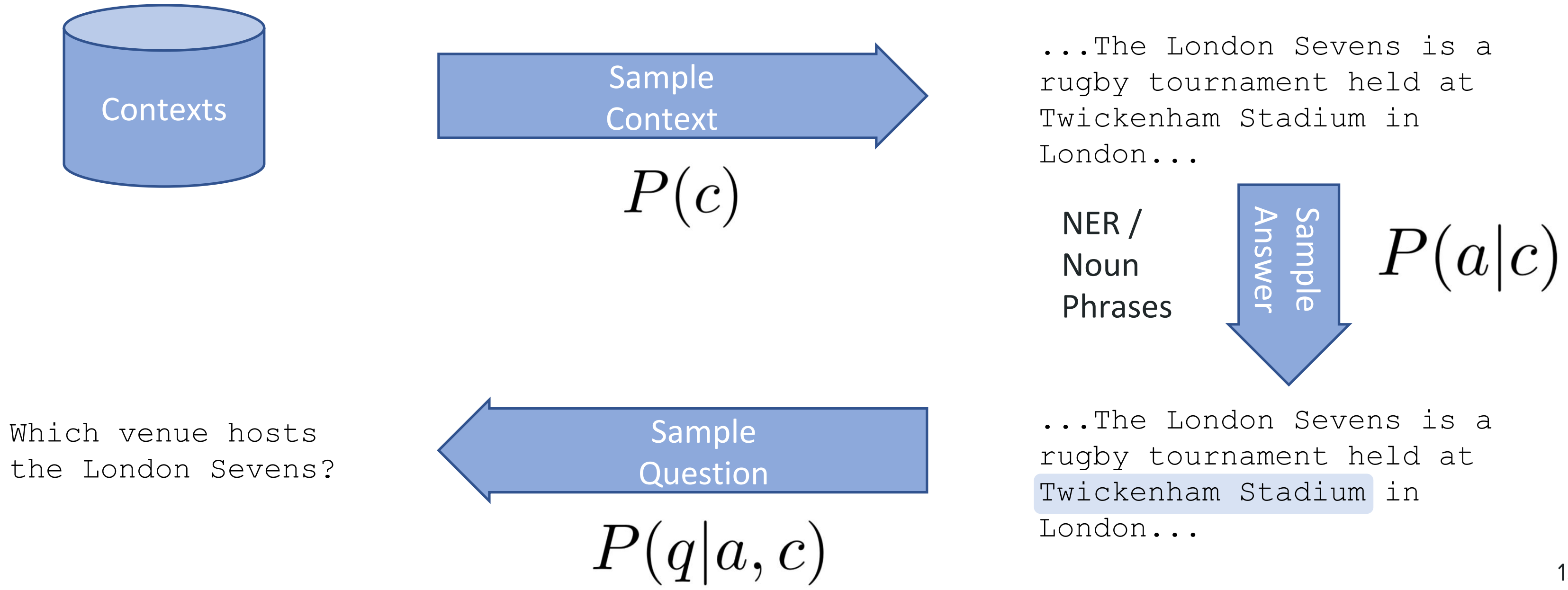
What if we could train unsupervised?

Unsupervised QA (Lewis et al, 2019)



Unsupervised QA (Lewis et al, 2019)

$$P(c, q, a) =$$



Unsupervised QA (Lewis et al, 2019)

Cloze Question q'

The London Sevens is a rugby tournament held at ____ in London.

Question Translation

Clozify

Which venue hosts the London Sevens?

...The London Sevens is a rugby tournament held at Twickenham Stadium in London...

Cloze Translation

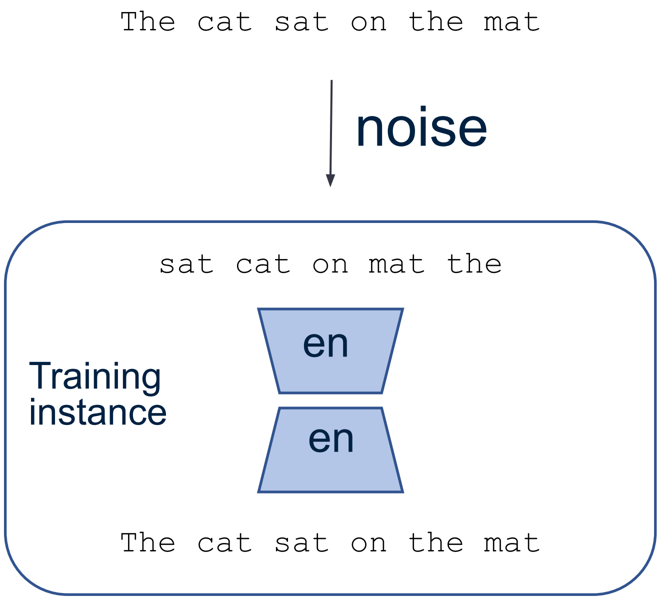
$$(c, q', a) \rightarrow (c, q, a)$$

- Naïve Baseline: Identity Cloze
  `The London Sevens is a rugby tournament held at _________ in London.`
- Hard baseline: Noisy Cloze
  `Where Sevens London BLANK rugby a held London ?`
- Rule-based: Statement-to-question [1]
  `Where was the London Sevens held ?`
- Unsupervised NMT [2]
  `Where is The London Sevens rugby tournament held at London ?`

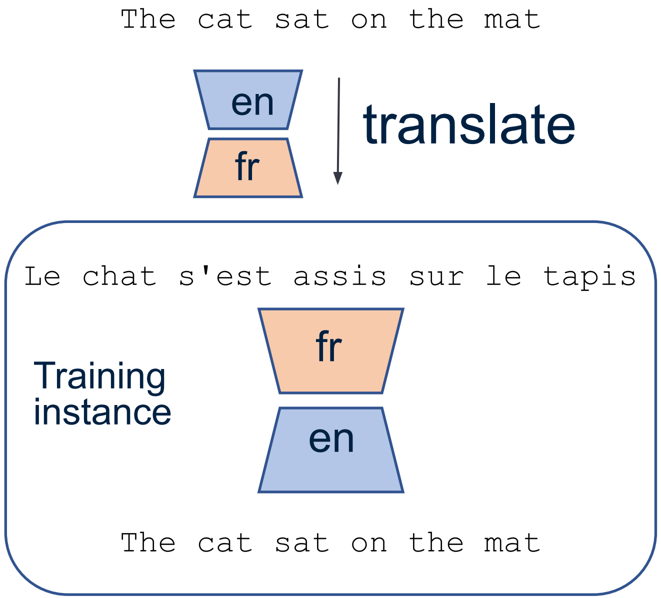
[1] M Heilman and N Smith, 2010
[2] G Lample et al. 2018

Unsupervised Machine Translation

Auto-encoder

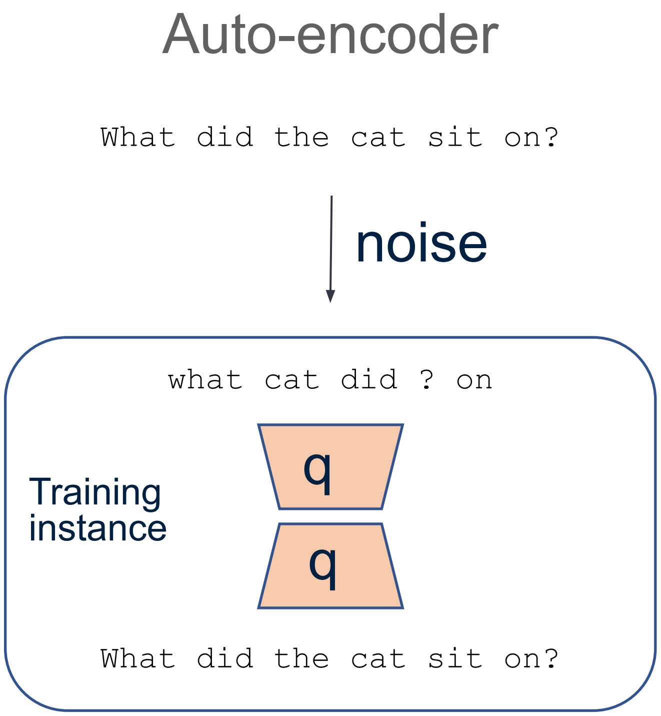


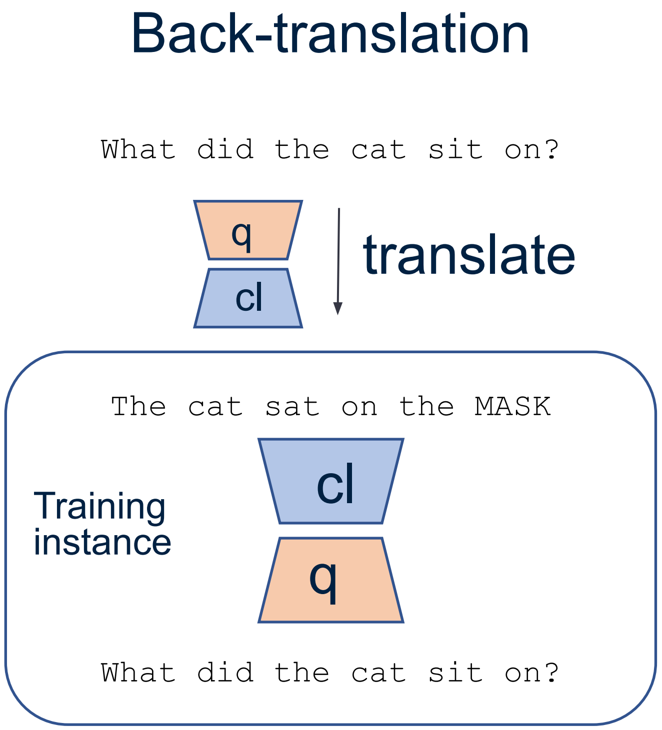
Back-translation



G Lample et al. 2018

Unsupervised Neural Cloze Translation

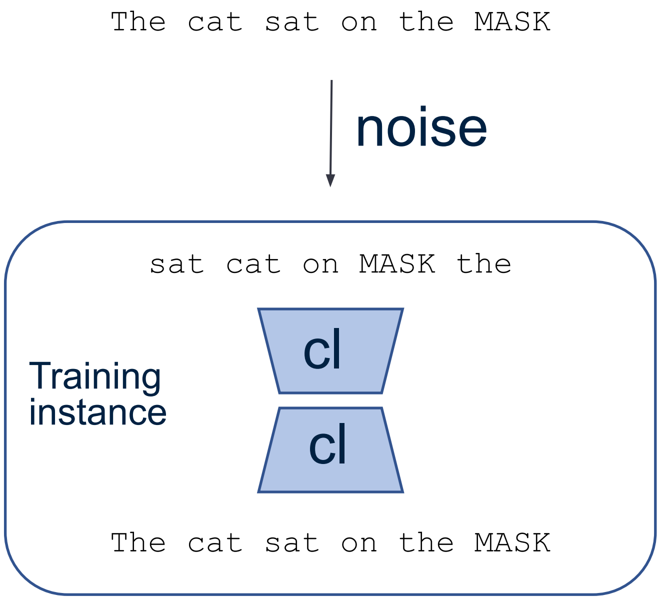




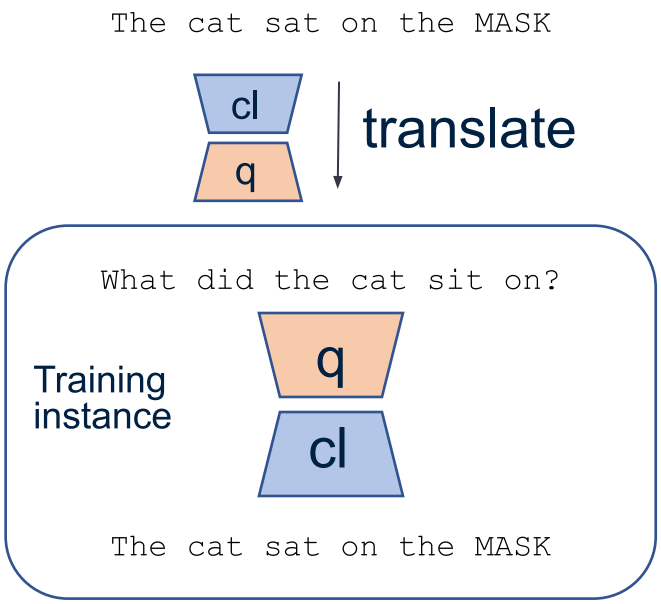
G Lample et al. 2018

Unsupervised Neural Cloze Translation

Auto-encoder



Back-translation



G Lample et al. 2018

Recent Work in Unsupervised QA

- ***Dhingra et al. 2018***: Generate (cloze question, context, answer) triples for EQA semi-supervision, also publish unsupervised setting
- ***Radford et al. 2019***: GPT-2, evaluate various zero-shot tasks including QA
- ***Chan et al. 2019***: KERMIT, evaluate on zero-shot cloze QA

Experiments

- Evaluate EQA performance without explicit supervision
- Explore impact of design decisions of data generator

- Context Generator: Paragraphs from English Wikipedia
- Cloze Question boundary: Sentence or sub-clause with “S” label
- 5M questions mined from common crawl, 5M clozes mined from Wikipedia
- Question Answering: BiDAF + Self-attention [1] and fine-tuning BERT [2]

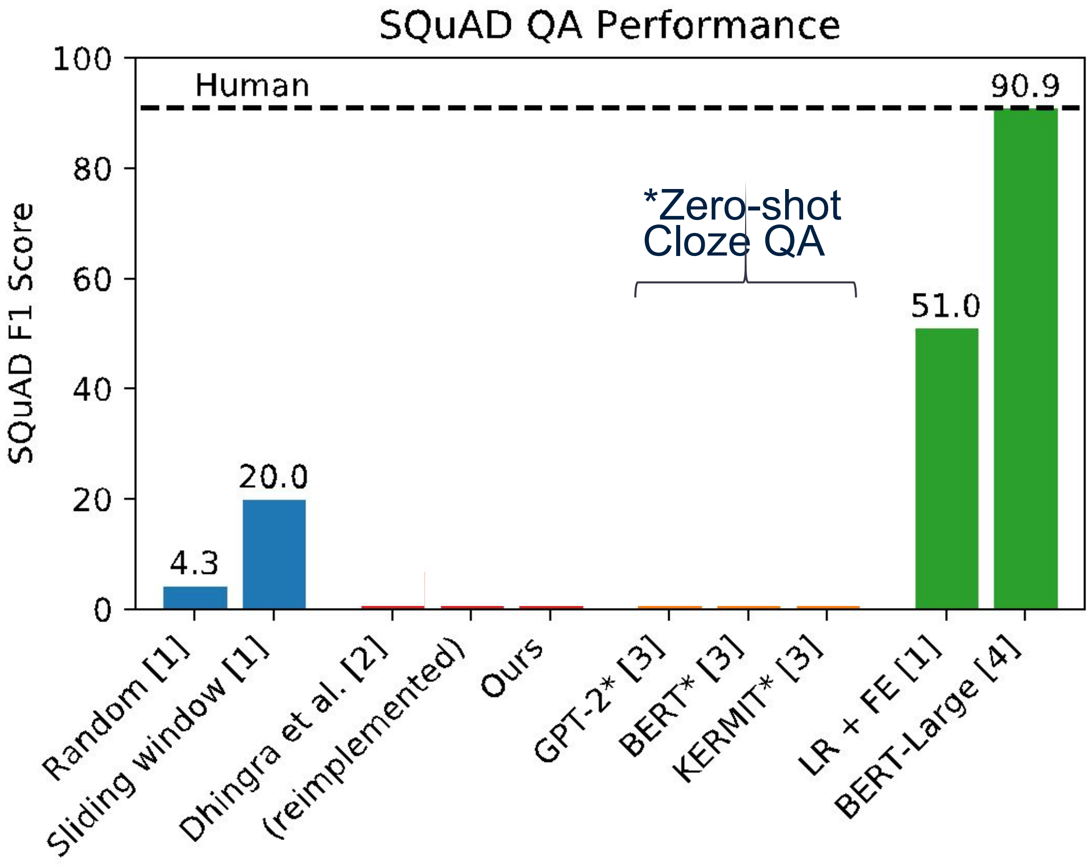
[1] J Devlin et al. 2019
[2] C Clark and M Gardner

Translation Examples

Cloze Question	Answer	Translated Question
WALA would be sold to the Des Moines-based **ORG** for $86 million	Meredith Corp	Who would buy the WALA Des Moines-based for $86 million?
The **NUMERIC** on Orchard Street remained open until 2009	second	How much longer did Orchard Street remain open until 2009?
he speaks **LANGUAGE**, English, and German	Spanish	What are we , English , and German?
Form a larger Mid-Ulster District Council in **TEMPORAL**	August	When is a larger Mid-Ulster District Council?
Form a larger Mid-Ulster District Council in **TEMPORAL**	August	When will a larger Mid-Ulster District Council be formed?

Results

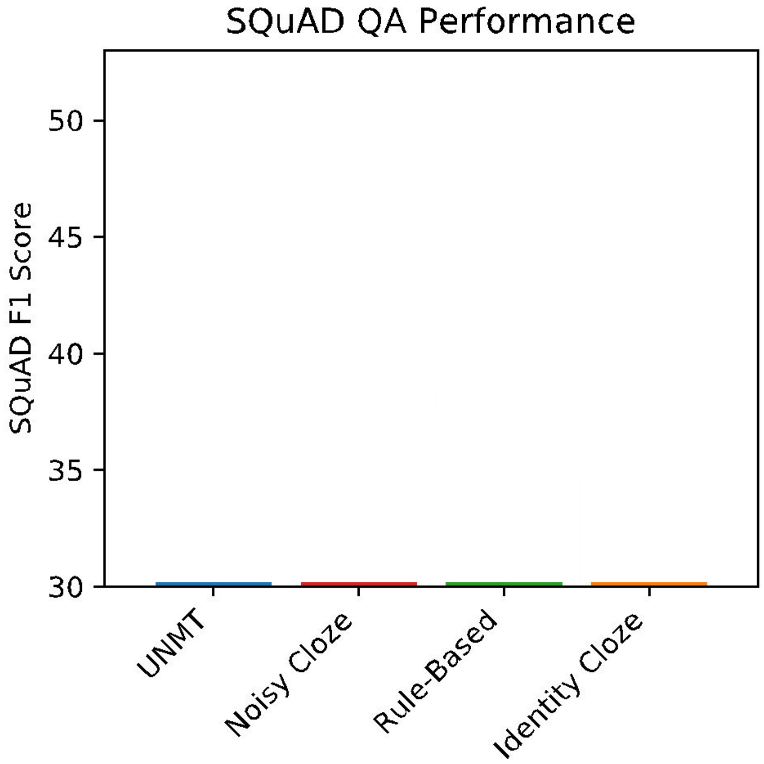
- Best results with:
  - Named entity answers
  - Unsupervised NMT questions
  - Wh* heuristic
  - Sub-clause boundaries
  - Bert-Large QA



[1] P Rajpurkar et al. 2016 [2] B Dhingra et al. 2018
[3] W Chan et al. 2019 [4] J Devlin et al. 2019

Question Translators

- UNMT ***best performing***
- But Noisy cloze ***competitive***
    - *Why?*
- *Rule-based [1]* ***lower than expected***
    - *Why?*



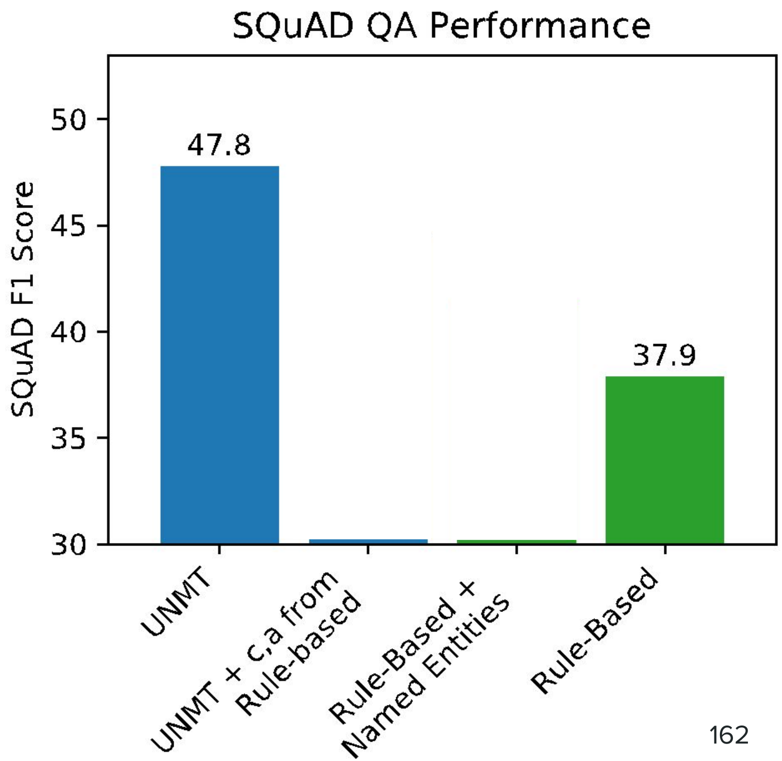
[1] M Heilman and N Smith. 2010

Question Translators (2)

- Does Rule-based system generate less variety of questions?
- Restrict ***our*** system to use contexts and answers from ***rule-based***

  ➪ -3.1 F1
- Is the answer distribution for rule-based system mismatched?
- Restrict ***rule-based*** system to use answers from ***our*** system

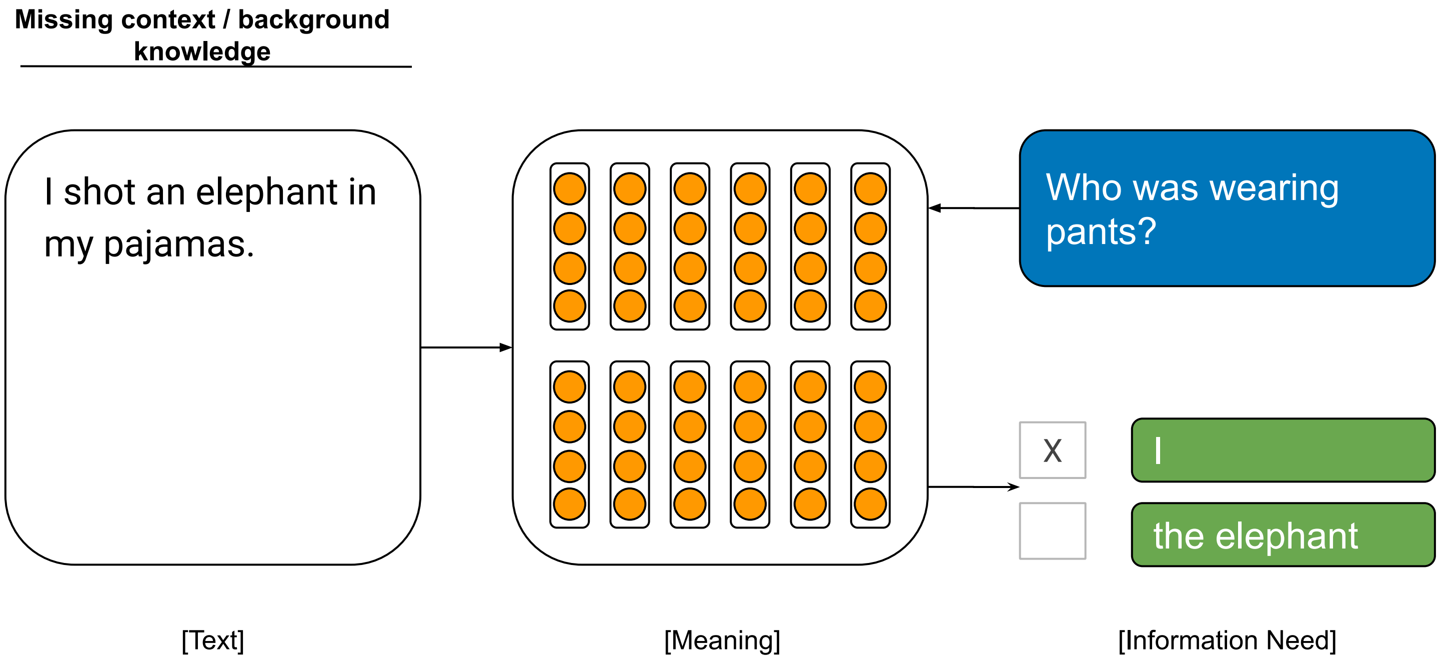
  ➪ +3.6 F1



Conclusion

- Outperform simple supervised models without explicit supervision
- Much scope for future work:
  - Questions without Answers
  - “Multi-hop” Questions
  - Other Question Answering tasks
- 4M UQA training datapoints, and code
  github.com/facebookresearch/UnsupervisedQA

Challenge II: Integrating Background Knowledge



Challenge II: Integrating Background Knowledge

- Approaches for leveraging common sense knowledge
  - Encyclopedic descriptions (Hill et al. 2016, Bahdanau et al. 2018)
  - Knowledge Bases (Yang and Mitchell 2017, Weissenborn et al. 2017, Mihaylov and Frank, 2018)
  - Example: Weissenborn et al. (2017):
    - condition context representations also on additional facts
    - Intuition: new background facts provide additional features ➜ refined vector representations

Challenge III: Integration of MR with Vision

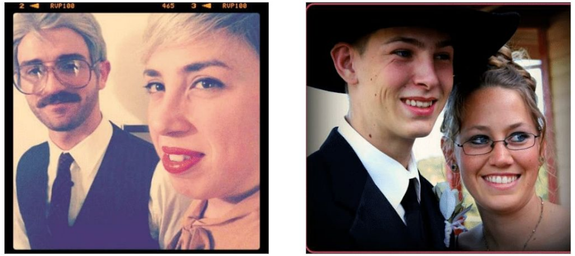
- End-to-end trainable encoders for questions, text
- Example: Visual QA

Question Complexity

Task	Question
Object Recognition	What is in the image?
Segmentation	Where are the things in the image?
Scene / Attribute Classification	What color is the dress?
Visual QA	Arbitrary Question

Who is wearing glasses?

man woman



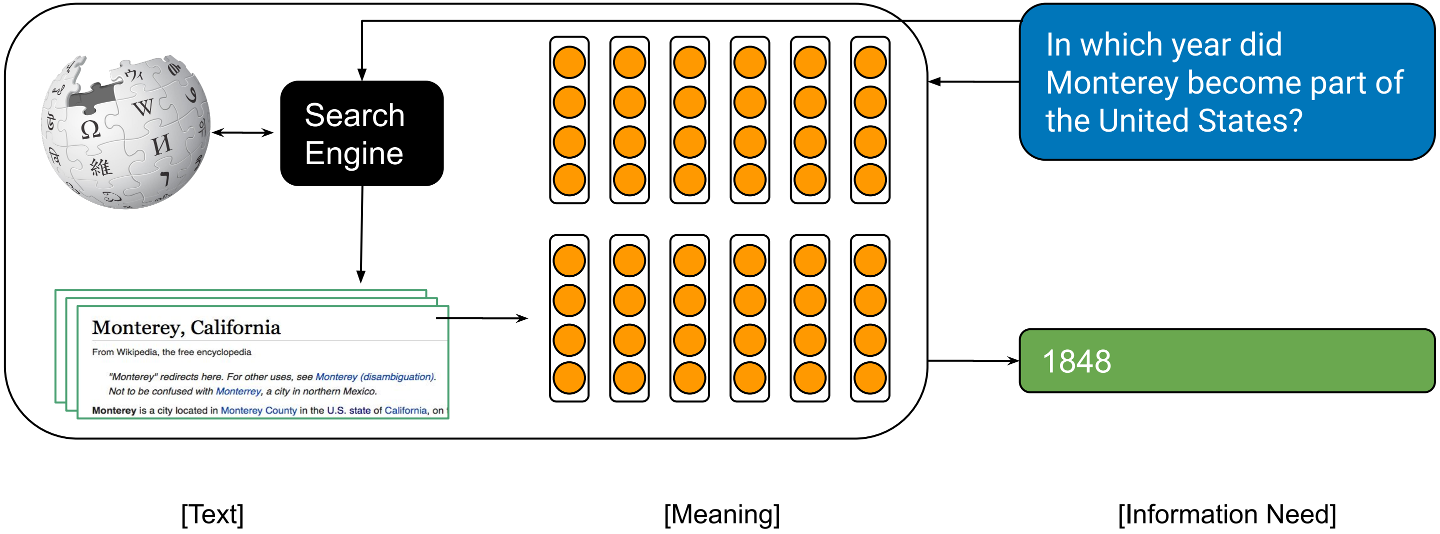
Is the umbrella upside down?

yes no

From: Goyal et al. (2017)

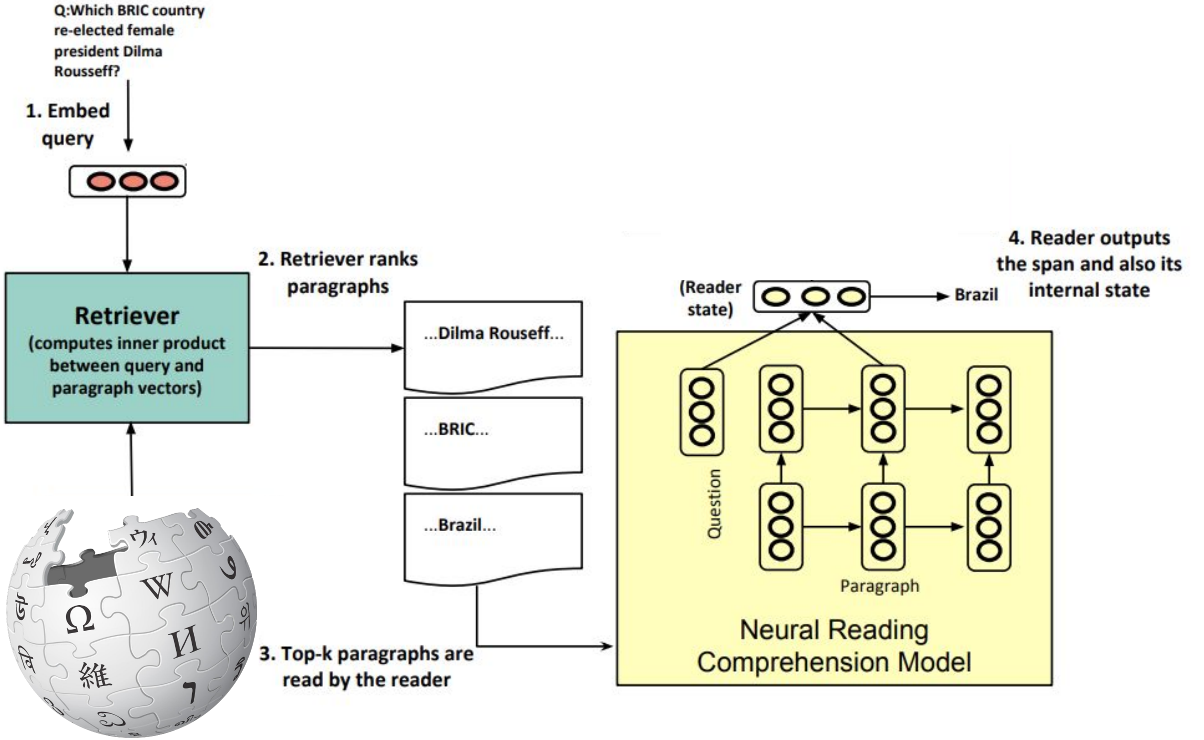
Challenge IV: End-to-End Machine Reading at Scale

Open-domain Question Answering, e.g. Chen et al. (2017)



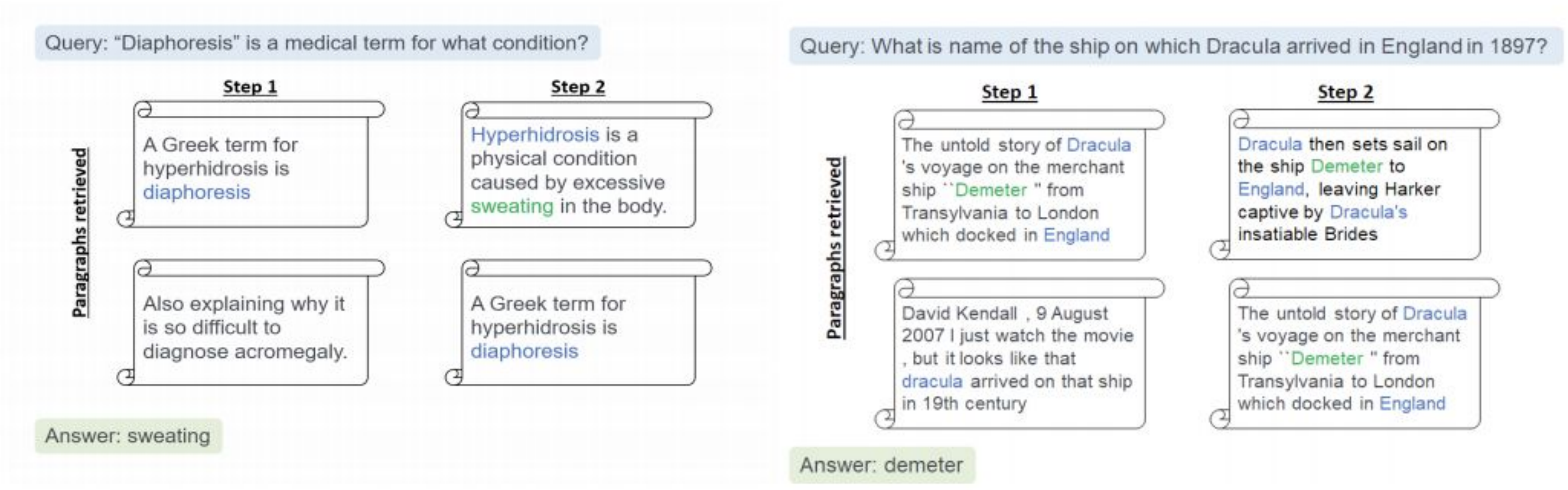
Current best: Multi-Step Retriever-Reader

Das et al., 2019



Current best: Multi-Step Retriever-Reader

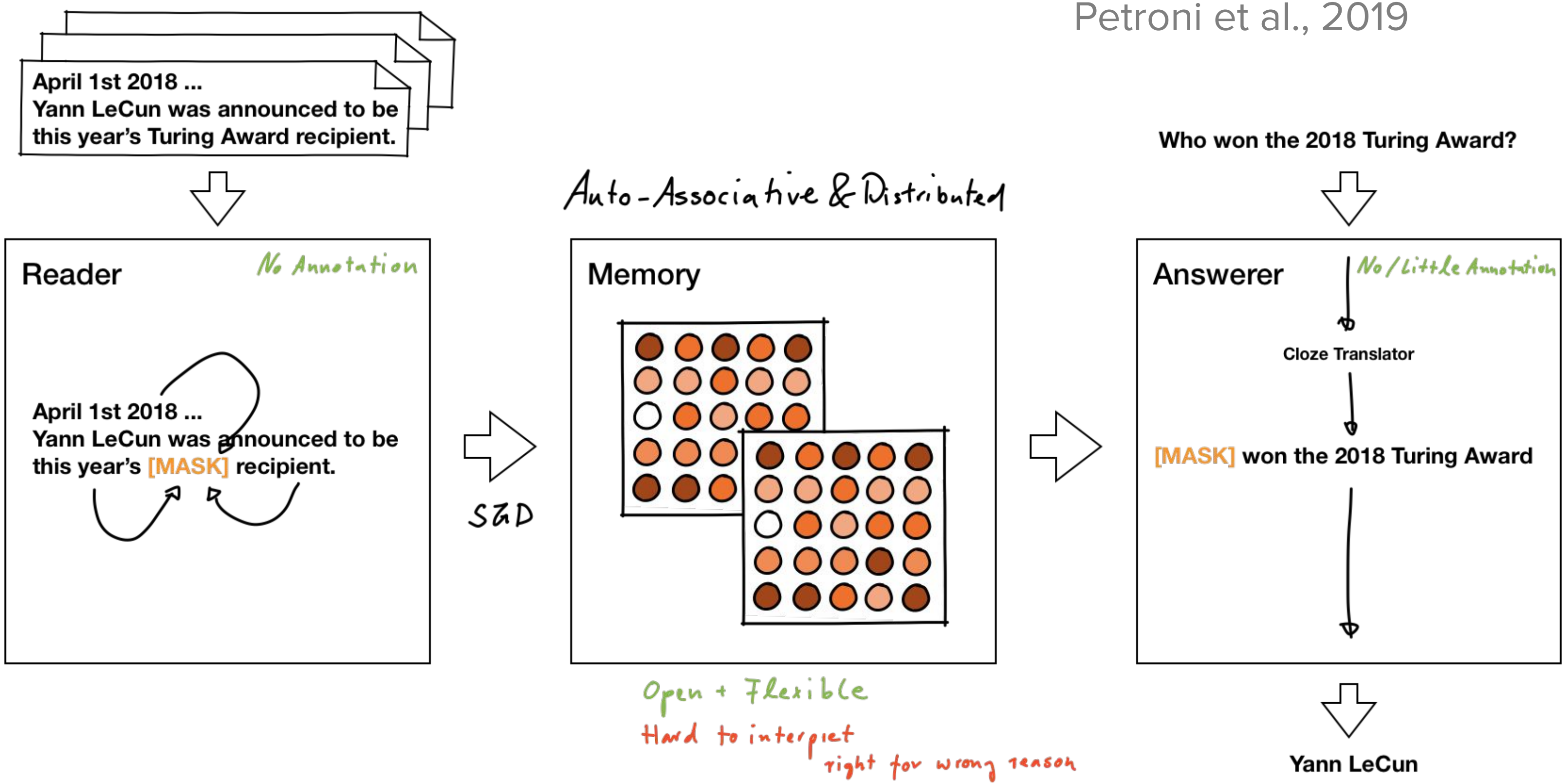
Das et al., 2019



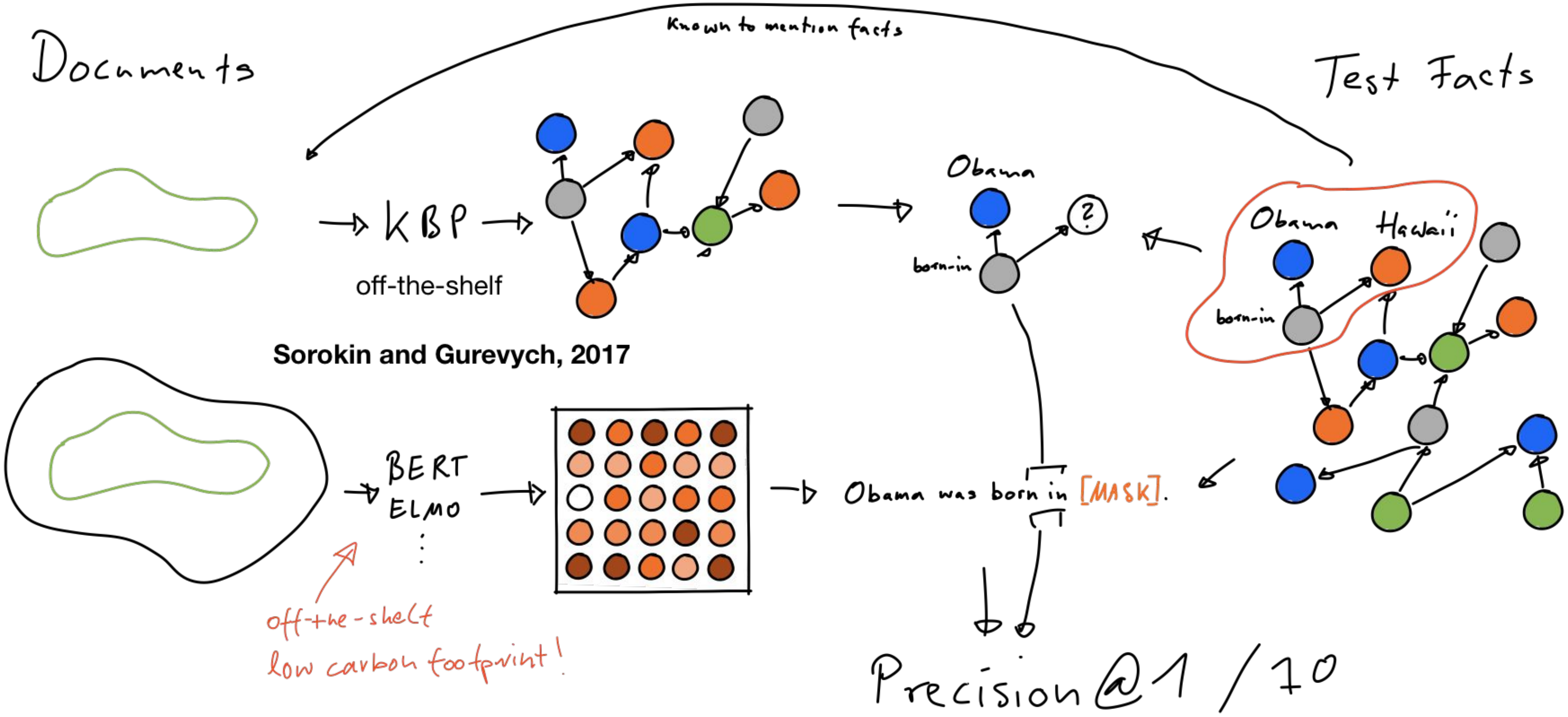
Between 40 and 60% of correct responses (for rather simple questions)

Language Models as Machine Readers (at super scale)?

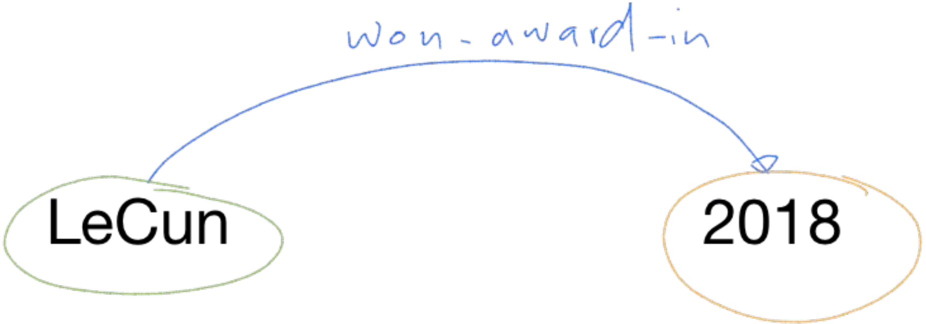
Petroni et al., 2019



Testing what the Language Model knows



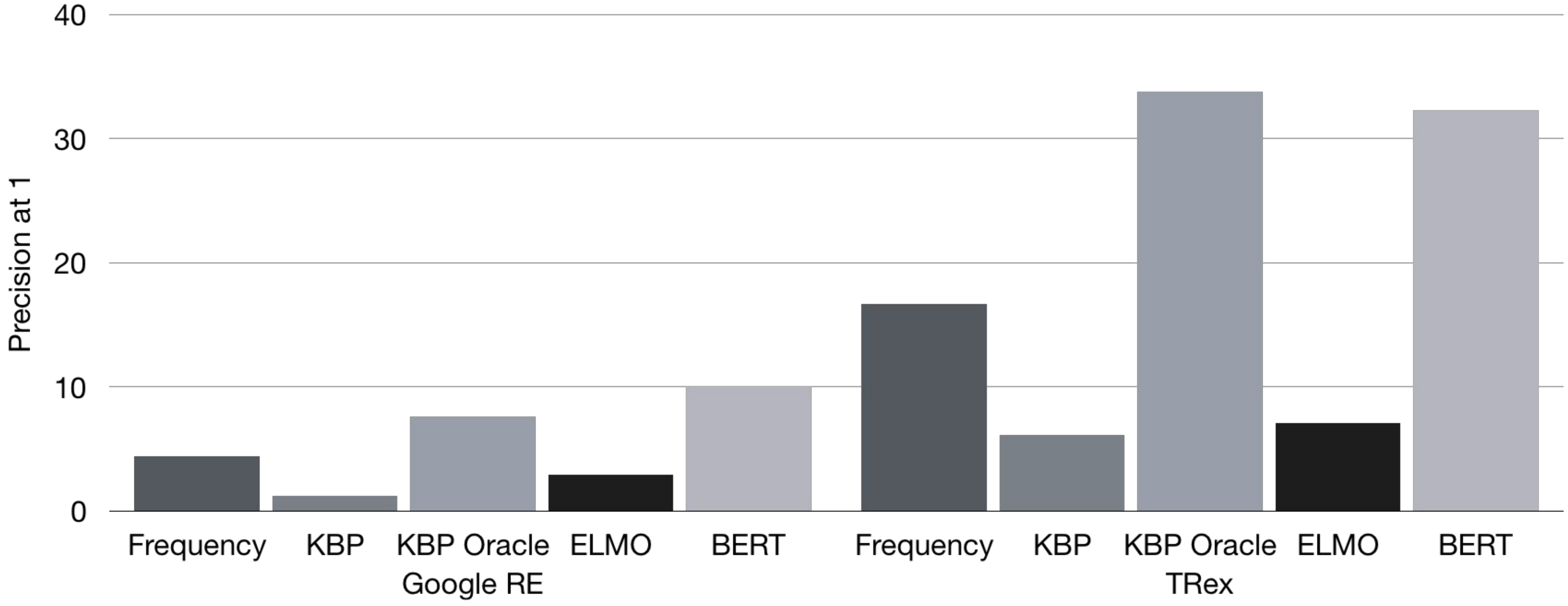
Traditional Machine Reading “Oracle”



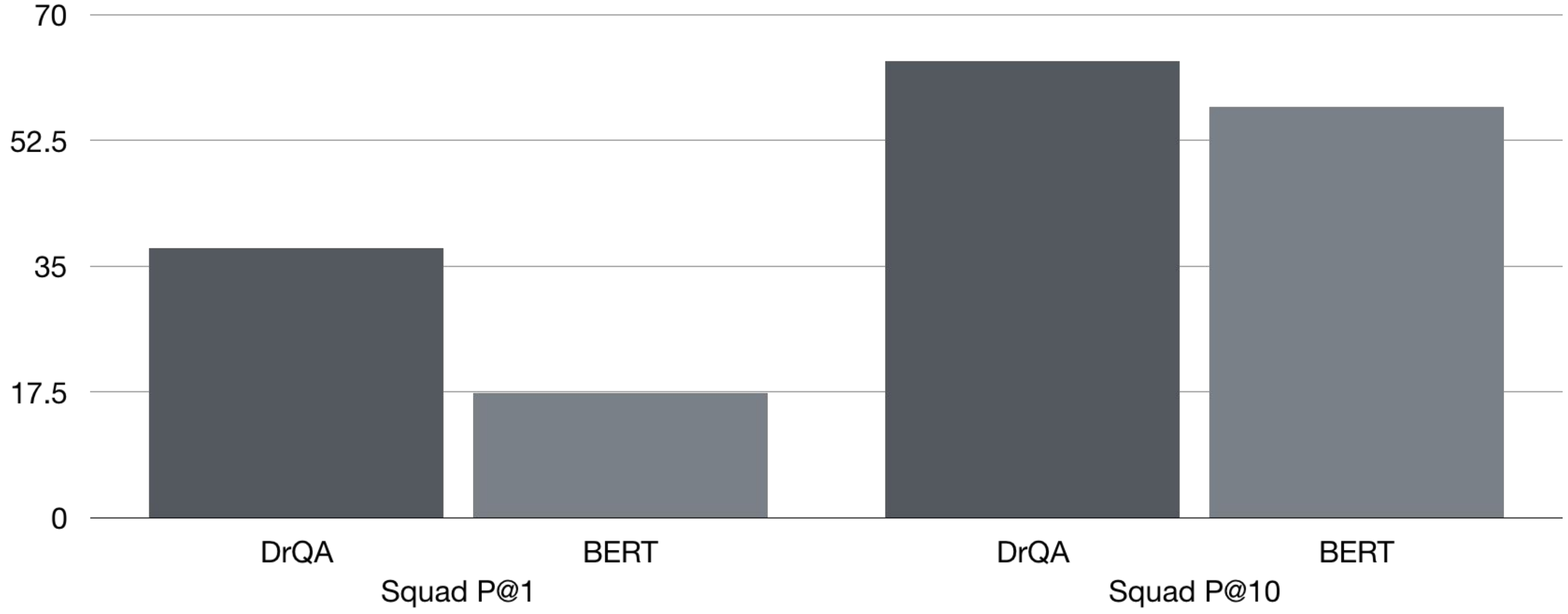
won-award-in

Yann LeCun was announced to be a 2018 Turing Award recipient.

Results on predicting Facts



Results on question answering



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?

Challenge V: 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*

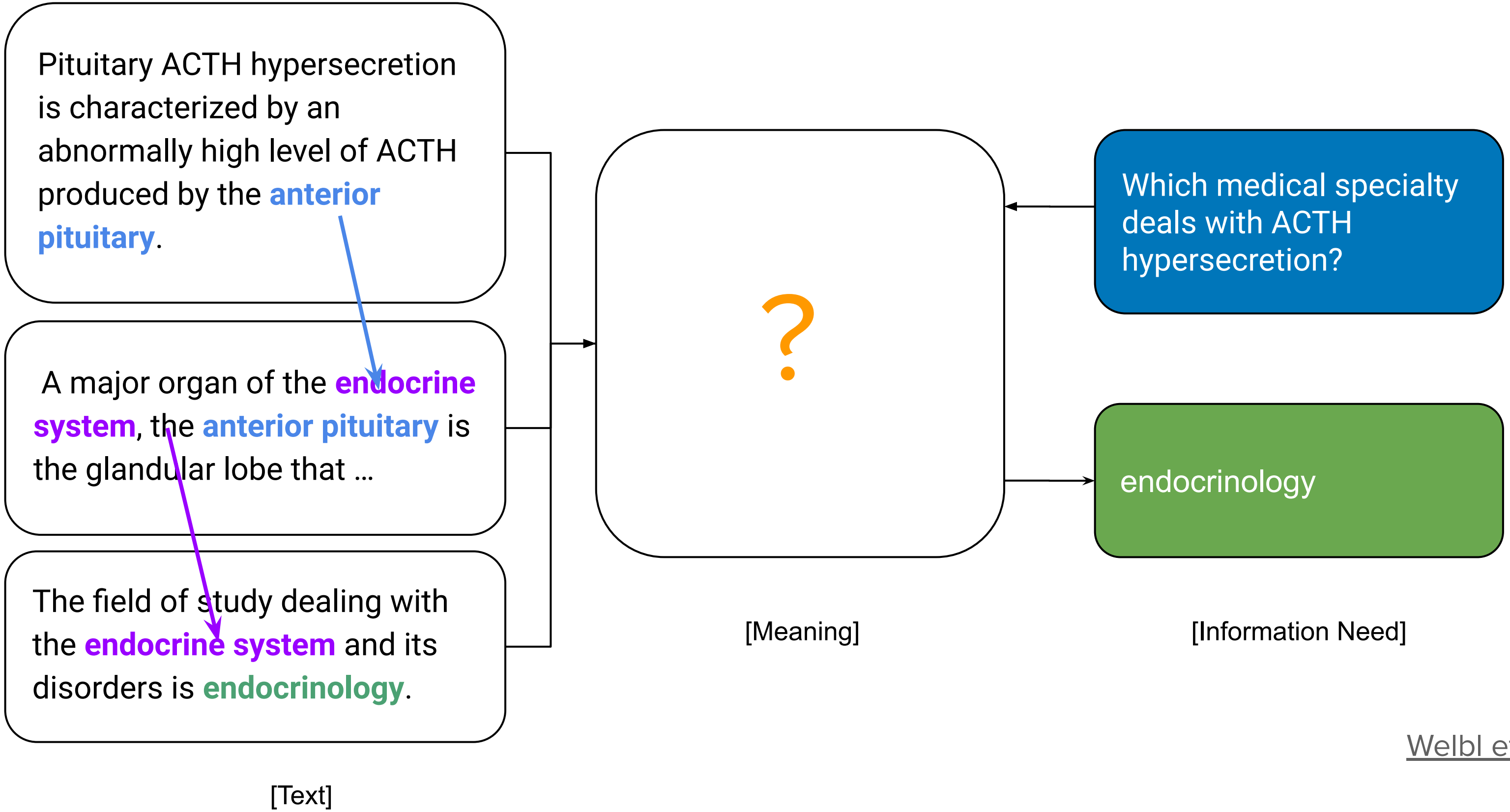
"We are not giving £20bn a year or £350m a week to Brussels - Britain pays **£276m** a week to the EU budget because of the rebate." *BBC Reality Check*

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

?

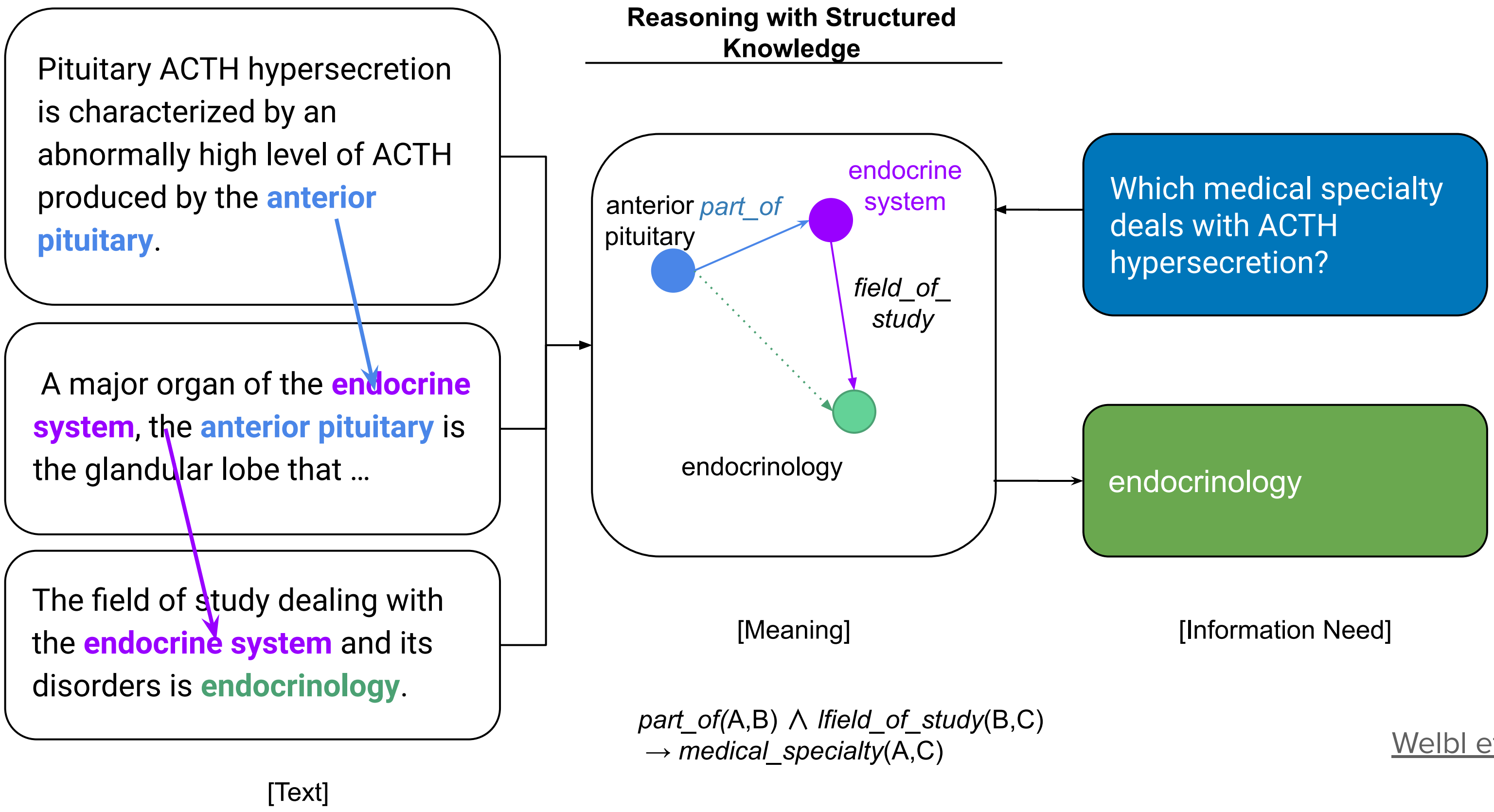
Trust into source, timeline, ...

Challenge VI: Reasoning with Text



Welbl et al. (2018)

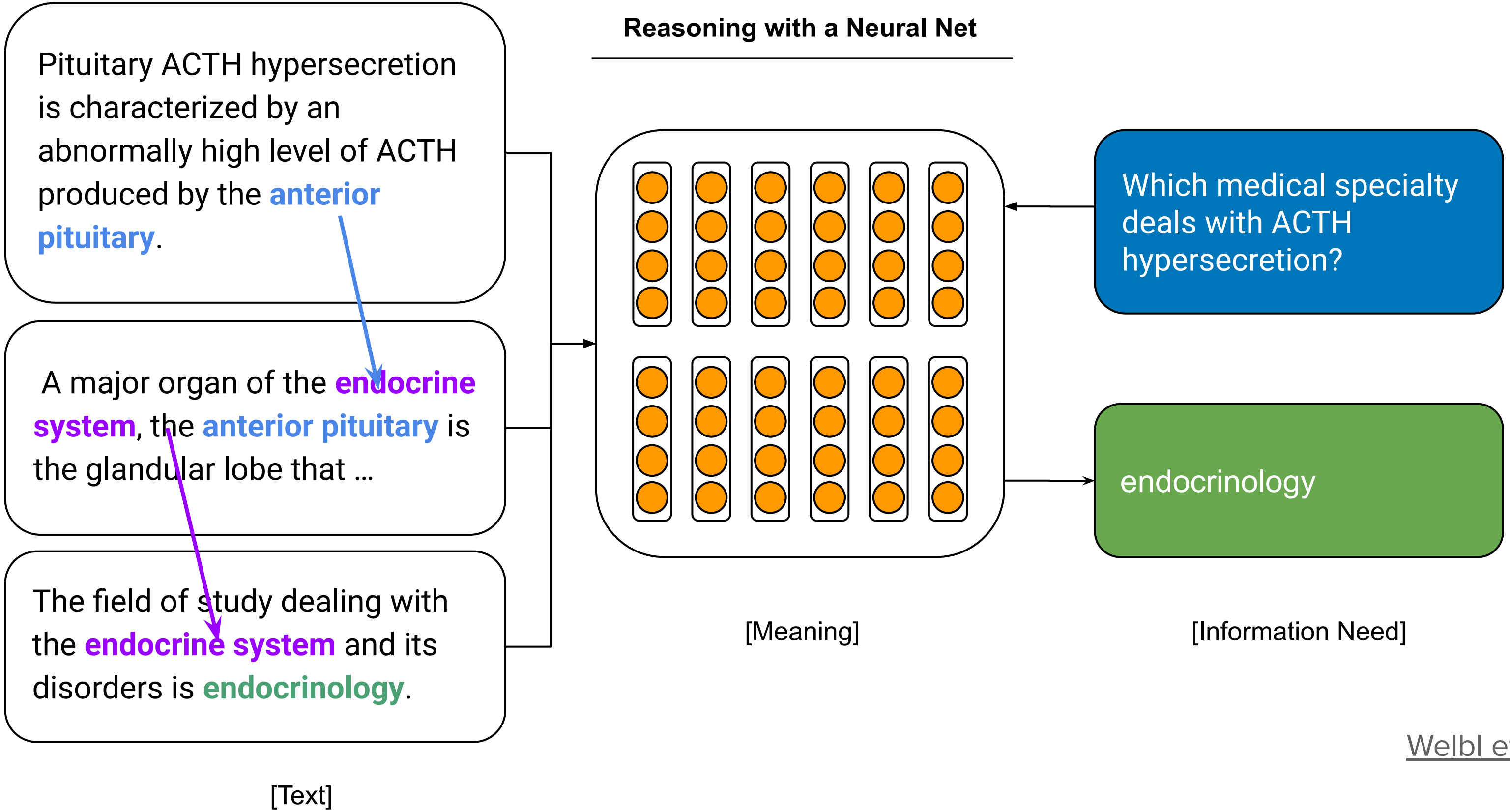
Challenge VI: Reasoning with Text



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

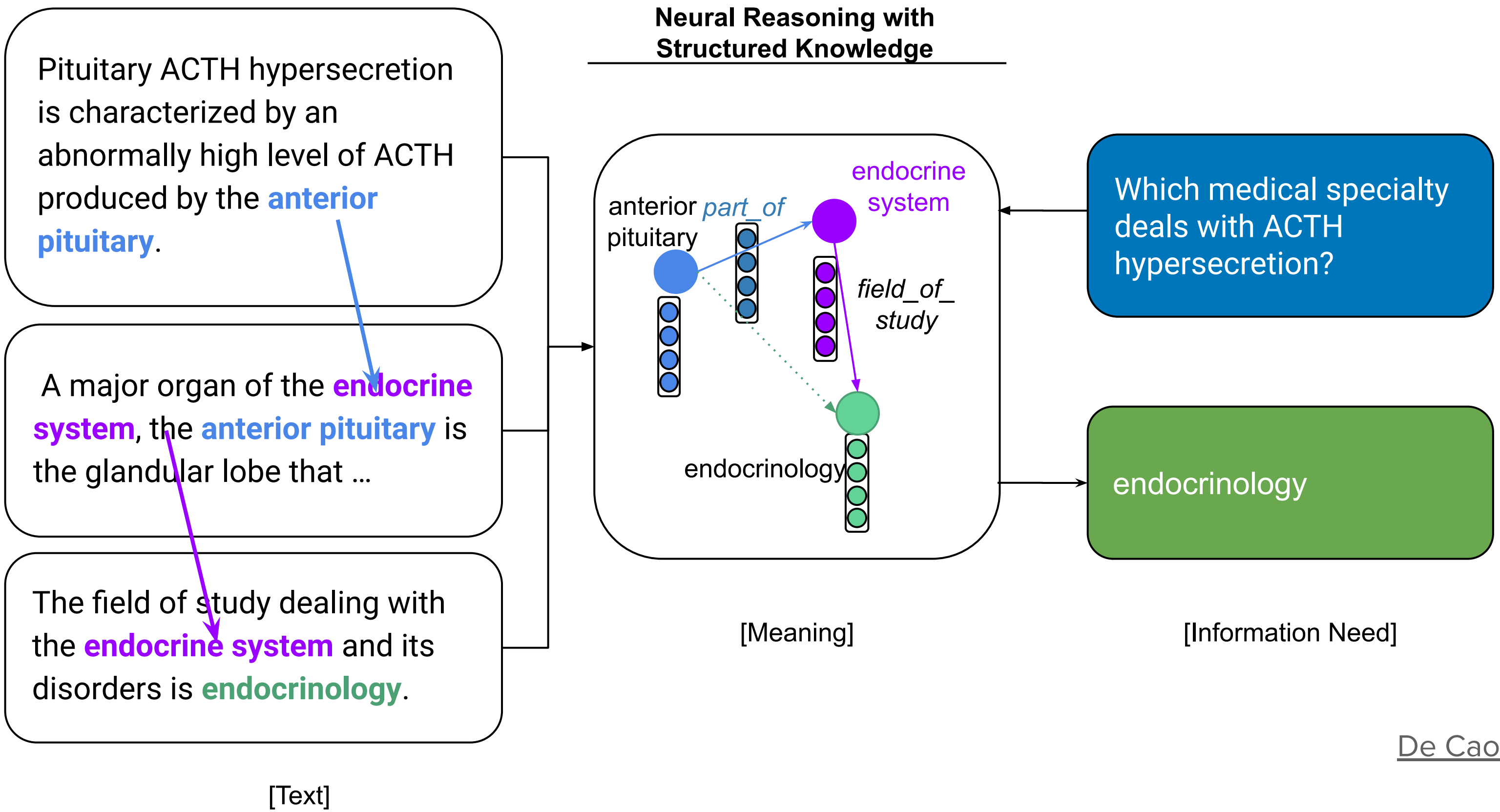
Welbl et al. (2018)

Challenge VI: Reasoning with Text



Welbl et al. (2018)

Challenge VI: Reasoning with Text



Challenge VII: Conversational Machine Reading

- Humans gather information by engaging in conversations involving a series of interconnected questions and answers.

- For machines to assist in information gathering, it is essential to enable them to answer conversational questions.

CoQA, QuAC

Jessica went to sit in her rocking chair. Today was her birthday and she was turning 80. Her granddaughter Annie was coming over in the afternoon and Jessica was very excited to see her. Her daughter Melanie and Melanie's husband Josh were coming as well.

Jessica had ....

Q1: Who had a birthday?
A1: Jessica

Q2: How old would she be?
A2: 80

Q3: Did she plan to have any visitors?
A3: Yes

Q4: How many?
A4: Three

CoQA: A Conversational Question Answering Challenge, Siva Reddy, et. al. 2018
QuAC : Question Answering in Context, Eunsol Choi, et. al. , 2018

QuAC

What is the origin of Daffy?	?→ first appeared in Porky's Duck Hunt
What was he like in that episode?	?→ assertive, unrestrained, combative
Was he the star?	?→ No, barely more than an unnamed bit player in this short
Who was the star?	??→ No answer

QuAC : Question Answering in Context, Eunsol Choi, et. al. , 2018

ShaRC

You'll carry on paying national insurance for the first 52 weeks you are abroad if you are working for an employer outside the EEA.

Do I need to carry on paying UK National Insurance?

FQ. Are you working for an employer outside the EEA?	Yes
FQ. Has it been less than 52 weeks since you are abroad?	Yes

A. Yes

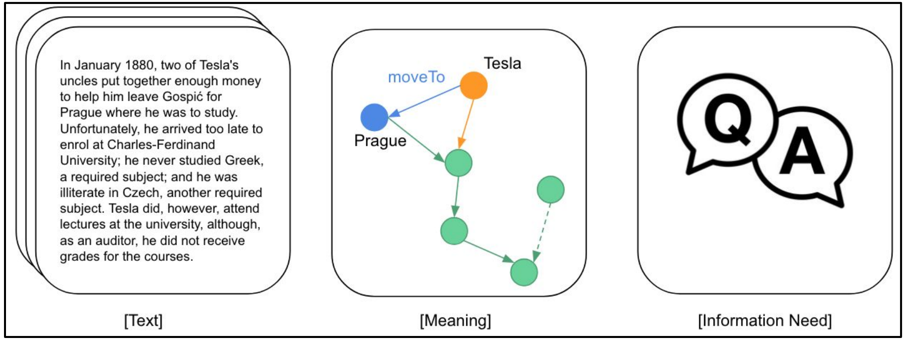
Shaping Answers with Rules through Conversation, Marzieh Saeidi et. al. , 2018

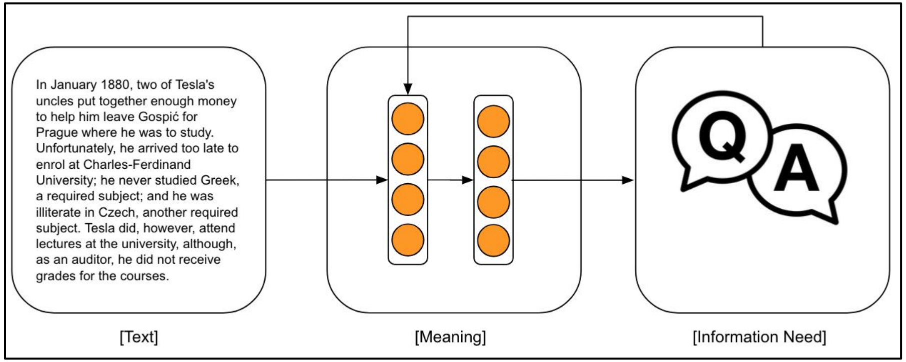
Conclusion

A Paradigm Shift

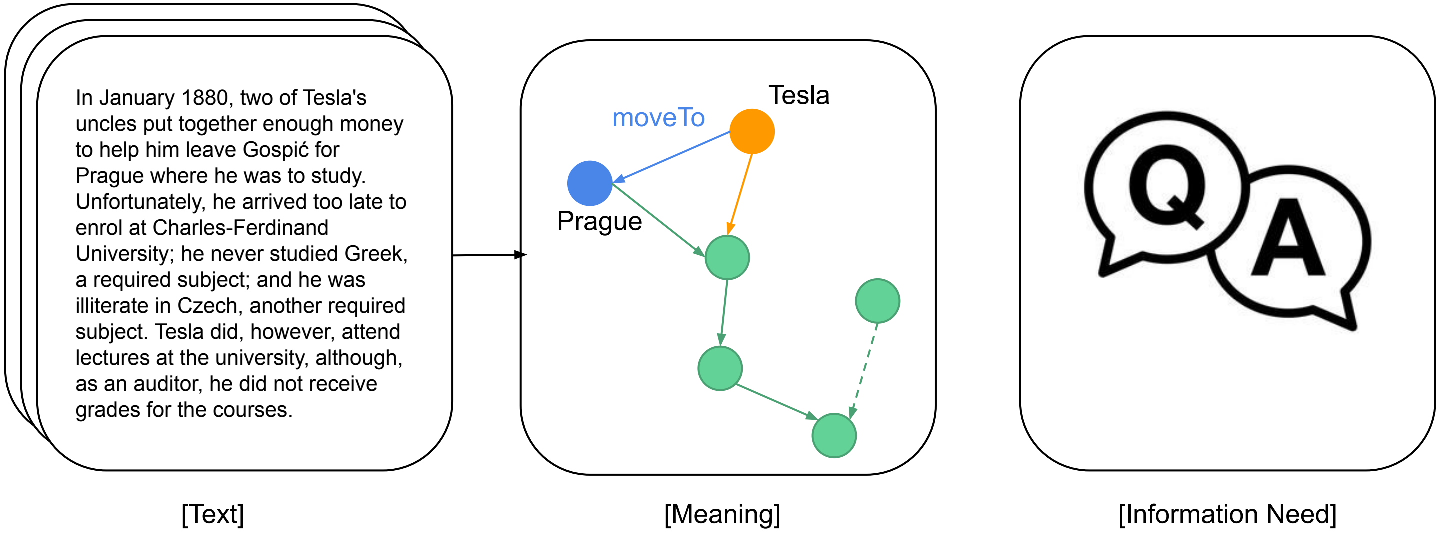
- Symbolic Meaning Representations
- ➡ Latent Vector Representations

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





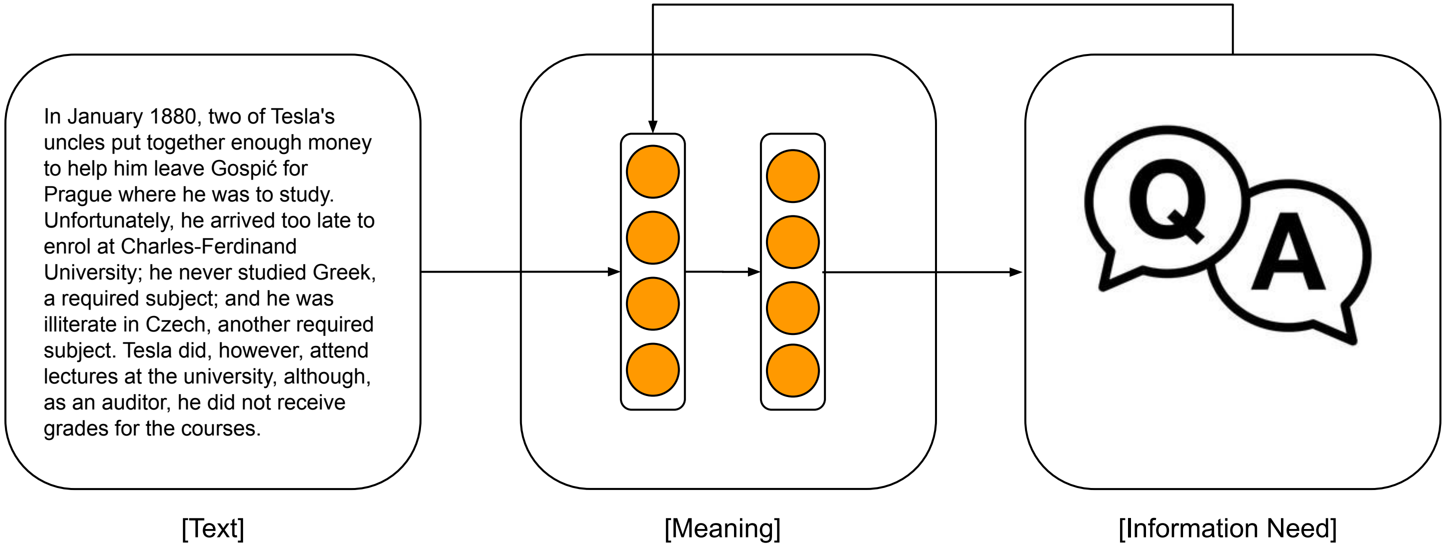
Automatic Knowledge Base Construction



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

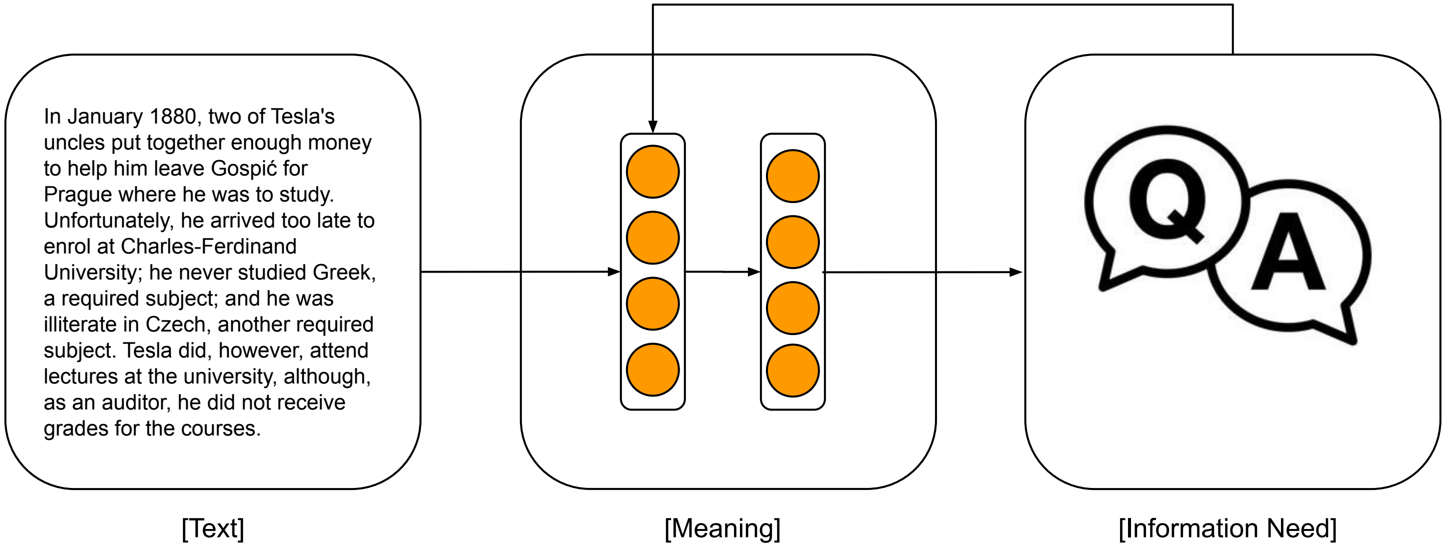
End-to-End Machine Reading



Distributed Representations

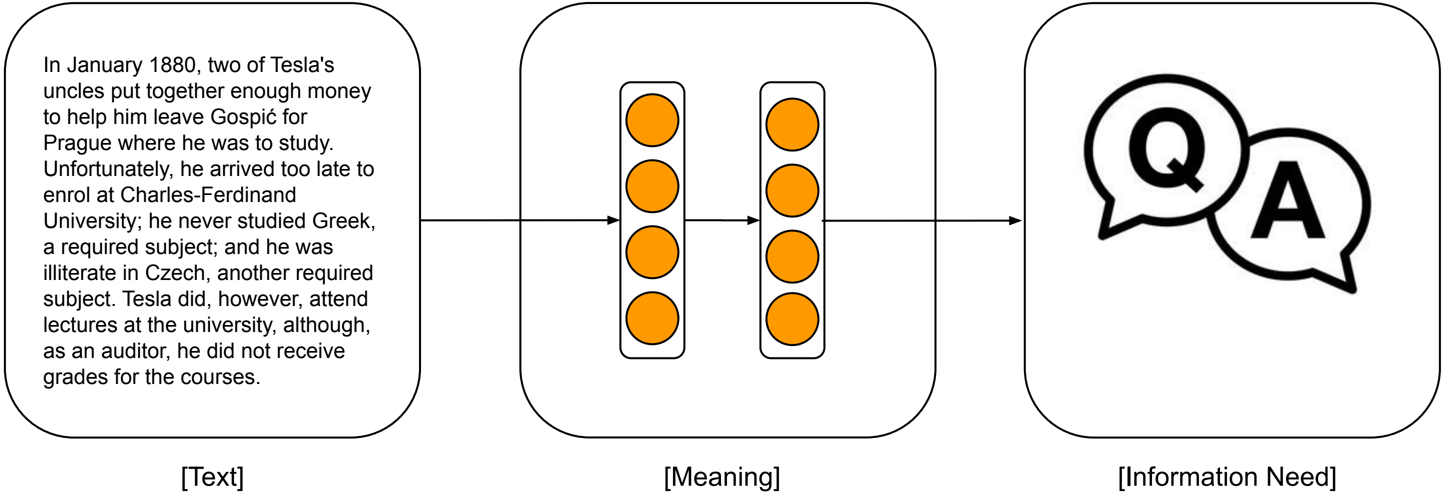
- Advantages
  - More robust to variation in language
  - No cascading errors
  - No domain expertise required
  - Multiple modalities (e.g., VQA) much easier
  - Easy annotation for end-to-end task (e.g., QA)
- Disadvantages
  - Scalability
  - Data efficiency
  - No interpretability
  - No support for reasoning
  - Representations not universal, but question-specific

End-to-End Machine Reading



End-to-End Machine Reading

universality?



Distributed Representations

- Advantages
  - More robust to variation in language
  - No cascading errors
  - No domain expertise required
  - Multiple modalities (e.g., VQA) much easier
  - Easy annotation for end-to-end task (e.g., QA)
- Disadvantages
  - Scalability
  - Data efficiency
  - No interpretability
  - No support for reasoning
  - Representations not universal, but question-specific [?]

Great research opportunities

References

- A Neural Probabilistic Language Model (Bengio et al. 2003, JMLR)
- A unified architecture for natural language processing. (Collobert & Weston 2008, ICML)
- Word representations: a simple and general method for semi-supervised learning. (Turian et al. 2010, ACL)
- Efficient Estimation of Word Representations in Vector Space. (Mikolov et al. 2013a, ICLR)
- Distributed Representations of Words and Phrases and their Compositionality. (Mikolov et al. 2013b, NIPS)
- GloVE: Global Vectors for Word Representation. (Pennington et al., 2014, EMNLP)
- Neural word embedding as implicit matrix factorization. (Levy & Goldberg 2014, NIPS)
- Character-Aware Neural Language Models. (Kim et al. 2016, AAAI)

- Blogs: http://ruder.io/word-embeddings-1/ , http://colah.github.io/posts/2015-01-Visualizing-Representations/

References

- QANet: Combining Local Convolution with Global Self-Attention for Reading Comprehension. (Yu et al. 2018, ICLR)
- Teaching machines to read and comprehend (Herrmann et al. 2015, NIPS)
- Attention is all you need. (Vaswani et al. 2017, NIPS)
- Long short-term memory-networks for machine reading. (Cheng et al. 2016, EMNLP)
- Gated self-matching networks for reading comprehension and question answering. (Wang et al. 2017, ACL)
- Improved semantic representations from tree-structured long short-term memory networks. (Tai et al. 2015, ACL)
- Recurrent Neural Network Grammars. (Dyer et al. 2016, NAACL)

References

- Adversarial Examples for Evaluating Reading Comprehension Systems (Jia et al. 2017, EMNLP)
- Know What You Don’t Know: Unanswerable Questions for SQuAD (Rajpurkar et al. 2018, ACL)
- Visual question answering: Datasets, algorithms, and future challenges (Kafle et al. 2017, Computer Vision and Image Understanding)
- Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering (Goyal et al. 2017, CVPR)
- Reading Wikipedia to Answer Open-Domain Questions (Chen et al. 2017, ACL)
- Event2Mind: Commonsense Inference on Events, Intents, and Reactions (Rashkin et al. 2018, arXiv)
- Semantically Equivalent Adversarial Rules for Debugging NLP Models (Ribeiro 2018, ACL)
- Understanding Neural Networks through Representation Erasure (Li et al. 2016, arXiv)
- HotFLip: White-Box Adversarial Examples for NLP (Ebrahimi et al. 2017, arXiv)
- Anchors: High-Precision Model-Agnostic Explanations (Ribeiro et al. 2018, AAAI)
- Deep contextualized word representations (Peters et al. 2018, NAACL)
- Learned in Translation: Contextualized Word Vectors (McCann et al. 2017, NIPS)
- Supervised Learning of Universal Sentence Representations from Natural Language Inference Data (Conneau et al. 2017, EMNLP)
- Efficient Estimation of Word Representations in Vector Space (Mikolov et al. 2013, NIPS)
- Simple and Effective Semi-Supervised Question Answering (Dhingra et al. NAACL 2018)
- Neural Domain Adaptation for Biomedical Question Answering (Wiese et al. 2017, CoNLL)
- Improving Language Understanding by Generative Pre-Training (Radford et al. 2018, arXiv)
- Neural Skill Transfer from Supervised Language Tasks to Reading Comprehension (Mihaylov et al. 2017, arXiv)
- Representing General Relational Knowledge in ConceptNet 5 (Speer and Havasi, LREC 2012)
- Learning to understand phrases by embedding the dictionary (Hill et al. 2016, TACL)
- Leveraging knowledge bases in lstms for improving machine reading (Yang et al. 2017, ACL)
- Knowledgeable Reader: Enhancing Cloze-Style Reading Comprehension with External Commonsense Knowledge. (Mihaylov and Frank, 2018, ACL)

- Reading Wikipedia to Answer Open-Domain Questions (Chen et al. 2017, ACL)

- Evidence aggregation for answer re-ranking in open-domain question answering (Wang et al. ICLR 2018)
- Marco Baroni and Gemma Boleda: https://www.cs.utexas.edu/~mooney/cs388/slides/dist-sem-intro-NLP-class-UT.pdf
- News article: https://www.independent.co.uk/infact/brexit-second-referendum-false-claims-eu-referendum-campaign-lies-fake-news-a8113381.html

References for Datasets

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

Thank You!

Backup or Old Slides

Why do we need compositional phrase representations in QA?

What city did Tesla move to in 1880?

In January 1880, two of Tesla's uncles put together enough money to help him leave Gospić for Prague where he was to study.

- **Goal**: similar representations for phrases with similar meaning, even with lexical / syntactic variation

"move from Gospić to Prague" ≈ "leave Gospić for Prague"

Synthesis: Symbolic vs. Subsymbolic Machine Reading

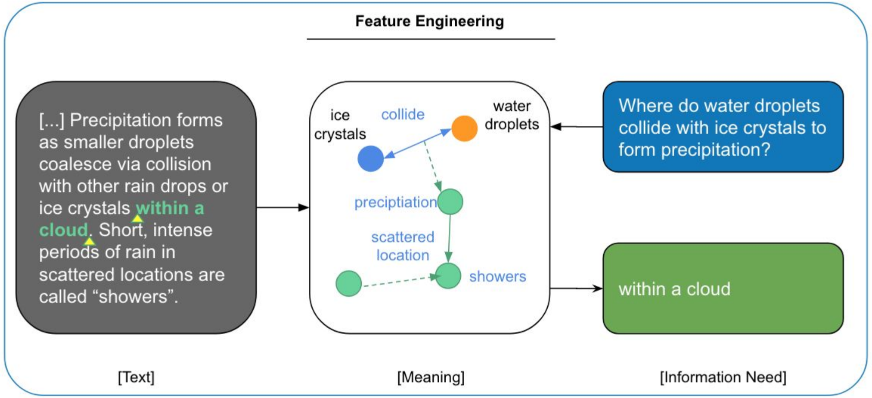
- A transferrable representation of text
  - that humans and machine can interface with.

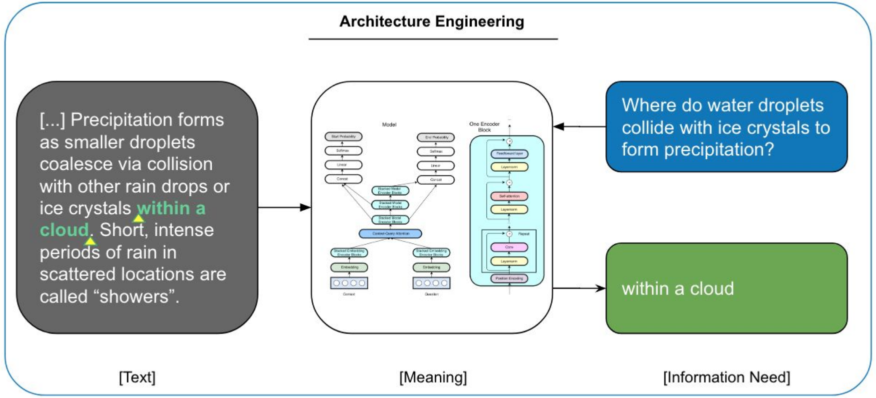
	Knowledge Base	**Neural Networks**
Knowledge Representation	structured / explicit	distributed / implicit
Means of Construction	Information Extraction	(Un)supervised Learning
Interface	Query Language	Vectors
Optimization	discrete	gradient-based

A Paradigm Shift

- Symbolic Meaning Representations
- ➡ Latent Vector Representations

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





Gains and Losses of this Shift

- Gains
  - Generalization and domain transferability (mainly due to unsupervised learning)
  - No domain expertise
  - Multiple modalities (e.g., VQA) much easier
  - Easy annotation for end-to-end task (e.g., QA)

- Losses
  - Ability to do reasoning
  - Data efficiency
  - Incorporating background knowledge
  - Scalability
  - Interpretability

Great research opportunities

Synthesis: Symbolic vs. Subsymbolic Machine Reading

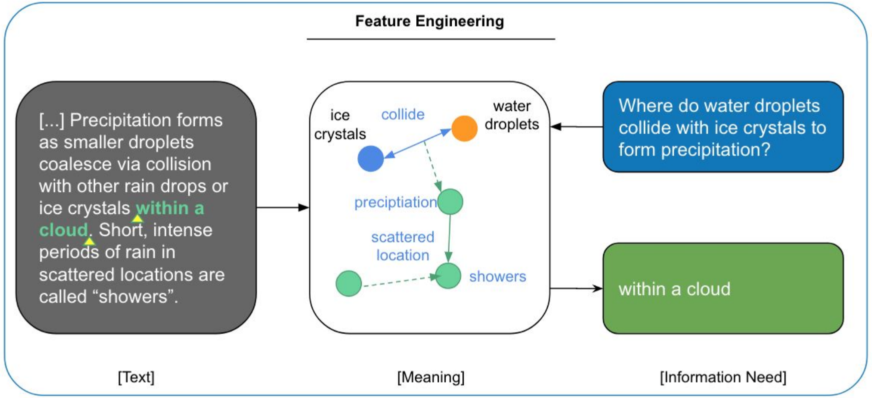
- A transferrable representation of text
  - that humans and machine can interface with.

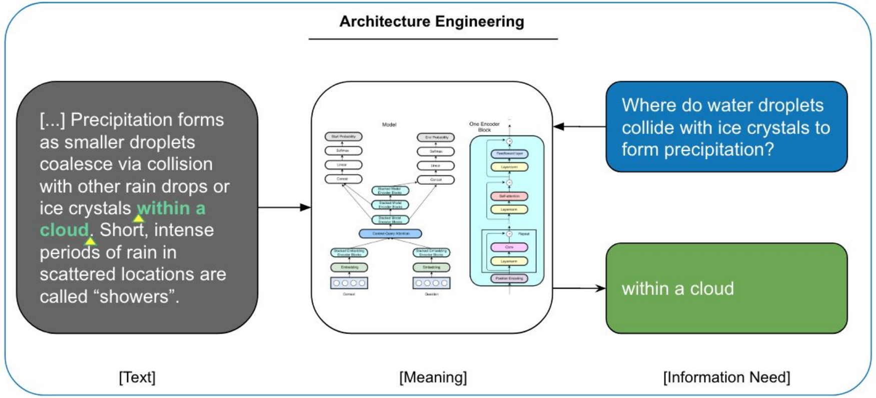
	Knowledge Base	**ELMo Vectors**
Knowledge Representation	structured / explicit	distributed / implicit
Means of Construction	Information Extraction	Applying Language Model
Interface	Query Language	Neural Net
Optimization	discrete	gradient-based

A Paradigm Shift

- Symbolic Meaning Representations
- ➡ Latent Vector Representations

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





A Synthesis ?!

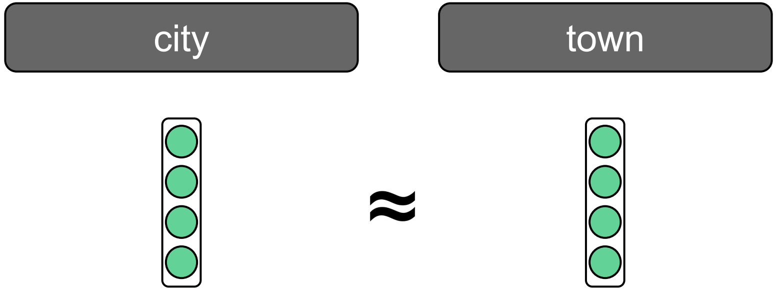
- Can we solve the challenges of end-to-end solutions that could be addressed more easily with intermediate symbolic meaning representations?

- Or can we find a way to synthesize the best of both worlds?

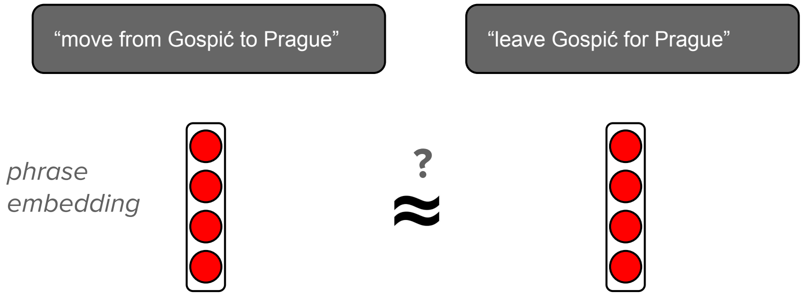
Best Practices

- Exploit pre-trained models:
  - (Minimum) word embeddings and language models
  - Modeling innovations such as (self-)attention
  -
- …
- Nice reference: ruder.io/deep-learning-nlp-best-practices/

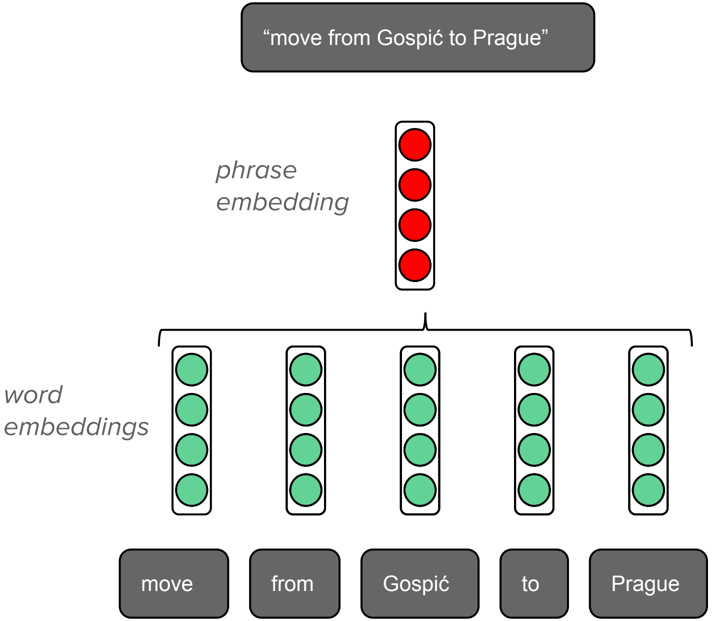
Similarity between words: word embeddings



Similarity between phrases?

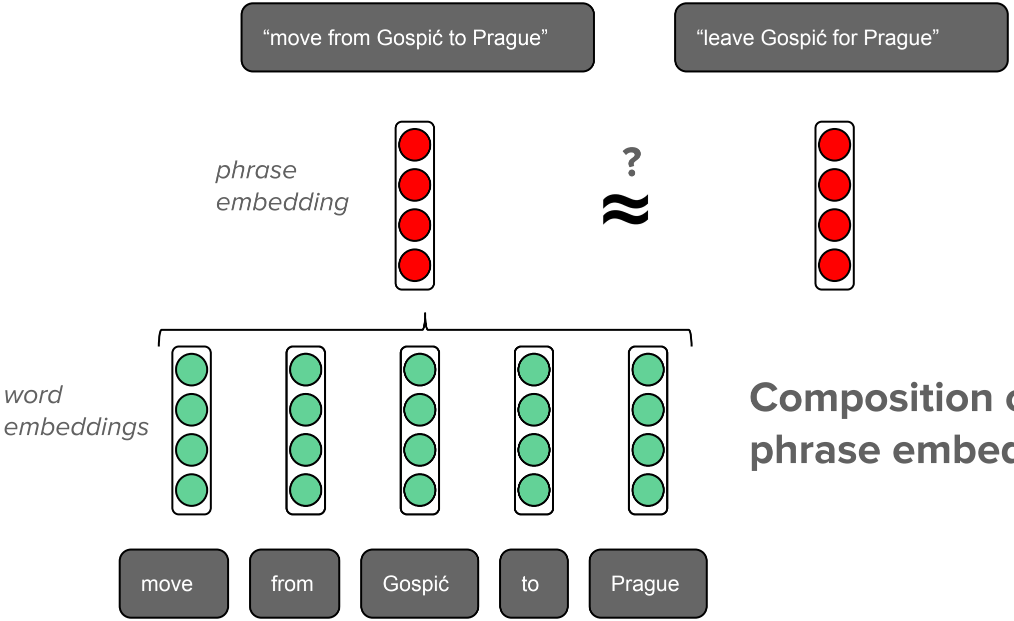


Similarity between phrases?



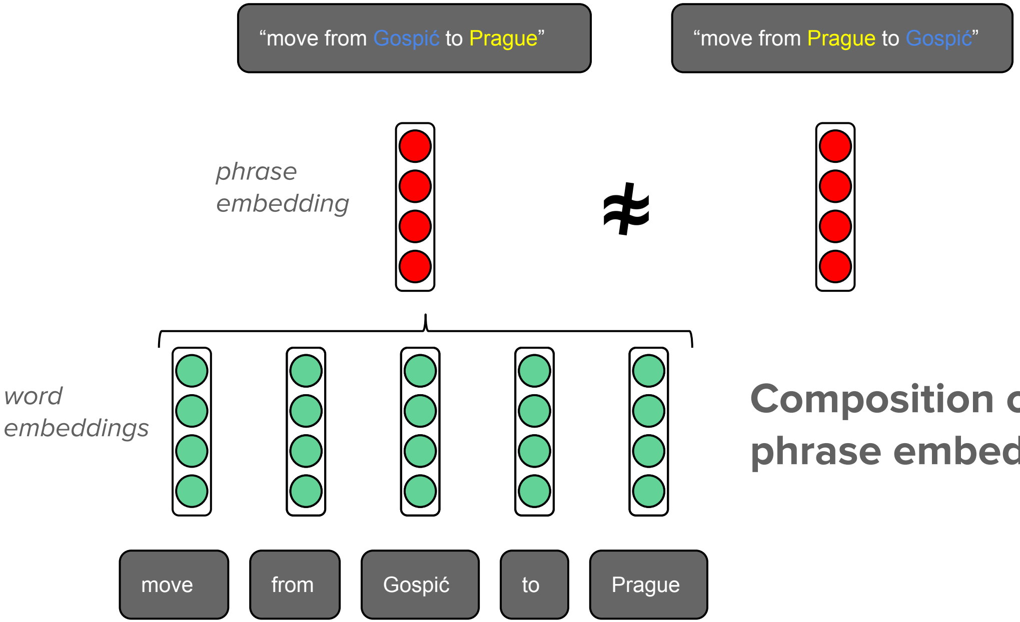
Composition of word into phrase embeddings

Similarity between phrases?



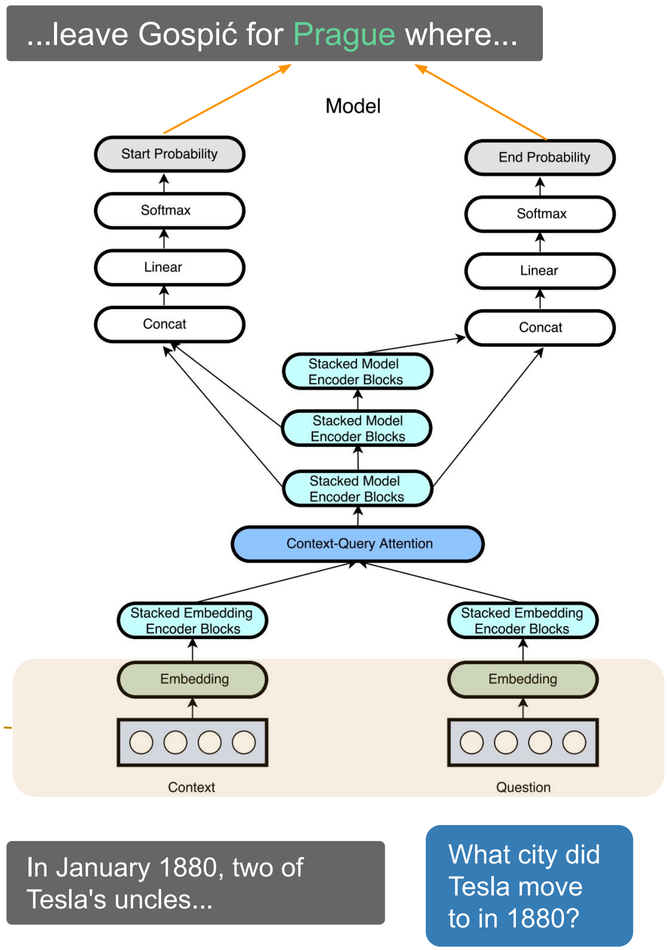
Composition of word into phrase embeddings

Similarity between phrases?



Word order matters!

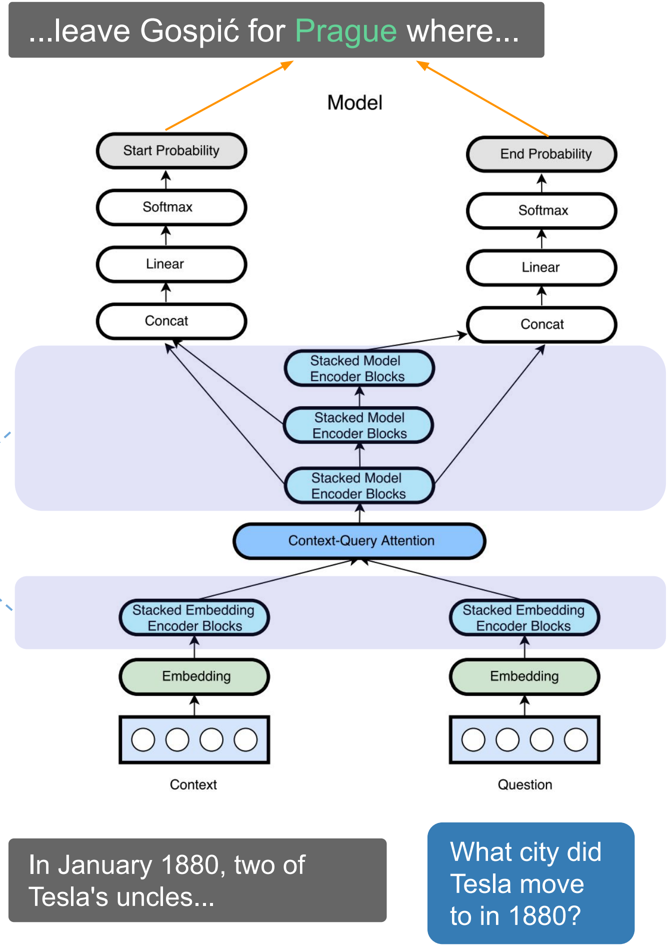
Composition of word into phrase embeddings



How to represent symbols?

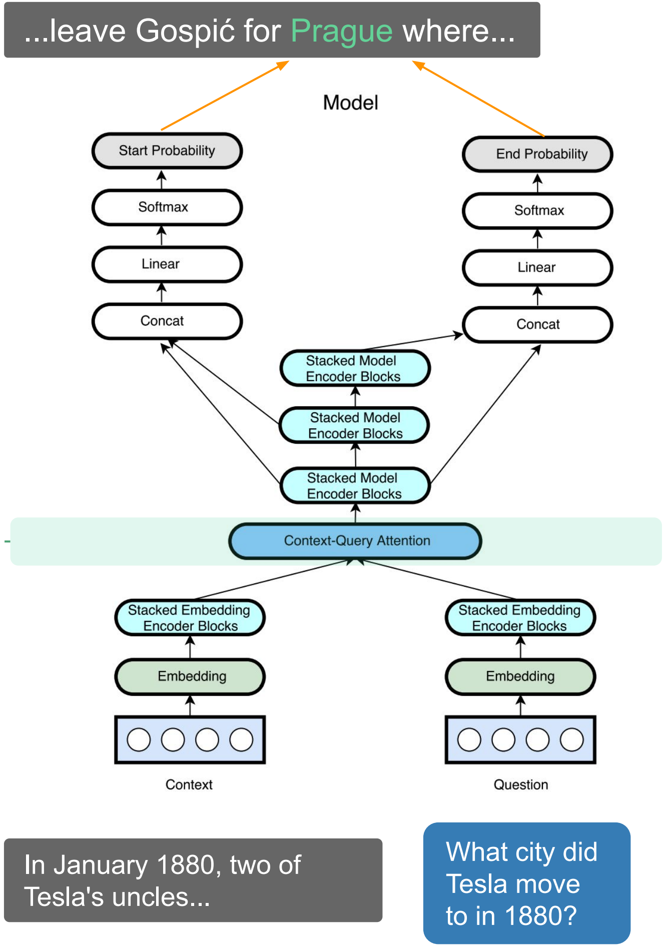
QANet, Yu et al. (2018)

How to condition word representations on one another



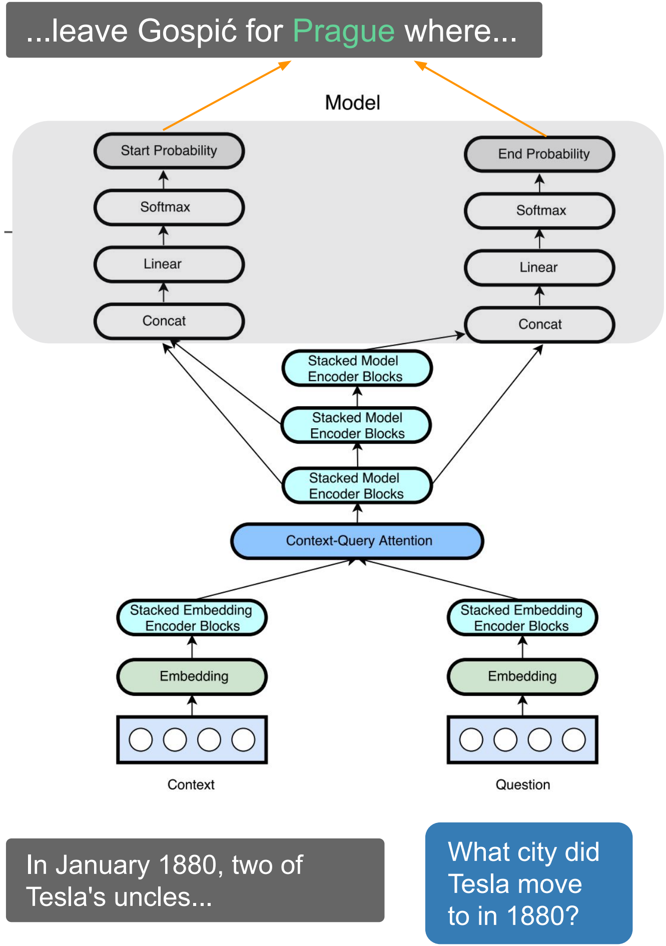
QANet, Yu et al. (2018)

sequence interaction between question and text



QANet, Yu et al. (2018)

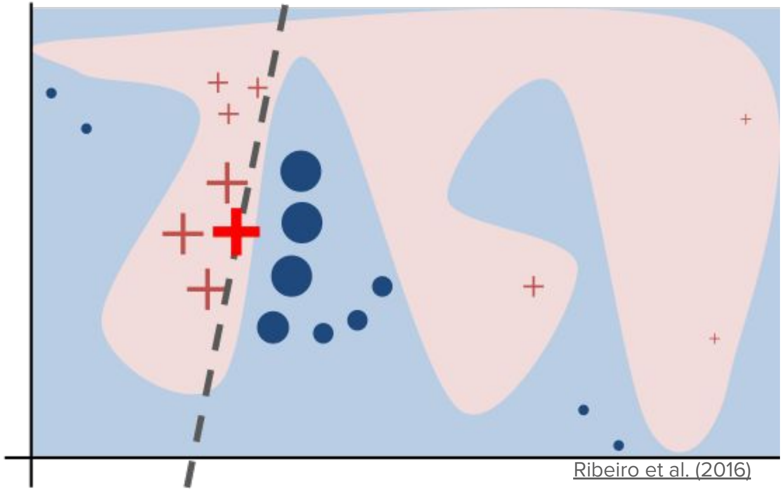
Span Scoring: linear projection, score for start and end position



QANet, Yu et al. (2018)

Model Diagnostics: Right for the Wrong Reason?

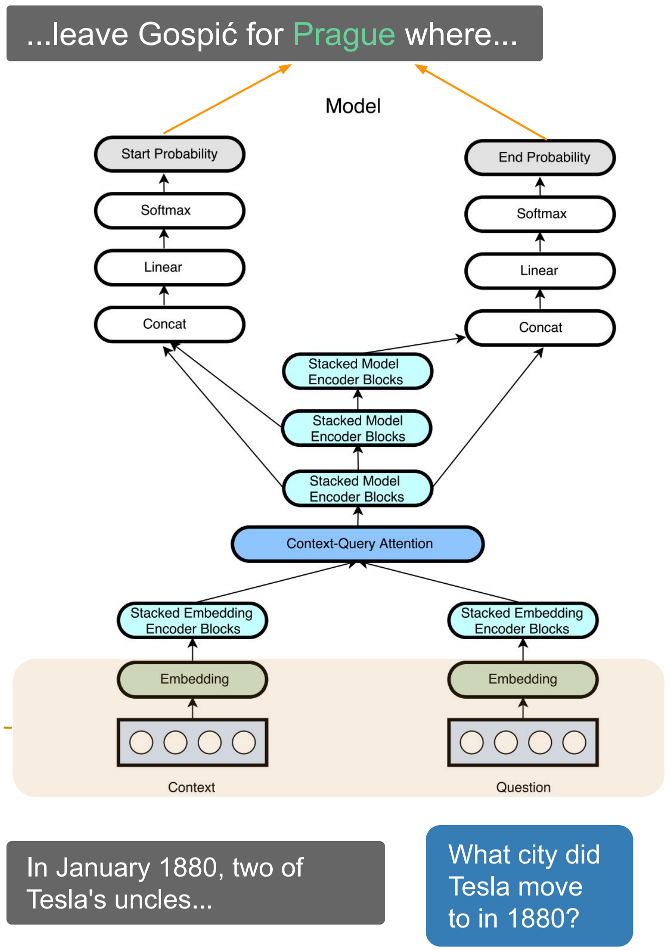
- Example 2: LIME (Ribeiro et al. 2016)
  - Idea: Find features that predictions are sensitive to
  - Local perturbations, fit linear model on predictions



Ribeiro et al. (2016)

Ribeiro et al. (2016)

- Alvarez-Melis and Jaakkola (2017): similar, but with sequences.



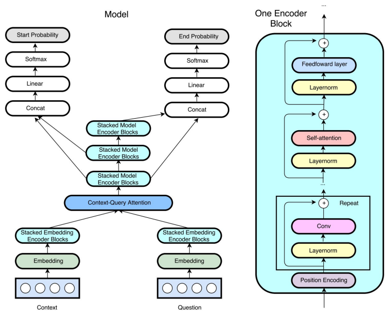
How to represent symbols?

QANet, Yu et al. (2018)

Architecture Engineering

Architecture Engineering

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



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

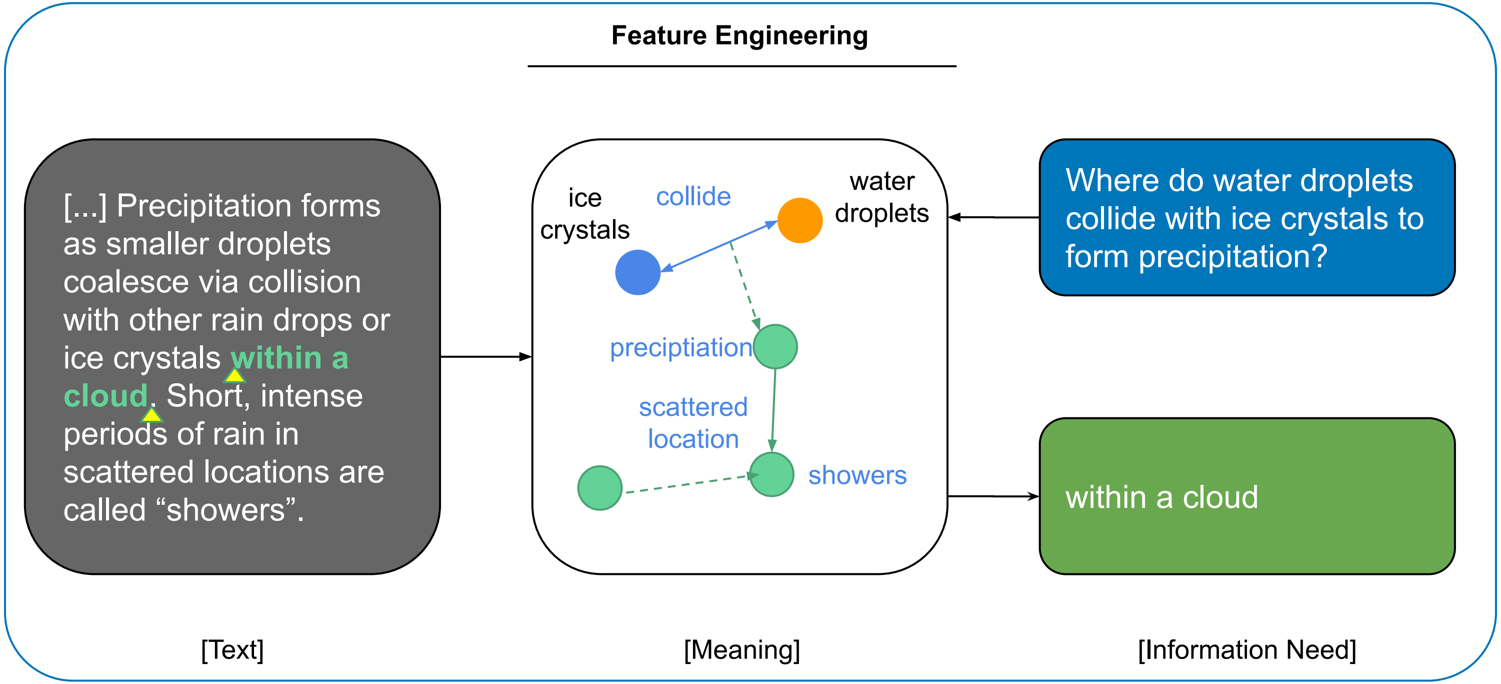
within a cloud

[Text]

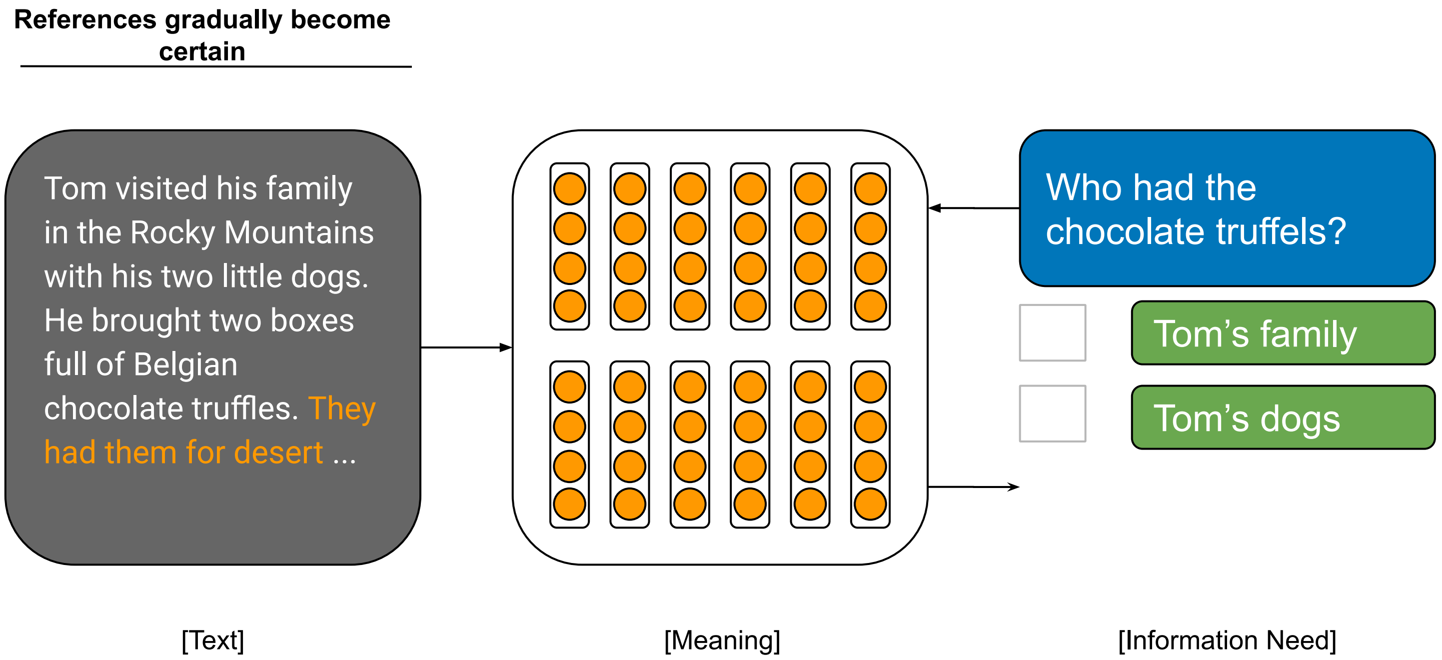
[Meaning]

[Information Need]

Architecture Engineering



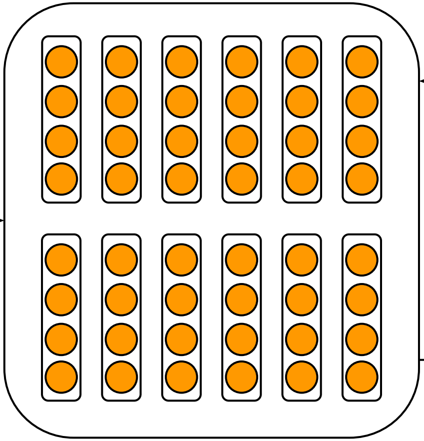
Challenge II: Ambiguity



Challenge II: Ambiguity

References gradually become certain

Tom visited his family in the Rocky Mountains with his two little dogs. He brought two boxes full of Belgian chocolate truffles. They had them for desert in their reunion barbeque.



Who had the chocolate truffels?

X Tom’s family

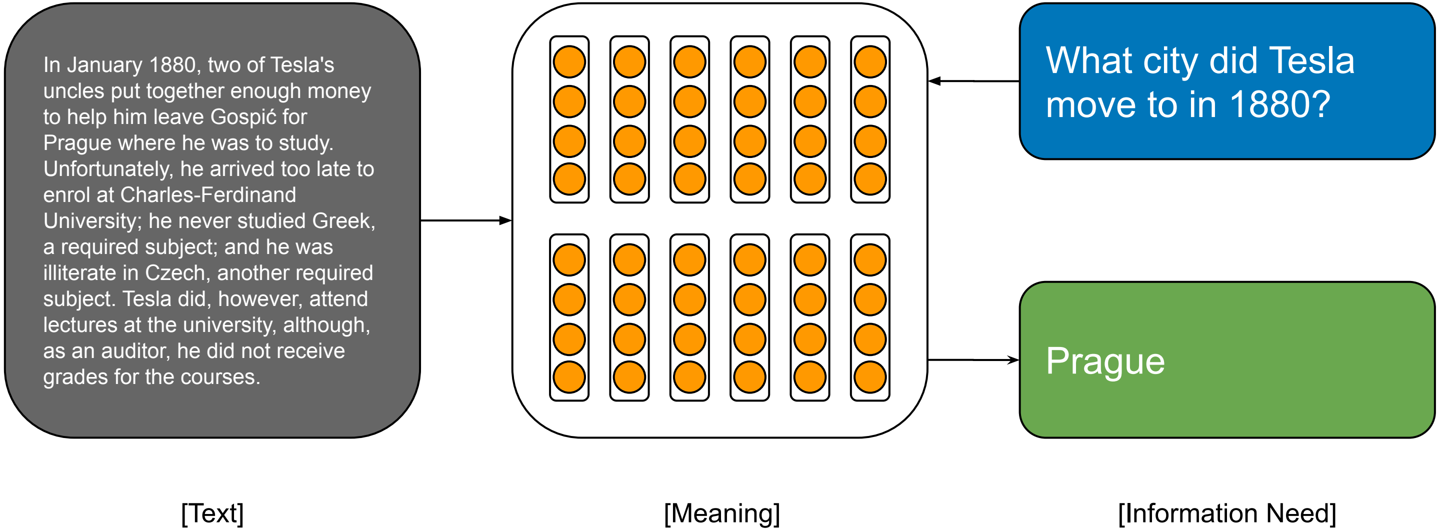
Tom’s dogs

[Text]

[Meaning]

[Information Need]

End-to-end Machine Reading for Question Answering



Representing Words in Context

Why do we need compositional representations in QA?

What **city** did Tesla move to in 1880?

In January 1880, two of Tesla's uncles put together enough money to help him leave Gospić for **Prague** where he was to study.

- **Goal**: similar representations for tokens in similar contexts, for instance through lexical / syntactic variation

"move from Gospić to **Prague**" ≈ "leave Gospić for **Prague**"

Similarity between contexts?

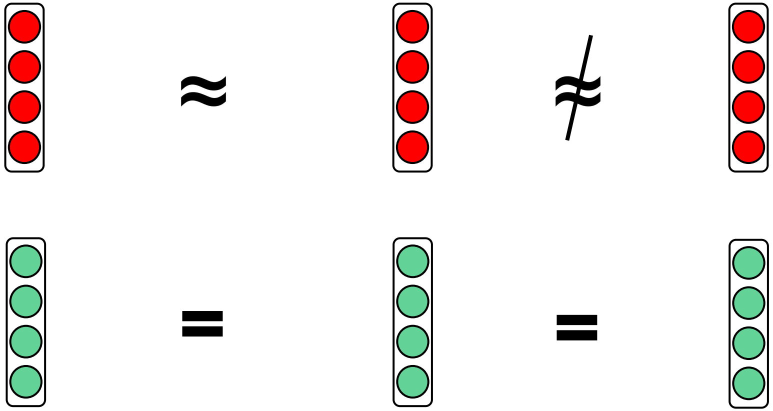
"move from Gospić to **Prague**"

"leave Gospić for **Prague**"

"leave **Prague** for Gospić"

Contextual representation of ***Prague***

Word representation of ***Prague***



Word order matters!

Word Similarity

“Words are defined by the company they keep.”

➜ Two words are similar if they appear in the same documents.

Term-Document matrix:

	d1	d2	d3	d4	…	d*M*
resident	2	0	0	0	…	1
street	0	1	0	1	…	0
city	4	2	0	1	…	1
…	…	…	…	…	…	…
town	1	1	0	1	…	1
mozarella	0	0	3	0	…	0
balsamico	0	0	1	0	…	0

Somewhat collinear, but very sparse