Weakly Supervised Natural Language Understanding

Ni Lao
mosaix.ai
11.11.2018

For completeness a large part of the tutorial is from previous works.
Thanks to Chen Liang for helping with creating these slides

Speaker Background

- **BS from Tsinghua U. EE (1999 - 2003)** Worked on the TsinghuAeolus system, and won world champion in RoboCup simulation league in 2001 and 2002
- **MS from Tsinghua U. CS (2003 - 2006)** Worked at Microsoft Research Asia on automatic OS diagnosis, Web search and product search
- **PhD from Carnegie Mellon U. (2006 - 2012)** Researched on IR, ML, NLP. Worked on the CMU JAVELIN QA system, and Never-Ending Learning (NELL) system
- **Research Scientist Google (2012 - 2017)** Researched on KG construction, semantic parsing. Worked on KG and Web-based QA products
- **Chief Scientist & Co-founder Mosaix.ai (2018 -)** Research on semantic parsing and text understanding for search and NLU services. And we are hiring!

Plan

Access slides and join discussions at
weakly-supervised-nlu google group

Mobile

Desktop

- ***Weak Supervision NLP***
  - NLP, AI, software 2.0
  - Semantics as a foreign language
  - Unsupervised learning
  - Knowledge representation (symbolism)
- ***Semantic Parsing Tasks***
  - *WebQuestionsSP, WikiTableQuestions*
- ***Neural Symbolic Machines*** (ACL 2017)
  - Compositionality (short term memory)
  - Scalable KB inference (symbolism)
  - RL vs MLE
- ***Memory Augmented Policy Optimization*** (NIPS 2018)
  - Experience replay (long term memory & optimal updating strategy)
  - Systematic exploration
  - Memory Weight Clipping (unbiased cold start strategy)

Natural Language Processing (NLP)

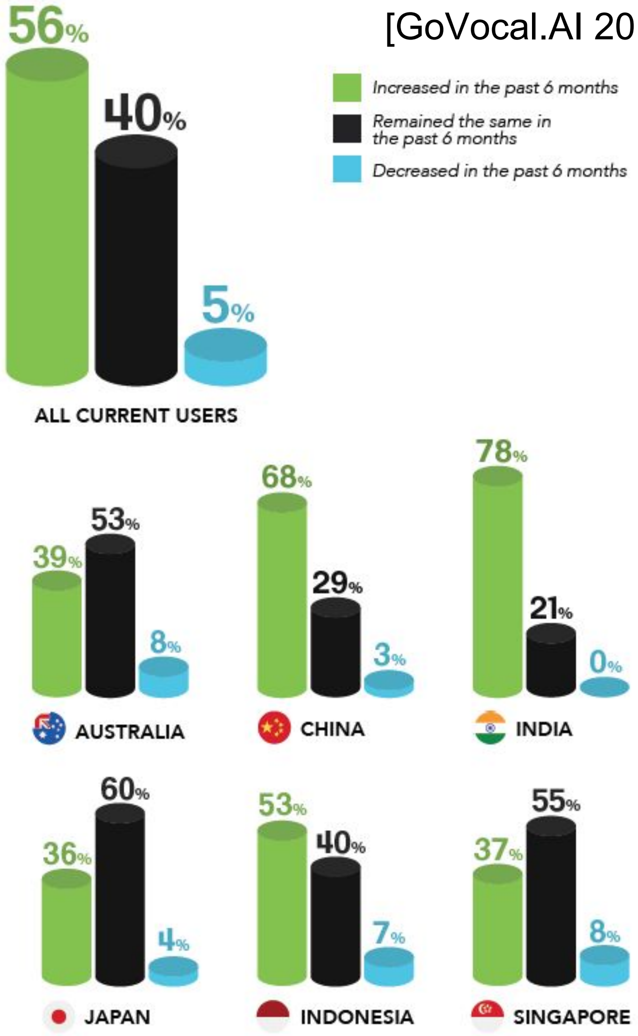
- Enables machines to understand and assists human
- A major problem of AI (AI-complete)
- Leads to new theories in cognitive science

There can be two underlying motivations for building a computational theory. The **technological goal** is simply to build better computers, and any solution that works would be acceptable. The **cognitive goal** is to build a computational analog of the human-language-processing mechanism; such a theory would be acceptable only after it had been verified by experiment.
-- James Allen,1987

Adoption Of Voice Technology

[GoVocal.AI 2018]

- Google's Speech Internationalization Project: From 1 to 300 Languages and Beyond -- Pedro J. Moreno, 2012
- My daughter adopted YouTube voice command since 2 years old
- 20% of the U.S. population has access to smart speakers -- Techcrunch, 2018
- Rising adoption in the Asia Pacific -- GoVocal.AI 2018



Language understanding for AI and humanity

- Experts with different views of AI agree on the potential of NLP

If you got a billion dollars to spend on a huge research project that you get to lead, what would you like to do?
-- r/CyberByte, 2015

NLP is fascinating, allowing us to focus on **highly-structured inference** problems, on issues that go to the core of "**what is thought**" but remain eminently practical, and on a technology that surely would **make the world a better place**.
-- Michael I Jordan

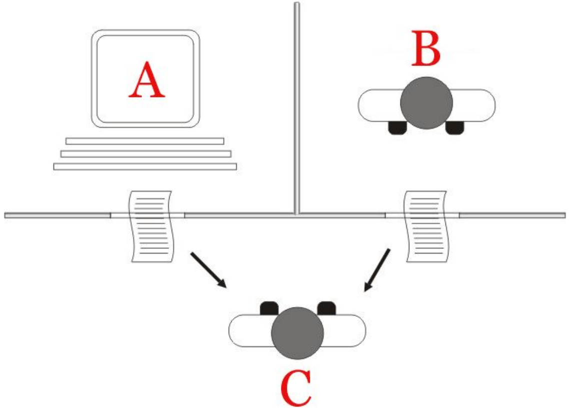
What kind of impact you hope deep learning has on our future?
-- Steve Paikin, 2016

I hope it allows Google to … search by the **content of the document** rather than by the words in the document … I hope it will make for **intelligent personal** systems, who can **answer questions** in a sensible way … It will make computers much easier to use. Because you'll be able to just **say to your computer** "print this damn thing"
-- Geoffrey Hinton

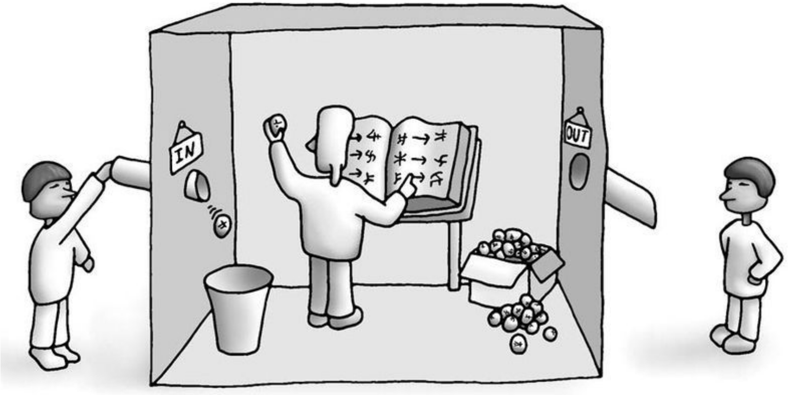
What is understanding?

"If they find a parrot who could **answer** to everything, I would claim it to be an **intelligent** being without hesitation.",
-- Alan Turing, 1950

Does the machine literally "**understand**" Chinese ? Or is it merely **simulating** the ability to understand Chinese?
-- John Searle, 1980



The Imitation Game



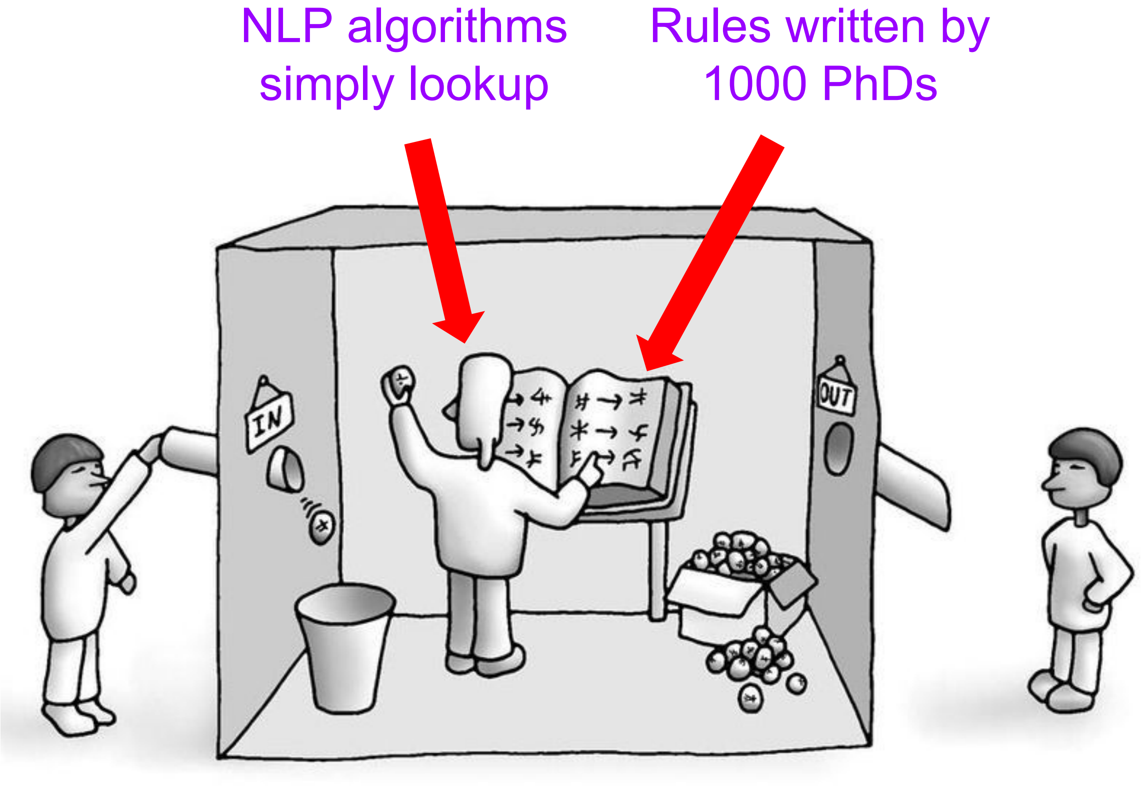
The Chinese Room Argument

Full Supervision NLP

- Traditionally NLP is a labor-intensive business

Applications
Discourse Processing
Semantic Parsing
Syntactic analysis
Morphological analysis

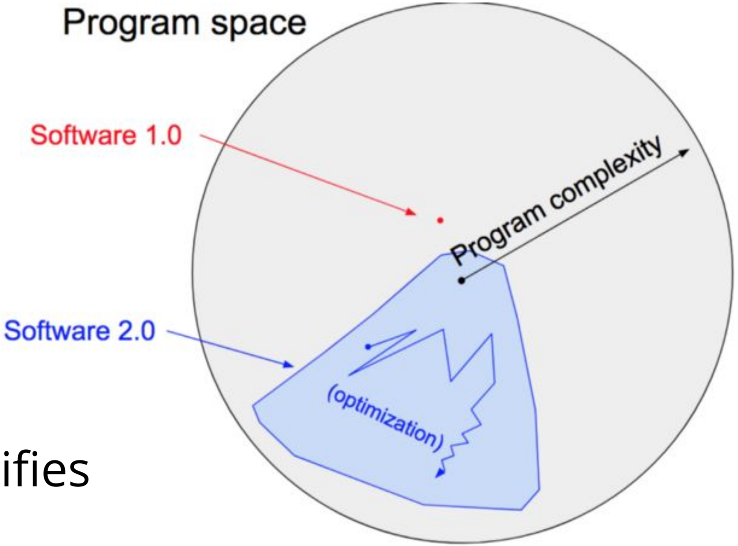
ASR	OCR
speech	text



Software 2.0

1. specify some goal on the behavior
    - e.g., “satisfy input output pairs of examples”,
    - e.g.,“win a game of Go”
2. write a rough skeleton of the code that identifies a subset of program space to search
    - e.g. a neural net architecture
3. use the computational resources at our disposal to search this space for a program that works.

Death of feature engineering. (The) users of the software will (play) a direct role in building it. Data labeling is a central component to system design.



[Karpathy 2017; Watson 2017; Ratner et al. 2018]

Where does knowledge come from?

- Can only cover the most popular semantics used by human

the world

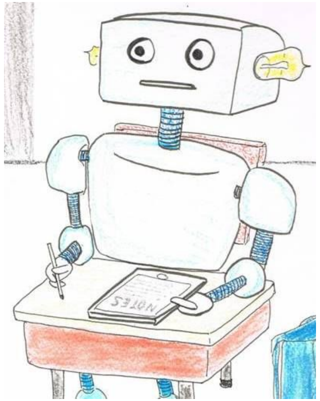
domain experts
(bottlenecks)

expert systems with
knowledge bases

Weak Supervision NLP

- Avoid the knowledge acquisition bottleneck with machine learning
- Then we can cover all possible semantics used by human

end to end examples
(e.g., QA pairs)



machine learning

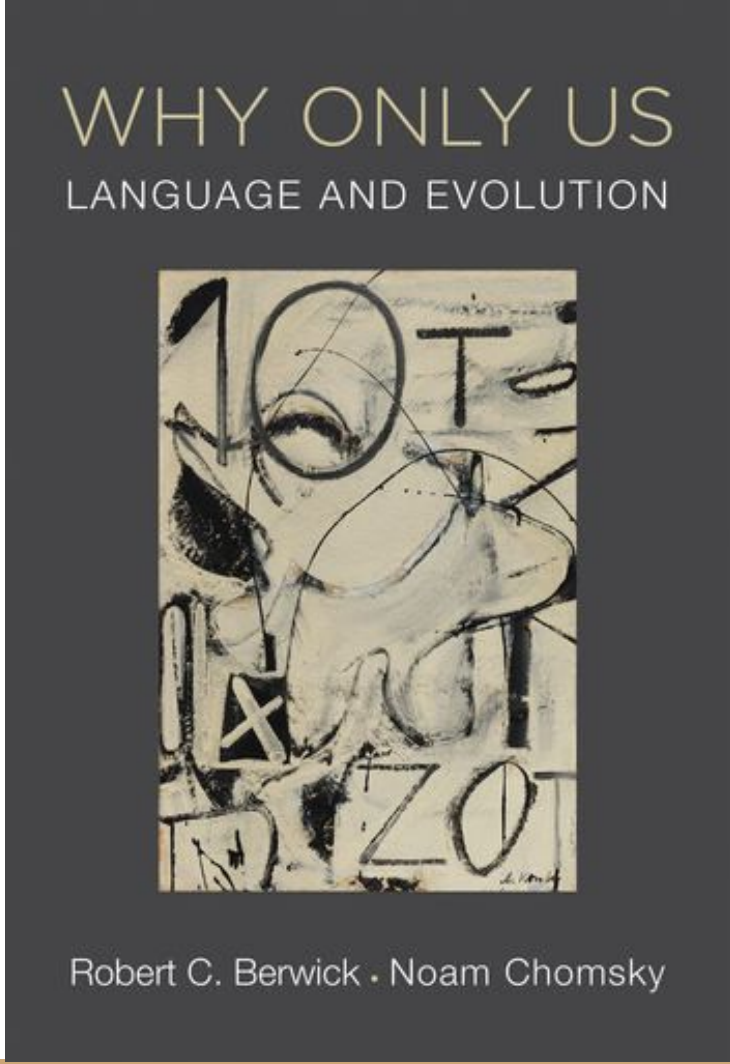
intelligent systems
with knowledge

Language & Reasoning

- The formalist view
  - Language was primarily invented for reasoning [Everaert+ 2015]

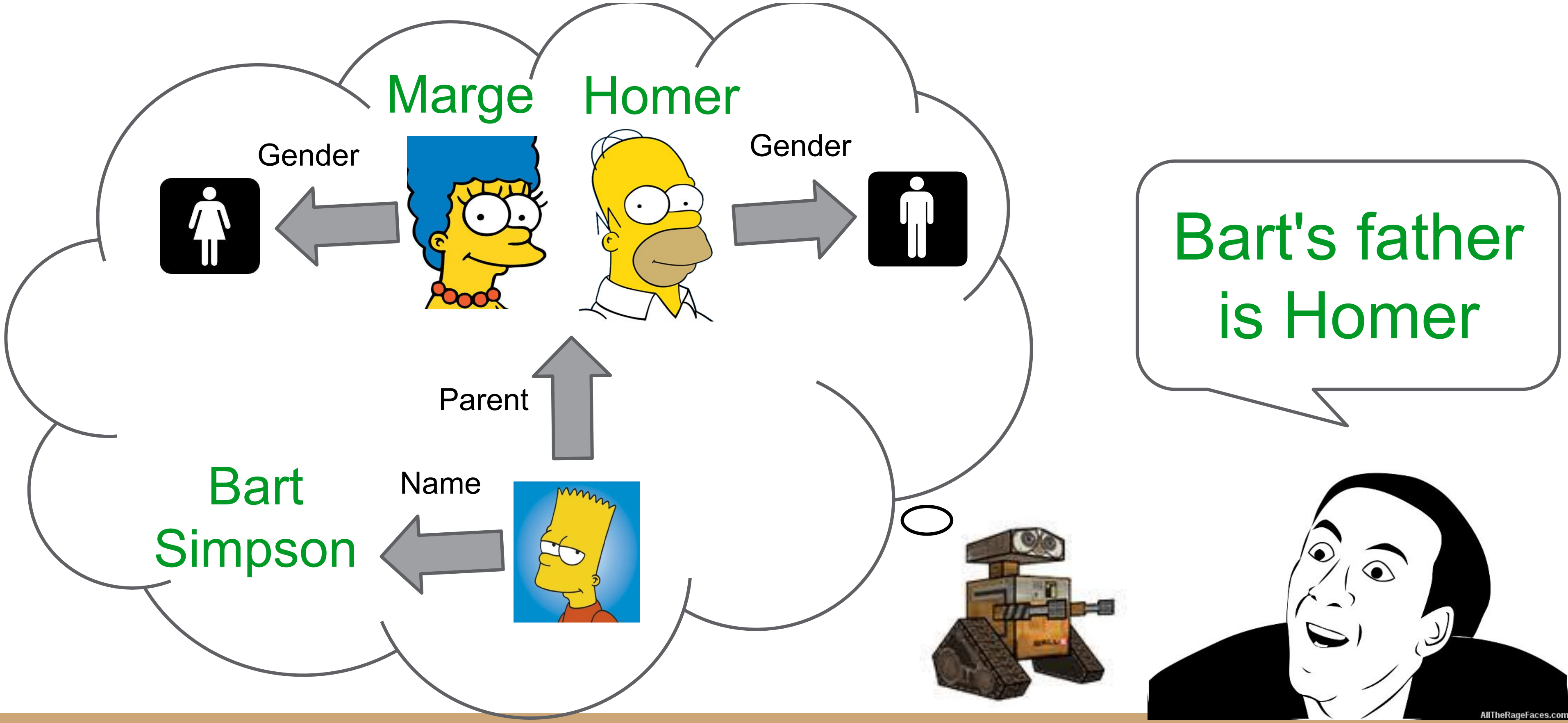
- The functionalist view
  - Language is for communication [Kirby 2017]

- Cognitive coupling hypothesis
  - sequential processing is “necessary for behaviours such as primate tool use, navigation, foraging and social action.” [Kolodny & Edelman 2018]



Reasoning is needed to understand text

[Liang+ 2017]



Semantics has a underlying language

“impressionist painters during the 1920s”

painters [/painting] !/art_forms

impressionist <visual_artist> x.[/associated_periods_or_movements = /impressionism]

<artist> during the 1920s x.[/date_of_work < 1930; /date_of_work > 1920]

Semantics has a underlying language

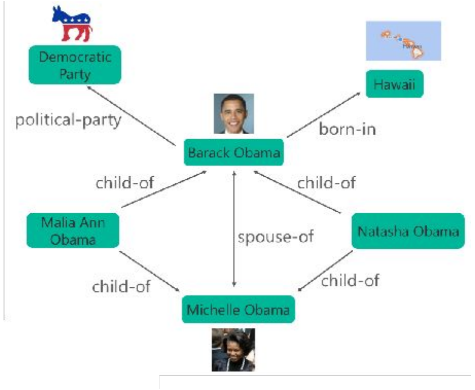
Semantic parsing

Largest city in US?

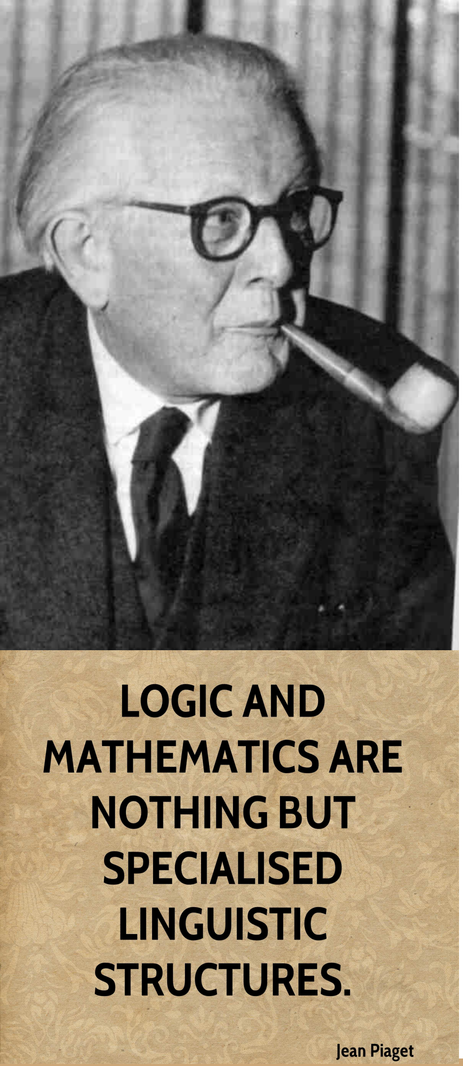
```
GO
(Hop V1 CityIn)
(Argmax V2 Population)
RETURN
```

Computation

NYC



Freebase, DBpedia, YAGO, NELL



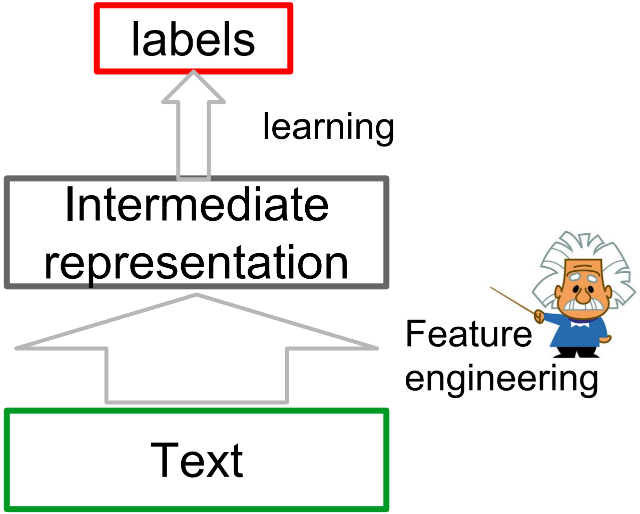
Semantics as a foreign language

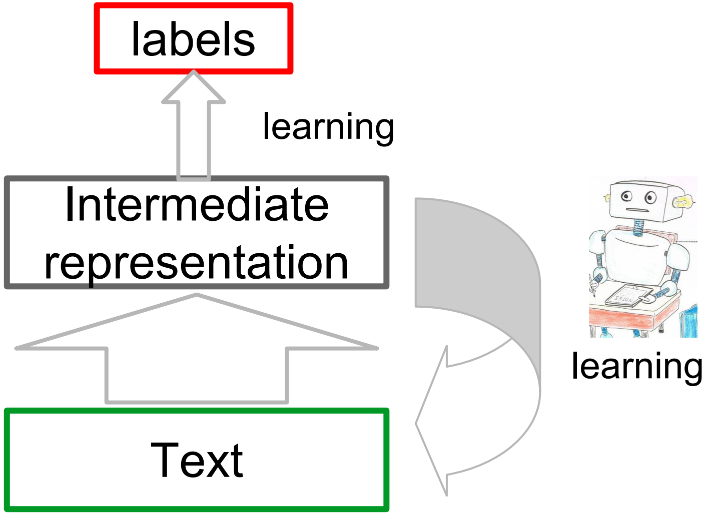
1) Natural languages are programming languages to control human behavior (either others or self)

2) For machines and human to understand each other, they just need translation models trained with control theory

NLU with Unsupervised Learning

- **Supervised** learning needs either feature engineering for compact representation or large amount of labeled data
- **Unsupervised** learning produces better representations and reduces the labeling cost, and it is easily transferable!





NIPS [LeCun 2016]

LeCun's Cake

- "Pure" Reinforcement Learning (cherry)
  - The machine predicts a scalar reward given once in a while.
  - **A few bits for some samples**
- Supervised Learning (icing)
  - The machine predicts a category or a few numbers for each input
  - Predicting human-supplied data
  - **10→10,000 bits per sample**
- Unsupervised/Predictive Learning (cake)
  - The machine predicts any part of its input for any observed part.
  - Predicts future frames in videos
  - **Millions of bits per sample**

LSA [Deer-wester+ 1990]
skip-gram [Mikolov + 2013]
Glove [Pennington+ 2014]

Pre-Trained Word Embeddings

- learning (**non-contextual**) word embeddings from text
  - matrix factorization on global **word-word co-occurrence counts** [1990]
  - **local context window** methods [2014]
  - weighted least squares on global **word-word co-occurrence counts** [2014]

- Co-occurrence statistics as an efficient approximation to the original text

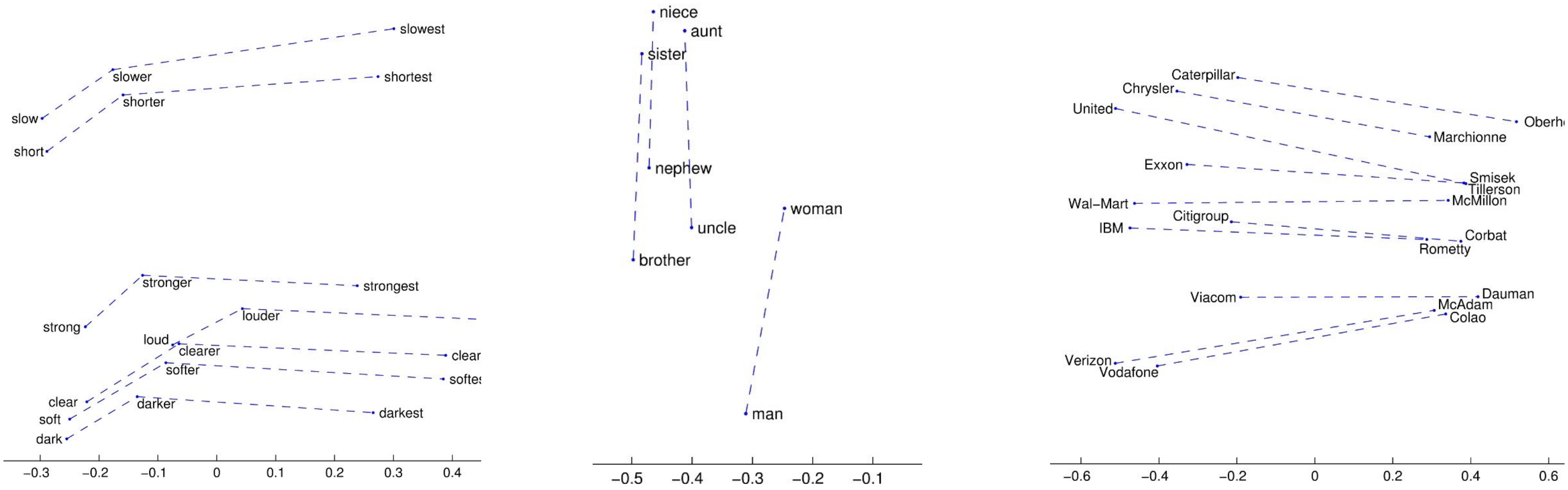
Probability and Ratio	$k = solid$	$k = gas$	$k = water$	$k = fashion$
$P(k\|ice)$	1.9×10^{-4}	6.6×10^{-5}	3.0×10^{-3}	1.7×10^{-5}
$P(k\|steam)$	2.2×10^{-5}	7.8×10^{-4}	2.2×10^{-3}	1.8×10^{-5}
$P(k\|ice)/P(k\|steam)$	8.9	8.5×10^{-2}	1.36	0.96

$$J = \sum_{i,j=1}^{V} f\left(X_{ij}\right)\left(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log X_{ij}\right)^2$$

Pre-Trained Word Embeddings

LSA [Deer-wester+ 1990]
skip-gram [Mikolov + 2013]
Glove [Pennington+ 2014]

Compact representations for **syntax**, **common sense** and **world knowledge**

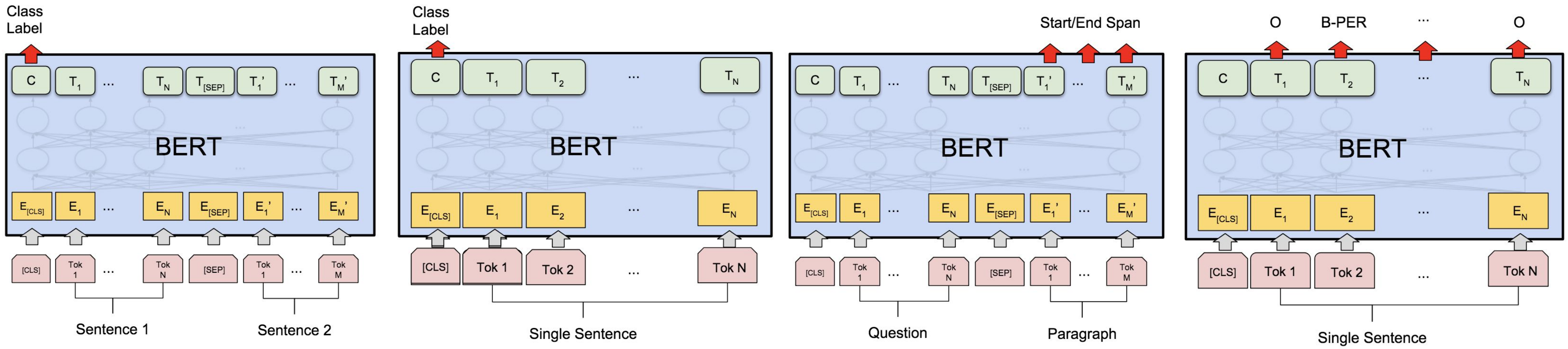


Glove [Pennington+ 2014]

[Devlin+ 2018]
[Radford+ 2018]
[Peters+ 2018]

Pre-Trained Sequence Models

With a lot more **computational power** we have **language modeling sequence models** which converts sequences of tokens to sequences of **contextual embeddings**



(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG

(b) Single Sentence Classification Tasks: SST-2, CoLA

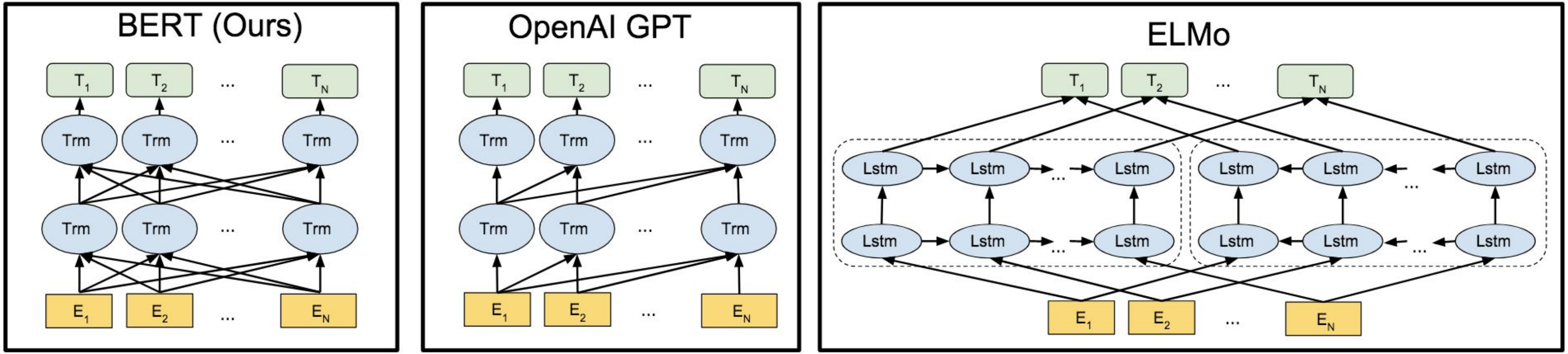
(c) Question Answering Tasks: SQuAD v1.1

(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

[Devlin+ 2018]
[Radford+ 2018]
[Peters+ 2018]

Pre-Trained Sequence Models

Differences in pre-training model architectures.
BERT uses a bidirectional Transformer. OpenAI **GPTuses** a left-to-right Transformer.
ELMo uses the concatenation of independently trained left-to-right and right-to-left LSTM.



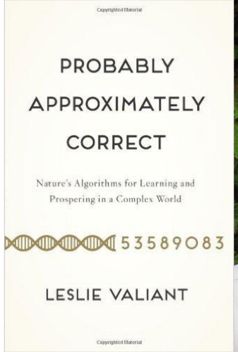
[Devlin+ 2018]
[Radford+ 2018]
[Peters+ 2018]

Pre-Trained Sequence Models

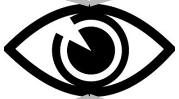
2018 is the year of Pre-Trained Sequence Models

System	MNLI-(m/mm)	QQP	QNLI	SST-2	CoLA	STS-B	MRPC	RTE	**Average**
	392k	363k	108k	67k	8.5k	5.7k	3.5k	2.5k	-
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.9	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	88.1	91.3	45.4	80.0	82.3	56.0	75.2
$\text{BERT}_{\text{BASE}}$	84.6/83.4	71.2	90.1	93.5	52.1	85.8	88.9	66.4	79.6
$\text{BERT}_{\text{LARGE}}$	**86.7/85.9**	**72.1**	**91.1**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**81.9**

Knowledge Representation & Scalability



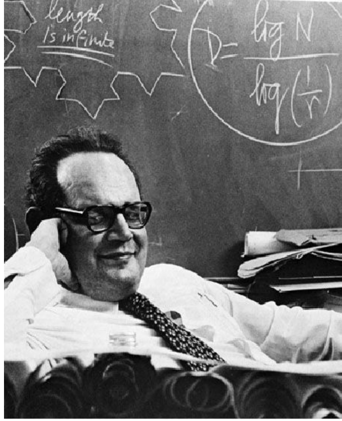
the world



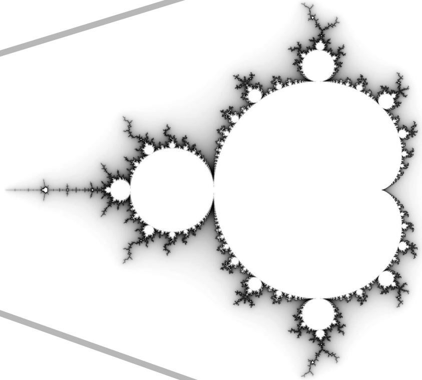
a small machine which copies the complexity of the world to the brain

a suitable representation

Mandelbrot Set



the nature of complex numbers



$$z_0 = 0$$

$$z_{n+1} = z_n^2 + c$$

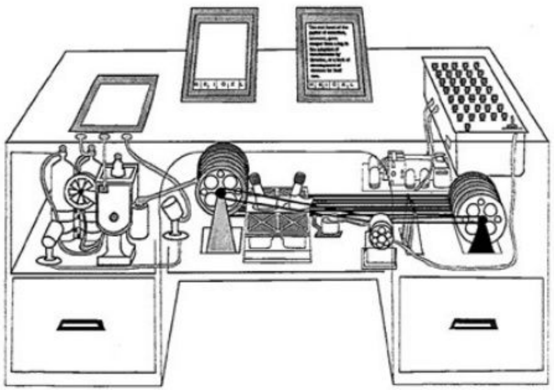
$$c \in M \iff \lim_{n \to \infty} |z_{n+1}| \leq 2$$

[Vannevar Bush 1945]

Internet as an external memory

- How information should be organized for scalability?

"AS WE MAY THINK"
(1945)

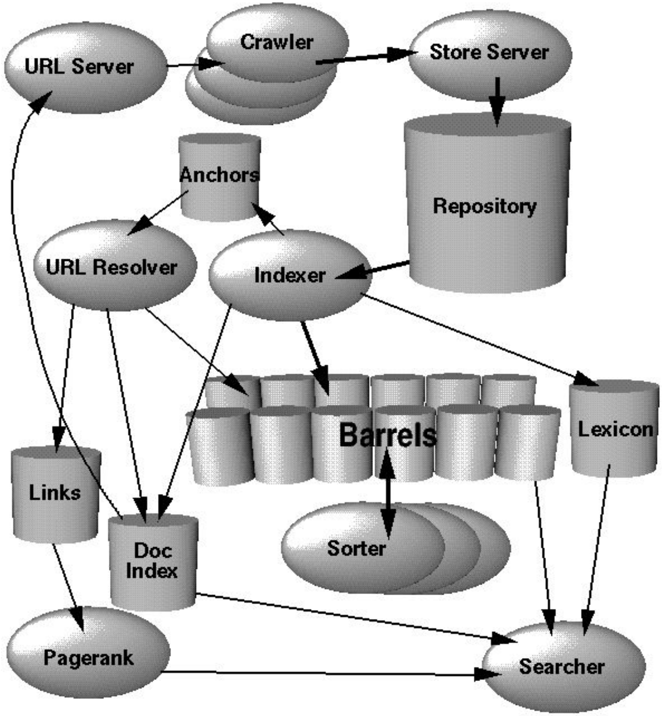


"Consider a future device for individual use, which is a sort of mechanized private file and library. It needs a name, and to coin one at random, memex will do. A memex is a device in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility. It is an enlarged intimate supplement to his memory.

[Brin & Page 1998]

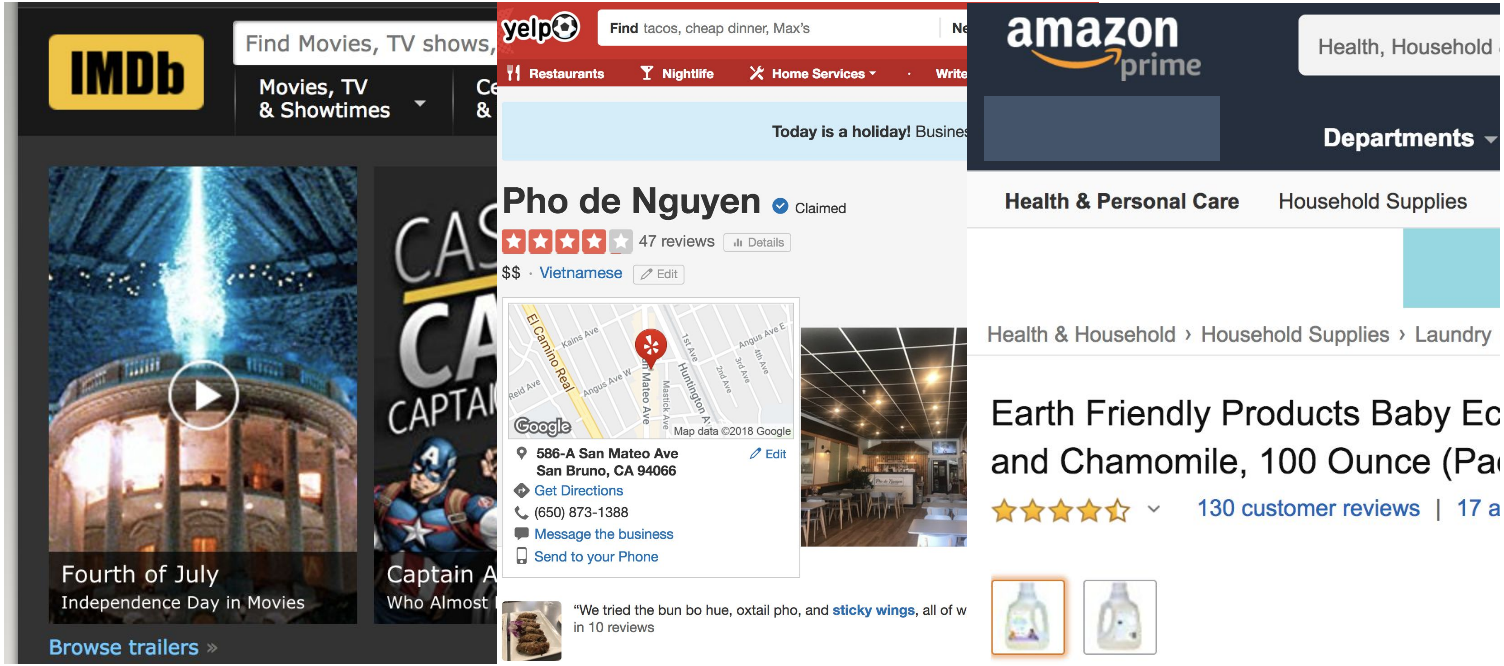
The scalability of modern search engines

- Can respond to user's requests within a fraction of a second
- But are weak at text understanding and complex reasoning



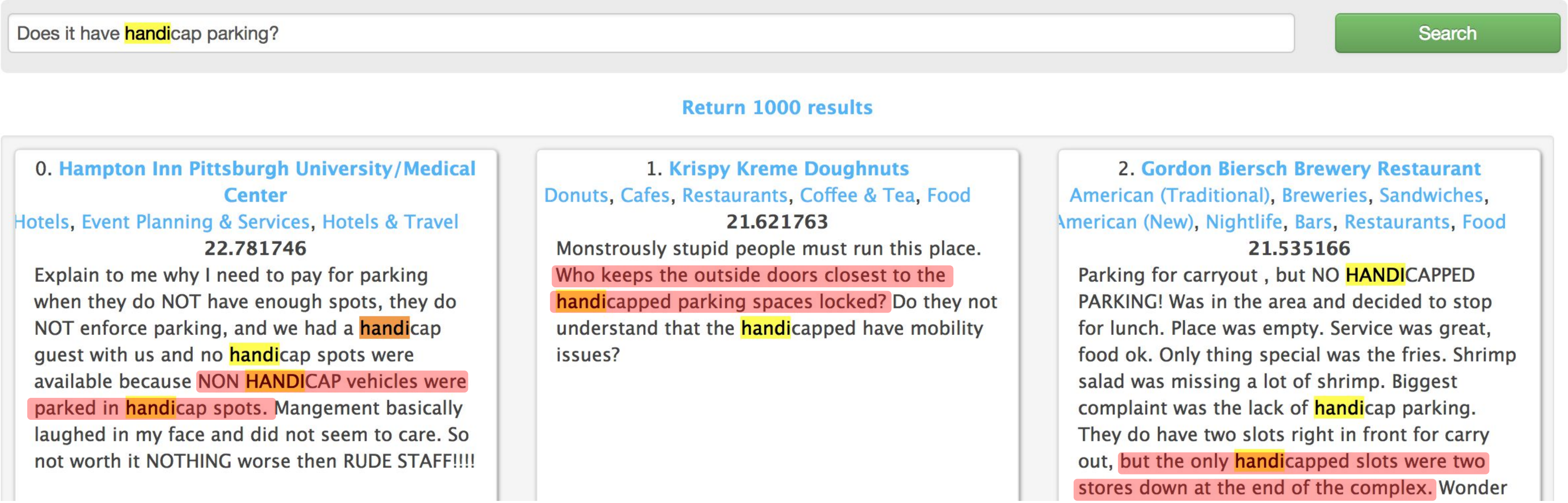
Can we search entities on the web?

- Multimedia, business, products have a lot of reviews and descriptions



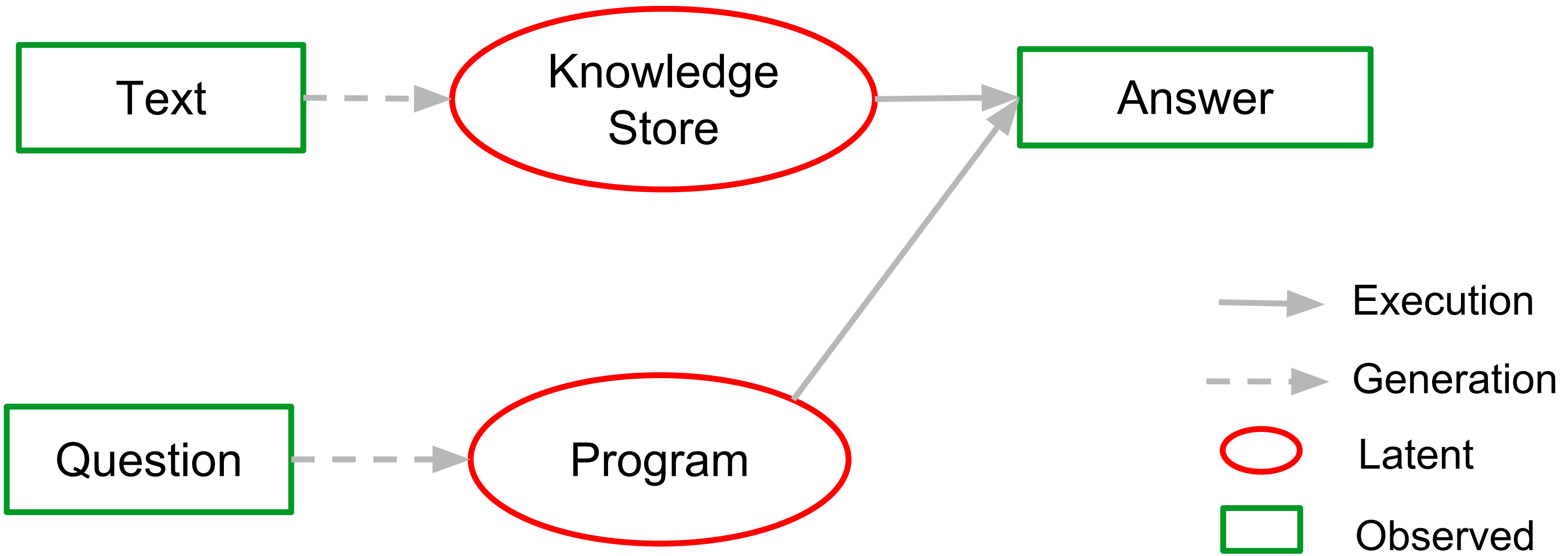
Traditional IR approach lacks understanding

- Need to interpret the meaning from the surface text



Question answering as a simple test bed

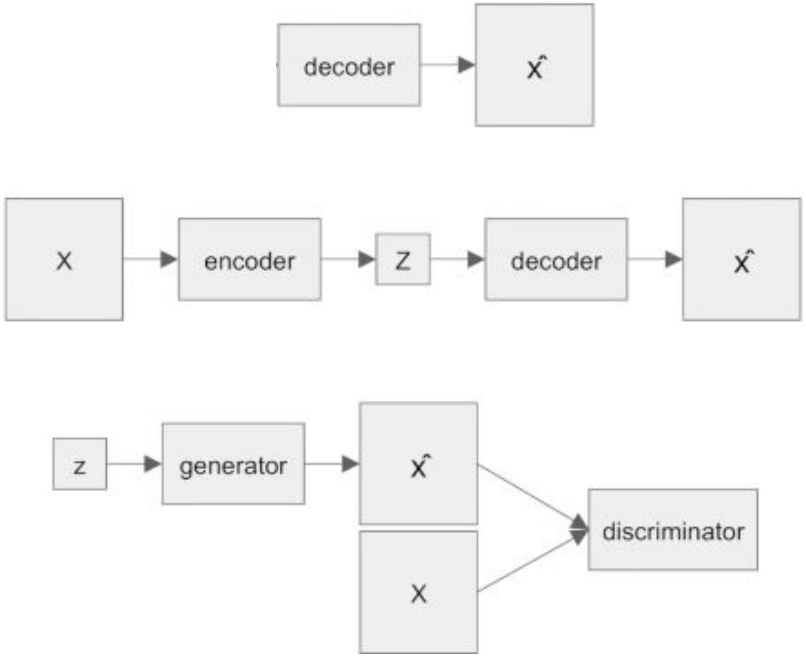
- A good semantic representation should support reasoning at scale



[Hochreiter & Schmidhuber 1997][Graves 2013][Sutskever, Vinyals, Le 2014][Cho+ 2014]
[Kingma, Welling 2014]
[Goodfellow+ 2014]

Three approaches to generative models

- Autoregression (e.g., LM), VAE, GAN



Autoregressive models (e.g. LM)
[Hochreiter & Schmidhuber 1997]
Graves [1308.0850]

Variational Autoencoders (VAE)
Kingma and Welling [1312.6114]

Generative Adversarial Networks (GAN)
Goodfellow et al. [1406.2661]

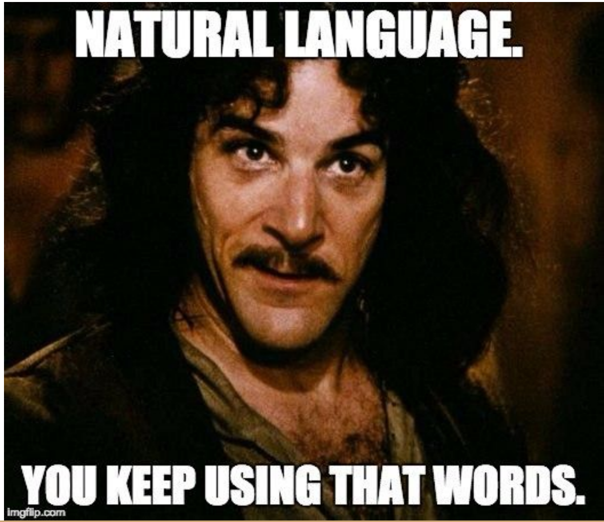
Blog [Karpathy+ 2016]

[Rajeswar+ 2017]

Immediately criticised when applied to text

- "I have a lot of respect for language. Deep-learning people seem not to"
- "They include such impressive natural language sentences as:"
    - * what everything they take everything away from
    - * how is the antoher headache
    - * will you have two moment ?
    - * This is undergoing operation a year .
- "These are not even grammatical!"
- The DNN bubble consists of models, which show great promises but not yet practical at this point

Blog [Goldberg 2017]

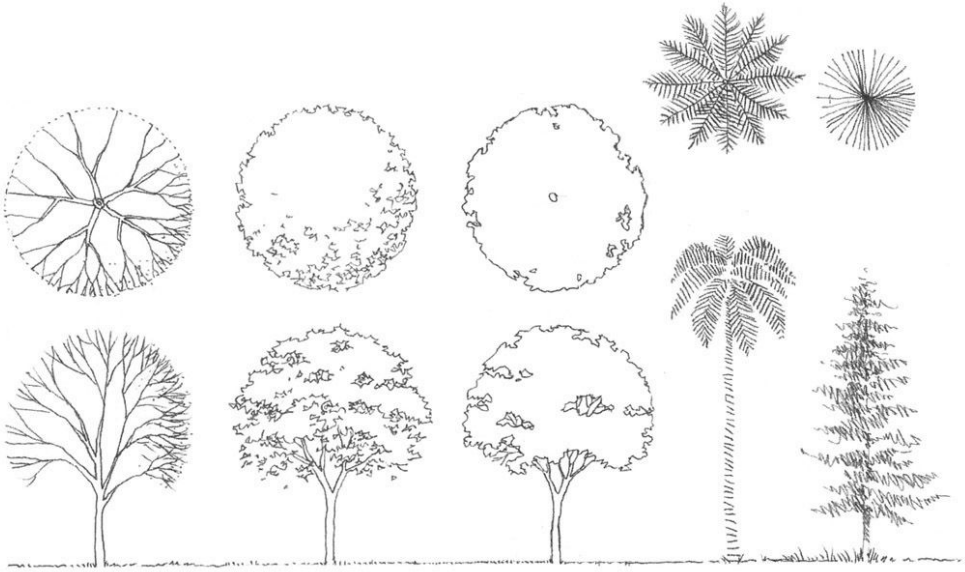


Seq2Seq [Sutskever, Vinyals, Le 2014]
VAE [Kingma & Welling 2014]
GAN [Goodfellow+ 2014]
ACL [Goldberg 2015]
ACL [Mooney 2015]

furious

statistician’s view v.s. linguist’s view

Given the power of deep learning anything can be mapped to a unit Gaussian ball



The world has real structures, which need to be represented by real structures

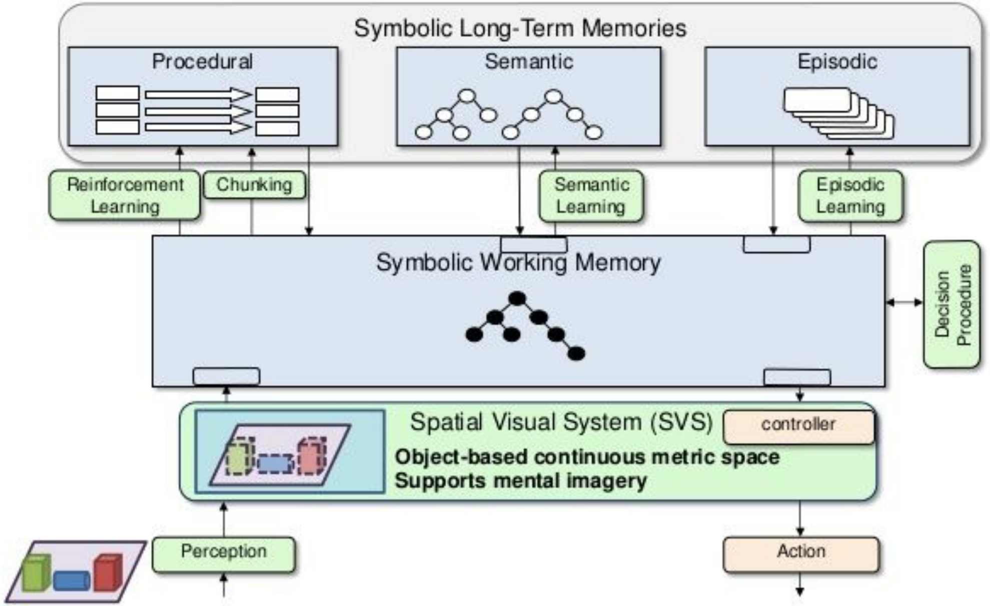
Scalability of mammal memory

- Very rapid adaptation (in just one or a few trials) is necessary for survival
  - E.g., associating taste of food and sickness
- Need fast responses based on large amount of knowledge
  - Needs good representation of knowledge
- However, good representation can only be learnt gradually
  - During sleeps to prevent interference with established associations

[Garcia+ 1966]
[Wickman 2012]
[Bartol+ 2015]

Scalability of Cognitive Architectures

- The design of mammalian brains is inspiring to NLP systems
  - they are solving similar problems
- The design has not changed much since 30 years ago
  - “We’ve totally solved it already … it is just a matter of job security”
    -- Nate Derbinsky, Northeastern U.
- Today we have
  - internet economy and data
  - computation and ML development



https://soar.eecs.umich.edu

Plan

Access slides and join discussions at
weakly-supervised-nlu google group

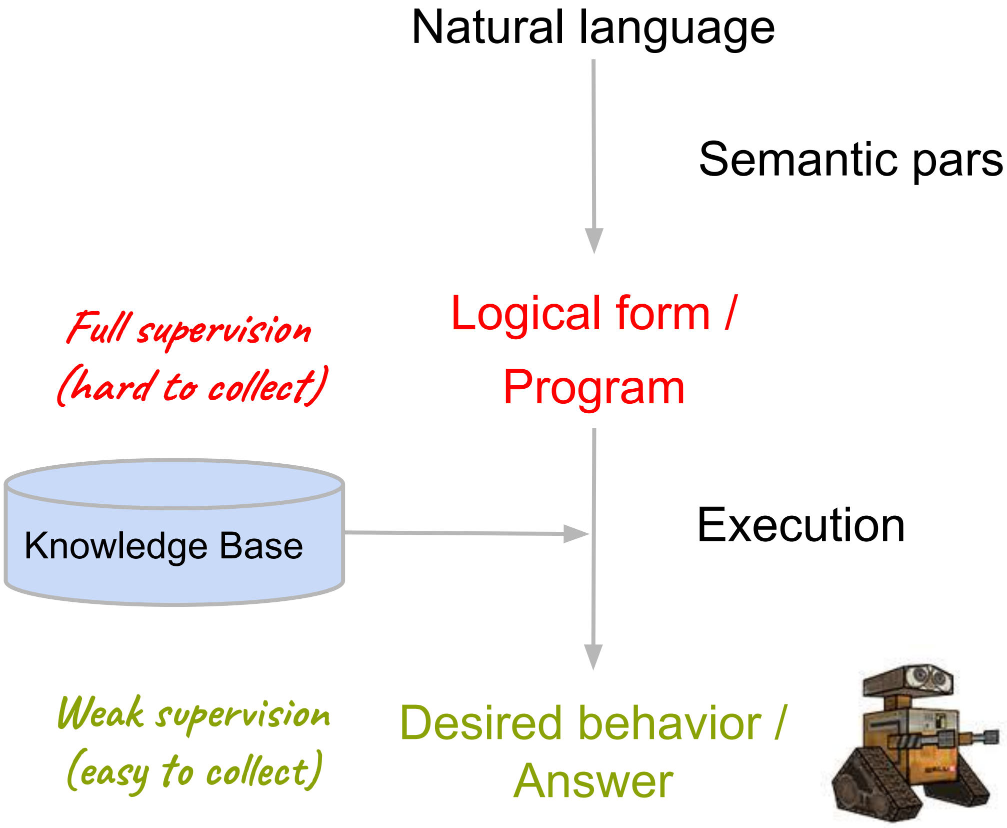
Mobile

Desktop

- ***Weak Supervision NLP***
  - NLP, AI, software 2.0
  - Semantics as a foreign language
  - Unsupervised learning
  - Knowledge representation (symbolism)
- ***Semantic Parsing Tasks***
  - *WebQuestionsSP, WikiTableQuestions*
- ***Neural Symbolic Machines*** (ACL 2017)
  - Compositionality (short term memory)
  - Scalable KB inference (symbolism)
  - RL vs MLE
- ***Memory Augmented Policy Optimization*** (NIPS 2018)
  - Experience replay (long term memory & optimal updating strategy)
  - Systematic exploration
  - Memory Weight Clipping (unbiased cold start strategy)

Semantic Parsing

- Natural language queries or commands are converted to computation steps on data and produce the expected answers or behavior



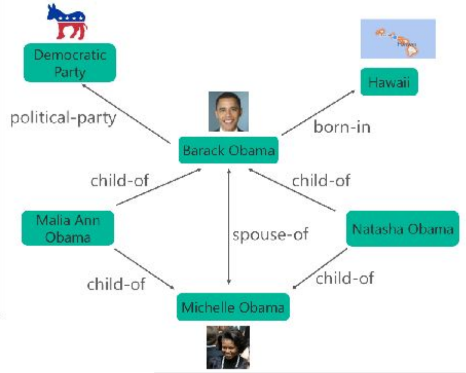
Related Works

- Training from full supervision is labor-intensive
  - DeepCoder [Balog, 2016]
  - NPI [Reed & Freitas, 2015]
  - Seq2Tree [Dong & Lapata, 2016]
  - …
- Traditional semantic parsing models require feature engineering
  - SEMPRE [Berant et al, 2013]
  - STAGG [Yih et al, 2015]
  - …
- End-to-end differentiable models cannot scale to large databases
  - Neural Programmer [Neekalantan et al, 2015]
  - Neural Turing Machines [Graves et al, 2014]
  - Neural GPU [Kaiser & Sutskever, 2015]
  - …
- Combine deep learning, symbolic reasoning and reinforcement learning
  - Neural Symbolic Machines (Liang, Berant, Le, Forbus, Lao, 2017)
  - Memory Augmented Policy Optimization (MAPO) (Liang, Norouzi, Berant, Le, Lao, 2018)

Question Answering with Knowledge Base

Largest city in US?

```
GO
(Hop V1 CityIn)
(Argmax V2 Population)
RETURN
```



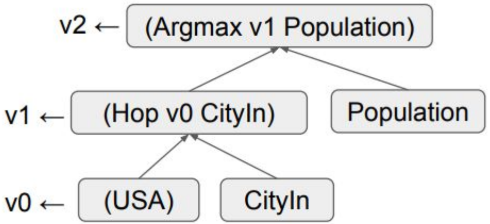
Freebase, DBpedia, YAGO, NELL

NYC

Paraphrase

Many ways to ask the same question, e.g.,
“What was the date that Minnesota became a state?”
“When was the state Minnesota created?”

Compositionality



Large Search Space (Optimization)

E.g., Freebase:
23K predicates,
82M entities,
417M triplets

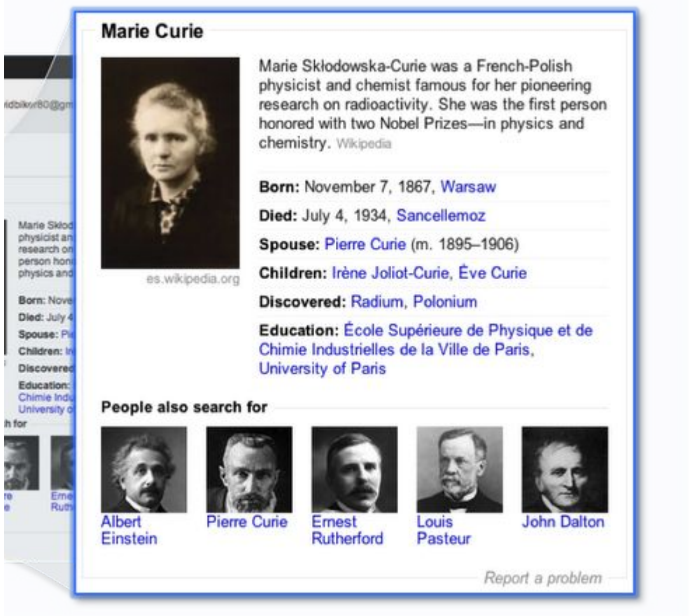
[Singhal, Google 2012]
[Qian, Bing 2013]
[Berant+ 2013]

WebQuestions: motivation

Motivation: Natural language interface to large structured knowledge-bases

Background: availability of many structured datasets (Google KG, Bing Satori, Freebase, DBPedia, Yelp, ...)

Introducing the Knowledge Graph: things, not strings



WebQuestions: getting questions

[Berant+ 2013]

Goal: collect large number of natural language queries

Strategy: breadth-first search over Google Suggest graph results in 1M queries

Where was Barack Obama born?

Where was __ born? **Google Suggest** Barack Obama, Lady Gaga, Steve Jobs

Where was Steve Jobs born?

Where was Steve Jobs __? **Google Suggest** born, raised, on the Forbes list

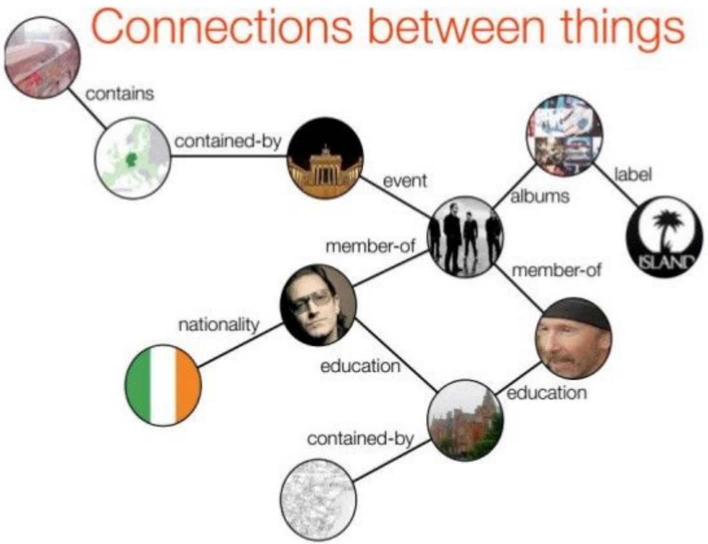
Where was Steve Jobs raised?

Freebase

Motivation: overcome the limitations of computers -- machines struggle to absorb knowledge in the way humans do.

Solution: a large collaborative knowledge base consisting of data (int triple format) composed mainly by its community members.

Later acquired by Google (discontinued)



A network of facts. © Metaweb/Google

41M entities (nodes)
19K properties (edge labels)
596M assertions (edges)

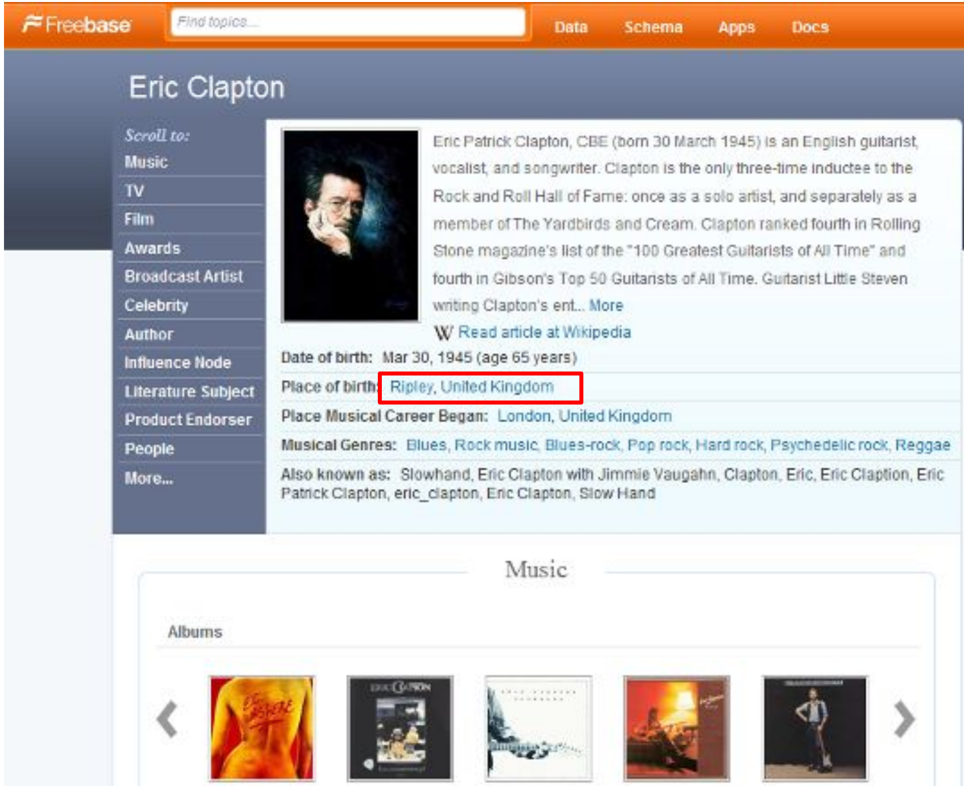
WebQuestions: getting Freebase answers

[Berant+ 2013]

Goal: obtain label from non-experts

Strategy: Amazon Mechanical Turk (AMT)

- Given a query e.g. “Eric Clapton hometown” detect if there is a single named entity “Eric Clapton”
- Ask the worker to pick one (or more) of the possible entities/values (“Ripley”) on the entity Freebase page
- Cost $0.03 per question



Freebase.com page circa 2011 © Metaweb/Google

[Berant et al, 2013; Yih et al, 2016]

WebQuestionsSP Dataset

- 5,810 questions from Google Suggest API & Amazon MTurk[1]
- Remove invalid QA pairs[2]
- 3,098 training examples, 1,639 testing examples remaining
- Open-domain and contains grammatical error
- Multiple entities as answer => macro-averaged F1

Grammatical error	Multiple entities
What do Michelle Obama do for a living?	writer, lawyer
What character did Natalie Portman play in Star Wars?	Padme Amidala
What currency do you use in Costa Rica?	Costa Rican colon
What did Obama study in school?	political science
What killed Sammy Davis Jr?	throat cancer

[Berant+ 2013]
[Lin+ 2012]
[Fader+ 2011]

SEMPRE: paraphrase

Collect entity pair observations for phrases and augmented predicates
(which partially solves the search problem)

ClueWeb09
1 billion docs

ReVerb

15M triples (of 15k phrases)

(*Barack Obama*, *was born in*, *Honolulu*)
(*Albert Einstein*, *was born in*, *Ulm*)
(*Barack Obama*, *lived in*, *Chicago*)
... ...

Freebase

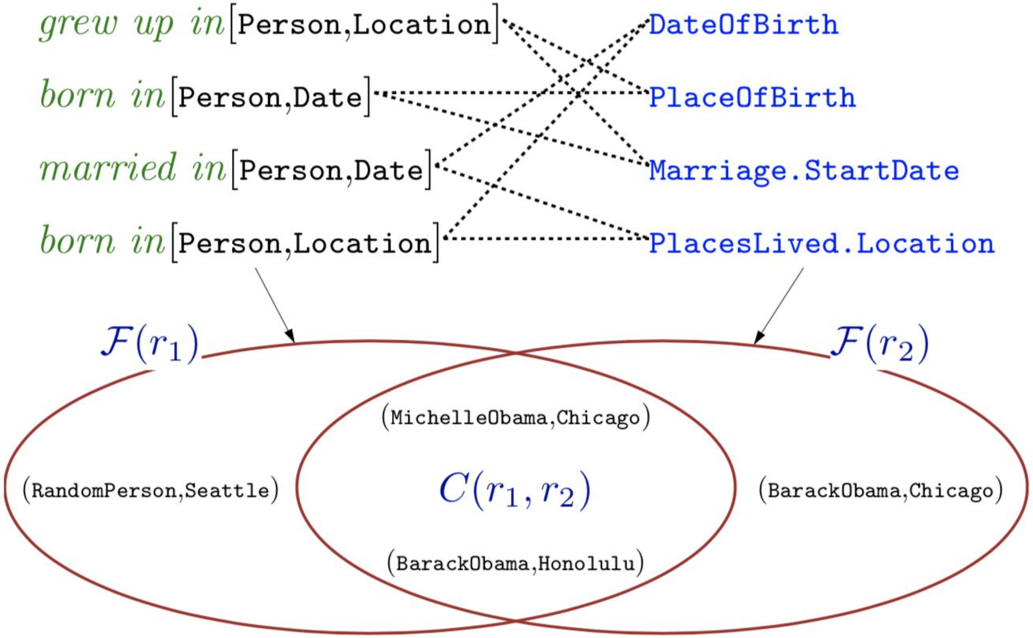
600M triples (of 60k augmented predicates)

(BarackObama, PlaceOfBirth, Honolulu)
(Albert Einstein, PlaceOfBirth, Ulm)
(BarackObama, PlacesLived.Location, Chicago)
... ...

[Berant+ 2013]

SEMPRE: paraphrase

- Construct lexicon and alignment features based on entity pair cooccurrences



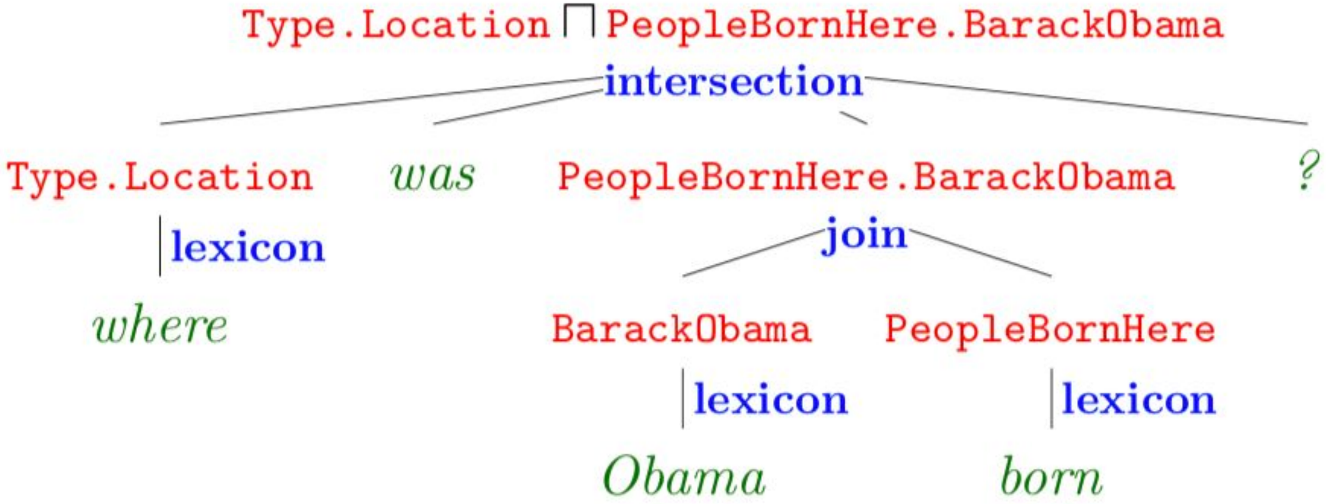
Alignment features

log-phrase-count:	log(15765)
log-predicate-count:	log(9182)
log-intersection-count:	log(6048)
KB-best-match:	0

SEMPRE: compositionality

The **semantic** structure (logic form) is coupled with **syntactic** structure (surface patterns) through composition rules

One derivation



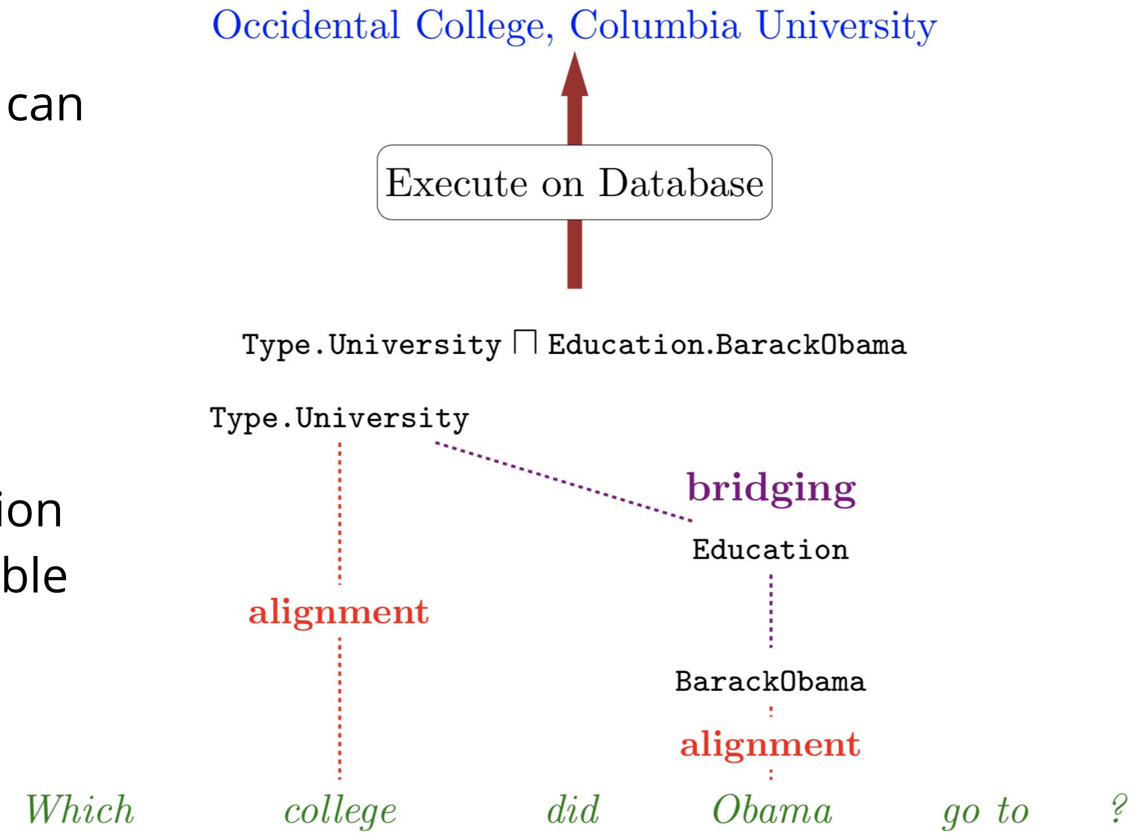
SEMPRE: search: bridging

[Berant+ 2013]

Motivation: individual words can be highly ambiguous

- What government does Chile have?
- What is Italy's language?
- Where is Beijing?
- What is the cover price of X-men?

Solution: the bridging operation generates predicates compatible with neighboring predicates



[Berant+ 2013]

SEMPRE: modeling

Log linear model over derivations *d* given utterance *x* with expert designed features

$$p_\theta(d \mid x) = \frac{\exp\{\phi(x,d)^\top \theta\}}{\sum_{d' \in D(x)} \exp\{\phi(x,d')^\top \theta\}}$$

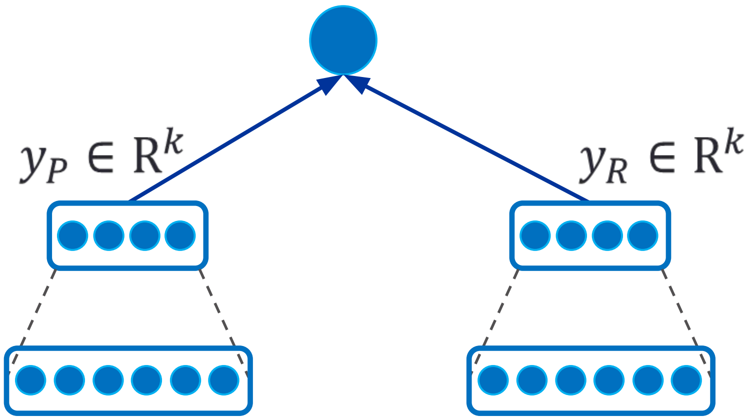
Category	**Description**
Alignment	Log of # entity pairs that occur with the phrase r_1 ($\lvert\mathcal{F}(r_1)\rvert$) Log of # entity pairs that occur with the logical predicate r_2 ($\lvert\mathcal{F}(r_2)\rvert$) Log of # entity pairs that occur with both r_1 and r_2 ($\lvert\mathcal{F}(r_1) \cap \mathcal{F}(r_2)\rvert$) Whether r_2 is the best match for r_1 ($r_2 = \arg\max_r \lvert\mathcal{F}(r_1) \cap \mathcal{F}(r)\rvert$)
Lexicalized	Conjunction of phrase w and predicate z
Text similarity	Phrase r_1 is equal/prefix/suffix of s_2 Phrase overlap of r_1 and s_2
Bridging	Log of # entity pairs that occur with bridging predicate b ($\lvert\mathcal{F}(b)\rvert$) Kind of bridging (# unaries involved) The binary b injected
Composition	# of intersect/join/bridging operations POS tags in join/bridging and skipped words Size of denotation of logical form

Table 1: Full set of features. For the alignment and text similarity, r_1 is a phrase, r_2 is a predicate with Freebase name s_2, and b is a binary predicate with type signature (t_1, t_2).

[Yih+ 2015]
[Shen+ 2014]
[Gabrilovich+ 2013]

STAGG: paraphrase

3 matching models based on char-ngram conv nets (Sent2vec [Shen+ 14])



Pattern-Chain:	who voiced meg on <e>	cast-actor
Question-EntPred:	who voiced meg on family guy	Meg Griffin cast-actor
ClueWeb12:	voiced homer on <e>	cast-actor

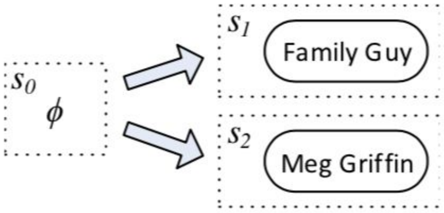
[Yih+ 2015]
[Yang & Chang 2015]

STAGG: search

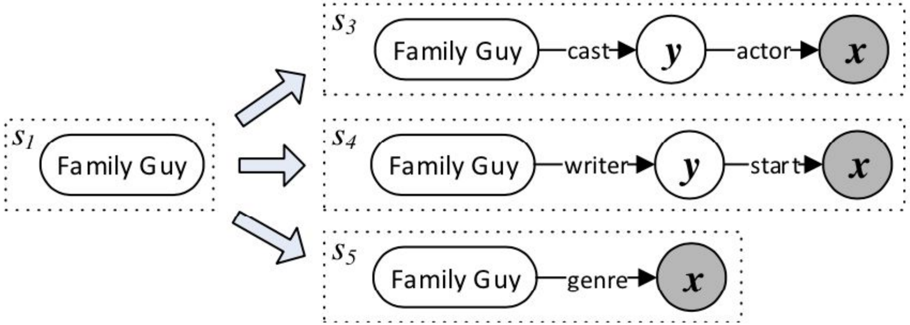
Staged Query Graph Generation

"Who first voiced Meg on Family Guy?"

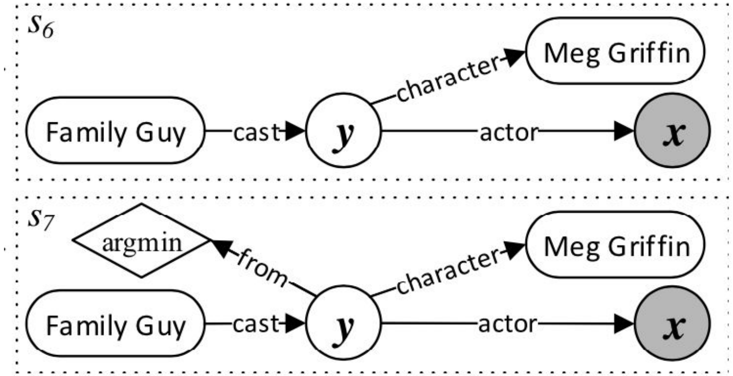
1) decide the topic entity



2) decide the core inference chain



3) add type/aggregation constraints



STAGG: search

[Yih+ 2015]

keeps up to N candidate states in the priority queue (N = 1000)

+10 special rules to restrict the type/aggregation constraints

> e.g. Consider "to" predicates (indicating the ending time of an event) only when the question contains "last", "latest" or "newest"

Domain knowledge is often very important for structure search problems. Will revisit in WikiTableQ

Algorithm 1 Staged query graph generation

Require: Priority queue H with limited size N

$s_o \leftarrow \phi$; $r_o \leftarrow -\infty$
H.add(s_o,r_o)
while H is not empty **do**
    $s, r \leftarrow H$.pop()
    if $r > r_o$ **then**
        $s_o \leftarrow s$; $r_o \leftarrow r$;
    end if
    for all $a \in \Pi(s)$ **do**
        $s' \leftarrow T(s, a)$
        H.add($s', \gamma(s')$)
    end for
end while
return s_o

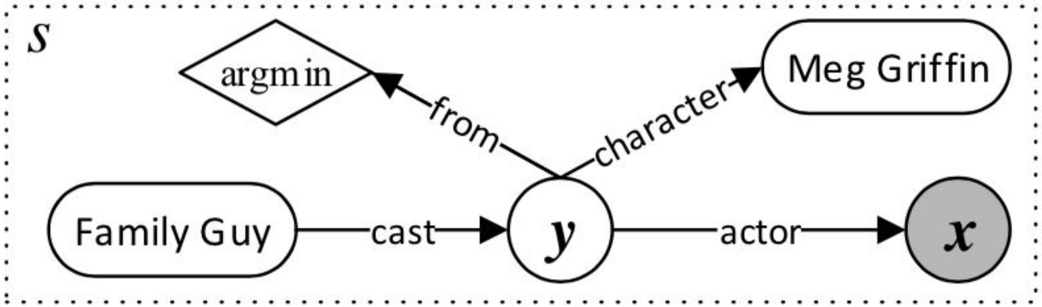
[Yih+ 2015]
[Burges 2010]

STAGG: modeling

Learning to rank model (for query graph candidates) based on expert designed features

+10 special features on the type/aggregation constraints

q = "Who first voiced Meg on Family Guy?"



(1) EntityLinkingScore(FamilyGuy, "Family Guy") = 0.9
(2) PatChain("who first voiced meg on <e>", cast-actor) = 0.7
(3) QuesEP(q, "family guy cast-actor") = 0.6
(4) ClueWeb("who first voiced meg on <e>", cast-actor) = 0.2
(5) ConstraintEntityWord("Meg Griffin", q) = 0.5
(6) ConstraintEntityInQ("Meg Griffin", q) = 1
(7) AggregationKeyword(argmin, q) = 1
(8) NumNodes(s) = 5
(9) NumAns(s) = 1

Brief Summary & Preview

DL leads to the death of feature engineering (but not domain knowledge)
DL makes it easier to leverage pre-trained unsupervised models

	SEMPRE/STAGG	**This talk**
Paraphrase (semi-supervised)	Co-occurrences collected from 1B docs (ClueWeb) and Freebase	Embeddings trained from 840B text tokens (GloVe)
Compositionality	Relies on syntactic structure (text spans) and domain specific rules to constrain the generation of logic forms	A (LISP like) language specify computations on KG
Search	priority queue & heuristic rules	RL & a program interpreter with syntactical & semantic checks
Modeling	Expert designed features	Deep learning

[Pasupat & Liang, 2015]

WikiTableQuestions: Dataset

Breadth

- No fixed schema: Tables at test time are not seen during training, needs to generalize based on column name.

Depth

- More compositional questions, thus require longer programs

- More operations like arithmetic operations and aggregation operations

Year	City	Country	Nations
1896	Athens	Greece	14
1900	Paris	France	24
1904	St. Louis	USA	12
...	...	...	...
2004	Athens	Greece	201
2008	Beijing	China	204
2012	London	UK	204

x_1: *"Greece held its last Summer Olympics in which year?"*
y_1: $\{2004\}$

x_2: *"In which city's the first time with at least 20 nations?"*
y_2: $\{\text{Paris}\}$

x_3: *"Which years have the most participating countries?"*
y_3: $\{2008, 2012\}$

x_4: *"How many events were in Athens, Greece?"*
y_4: $\{2\}$

x_5: *"How many more participants were there in 1900 than in the first year?"*
y_5: $\{10\}$

[Pasupat & Liang, 2015]
[Liang+ 2018]

WikiTableQuestions: semantics

Function	Arguments	Returns	Description
(**hop** v p)	**v**: a list of rows. **p**: a column.	a list of cells.	Select the given column of the given rows.
(**argmax** v p) (**argmin** v p)	**v**: a list of rows. **p**: a number or date column.	a list of rows.	From the given rows, select the ones with the largest / smallest value in the given column.
(**filter**$_{>}$ v q p) (**filter**$_{\geqslant}$ v q p) (**filter**$_{<}$ v q p) (**filter**$_{\leqslant}$ v q p) (**filter**$_{=}$ v q p) (**filter**$_{\neq}$ v q p)	**v**: a list of rows. **q**: a number or date. **p**: a number or date column.	a list of rows.	From the given rows, select the ones whose given column has certain order relation with the given value.
(**filter**$_{in}$ v q p) (**filter**$_{!in}$ v q p)	**v**: a list of rows. **q**: a string. **p**: a string column.	a list of rows.	From the given rows, select the ones whose given column contain / do not contain the given string.

WikiTableQuestions: semantics

[Pasupat & Liang, 2015]
[Liang+ 2018]

(**first** v) (**last** v)	**v**: a list of rows.	a row.	From the given rows, select the one with the smallest / largest index.
(**previous** v) (**next** v)	**v**: a row.	a row.	Select the row that is above / below the given row.
(**count** v)	**v**: a list of rows.	a number.	Count the number of given rows.
(**max** v p) (**min** v p) (**average** v p) (**sum** v p)	**v**: a list of rows. **p**: a number column.	a number.	Compute the maximum / minimum / average / sum of the given column in the given rows.
(**mode** v p)	**v**: a list of rows. **p**: a column.	a cell.	Get the most common value of the given column in the given rows.
* (**same_as** v p)	**v**: a row. **p**: a column.	a list of rows.	Get the rows whose given column is the same as the given row.
(**diff** v0 v1 p)	**v0**: a row. **v1**: a row. **p**: a number column.	a number.	Compute the difference in the given column of the given two rows.

WikiTableQuestions: search

[Pasupat & Liang, 2015]
[Zhang, Pasupat & Liang, 2017]
[Liang+ 2018]

Feature engineering is dead.
It is survived by **program space design.**

```
(when t-alternative
  (rule $AnchoredOr ($LEMMA_TOKEN) (FilterTokenFn lemma and or) (anchored 1))
...
(when t-movement
  (rule $AnchoredMovement ($LEMMA_TOKEN) (FilterTokenFn lemma next previous after before above below) (anchored 1))
...
(when t-comparison
  (rule $AnchoredMore ($LEMMA_TOKEN) (FilterTokenFn lemma more than least above after) (anchored 1))
...
(when t-superlative
  (rule $SuperlativeTrigger ($LEMMA_TOKEN) (FilterPosTagFn token JJR JJS RBR RBS) (anchored 1))
  (rule $SuperlativeTrigger ($LEMMA_TOKEN) (FilterTokenFn lemma top first bottom last) (anchored 1))
...
(when t-arithmetic
  (rule $AnchoredSub ($LEMMA_TOKEN) (FilterTokenFn lemma difference between and much) (anchored 1))
…
...
```

[Liang+, 2018]

WikiTableQuestions: example solutions

Superlative	
nt-13901: the most points were scored by which player?	
(argmax all_rows r.points-num)	Sort all rows by column 'points' and take the first row.
(hop v0 r.player-str)	Output the value of column 'player' for the rows in v0.
Difference	
nt-457: how many more passengers flew to los angeles than to saskatoon?	
(filter$_{in}$ all_rows ['saskatoon'] r.city-str)	Find the row with 'saskatoon' matched in column 'city'.
(filter$_{in}$ all_rows ['los angeles'] r.city-str)	Find the row with 'los angeles' matched in column 'city'.
(diff v1 v0 r.passengers-num)	Calculate the difference of the values in the column 'passenger' of v0 and v1.

[Liang+, 2018]

More examples

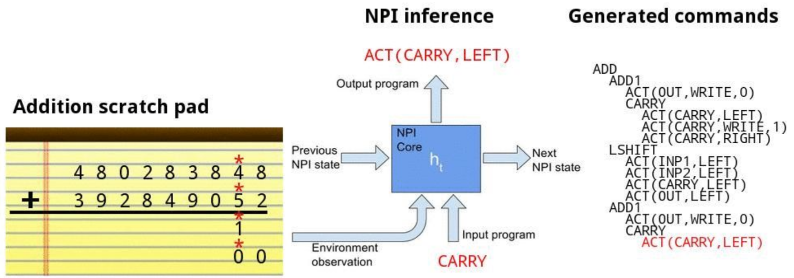
Before / After	
nt-10832: which nation is before peru?	
(filter$_{in}$ all_rows ['peru'] r.nation-str)	Find the row with 'peru' matched in 'nation' column.
(previous v0)	Find the row before v0.
(hop v1 r.nation-str)	Output the value of column 'nation' of v1.

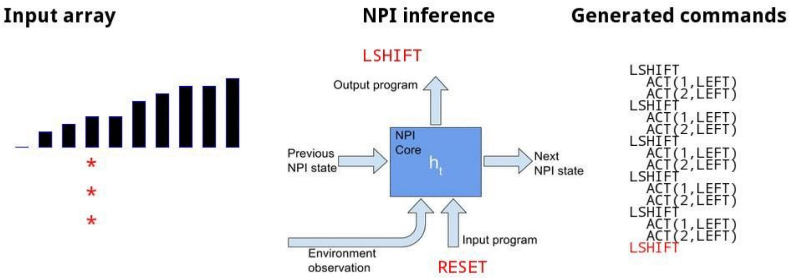
Compare & Count	
nt-647: in how many games did sri lanka score at least 2 goals?	
(filter$_{\geq}$ all_rows [2] r.score-num)	Select the rows whose value in the 'score' column >= 2.
(count v0)	Count the number of rows in v0.

Exclusion	
nt-1133: other than william stuart price, which other businessman was born in tulsa?	
(filter$_{in}$ all_rows ['tulsa'] r.hometown-str)	Find rows with 'tulsa' matched in column 'hometown'.
(filter$_{!in}$ v0 ['william stuart price'] r.name-str)	Drop rows with 'william stuart price' matched in the value of column 'name'.
(hop v1 r.name-str)	Output the value of column 'name' of v1.

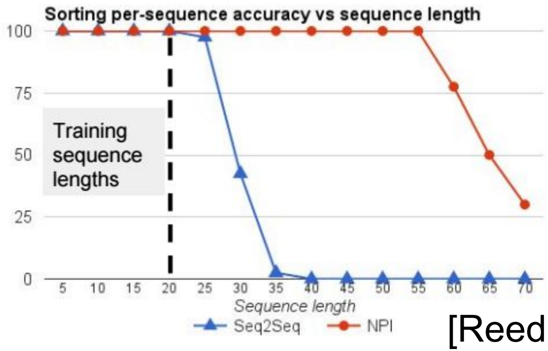
Neural Program Induction & Scalabilities

- Impressive works to show NN can learn addition and sorting, but...



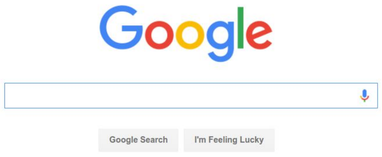


- The learned operations are not as scalable and precise.



[Reed & Freitas 2015]

- Why not use existing modules that are scalable, precise and interpretable?



[Zaremba & Sutskever 2016]

Plan

Access slides and join discussions at
weakly-supervised-nlu google group

Mobile

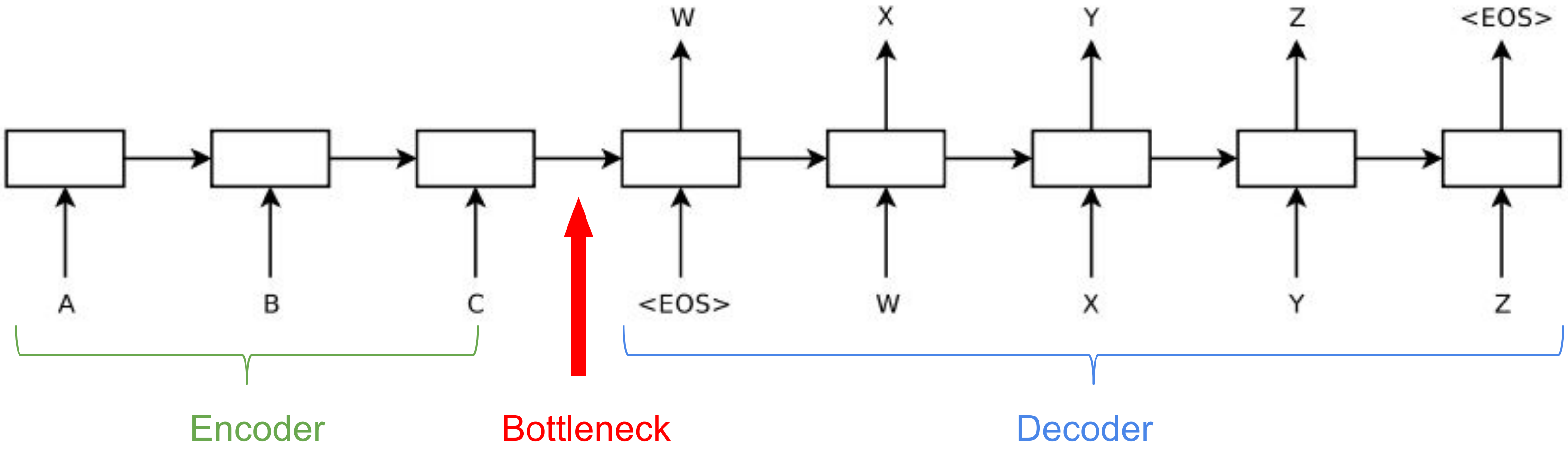
Desktop

- ***Weak Supervision NLP***
  - NLP, AI, software 2.0
  - Semantics as a foreign language
  - Unsupervised learning
  - Knowledge representation (symbolism)
- ***Semantic Parsing Tasks***
  - *WebQuestionsSP, WikiTableQuestions*
- ***Neural Symbolic Machines*** (ACL 2017)
  - Compositionality (short term memory)
  - Scalable KB inference (symbolism)
  - RL vs MLE
- ***Memory Augmented Policy Optimization*** (NIPS 2018)
  - Experience replay (long term memory & optimal updating strategy)
  - Systematic exploration
  - Memory Weight Clipping (unbiased cold start strategy)

[Sutskever, Vinyals, Le 2014]

A bit background on Seq2Seq

- Separate a sequence model in to encoder and decoder
- Improves a phrase-based SMT system by re-ranking top candidates
- Cannot perform well by itself due to the information bottleneck

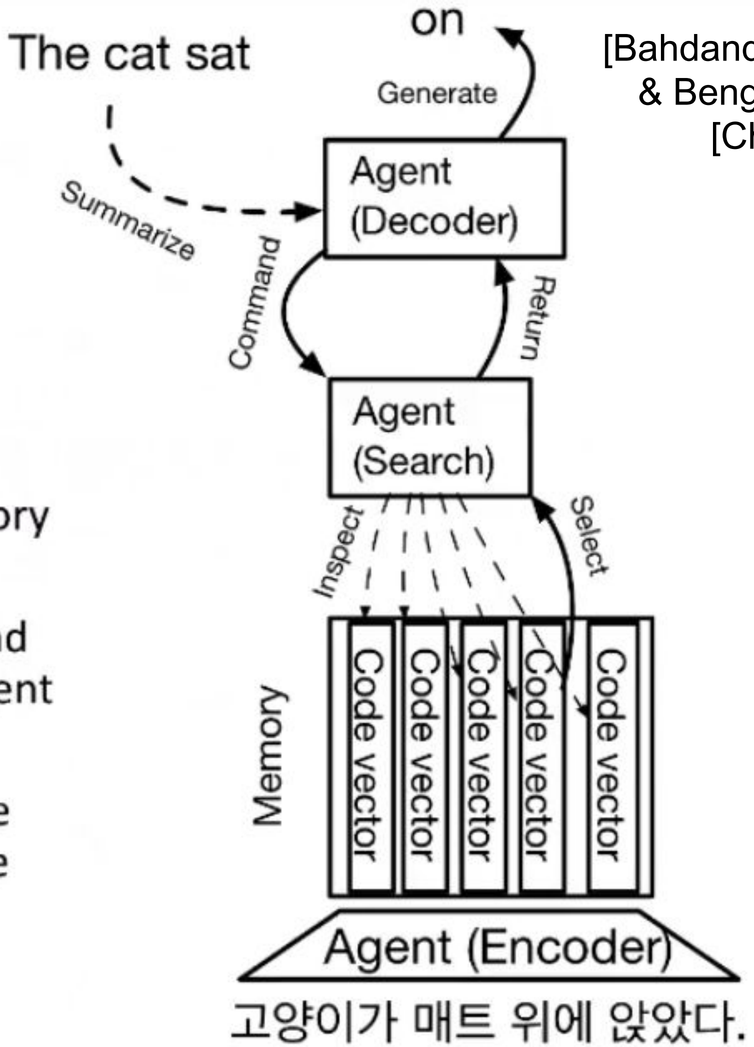


[Bahdandau, Cho & Bengio 2015]
[Cho 2017]

Re-thinking sequence-to-sequence learning

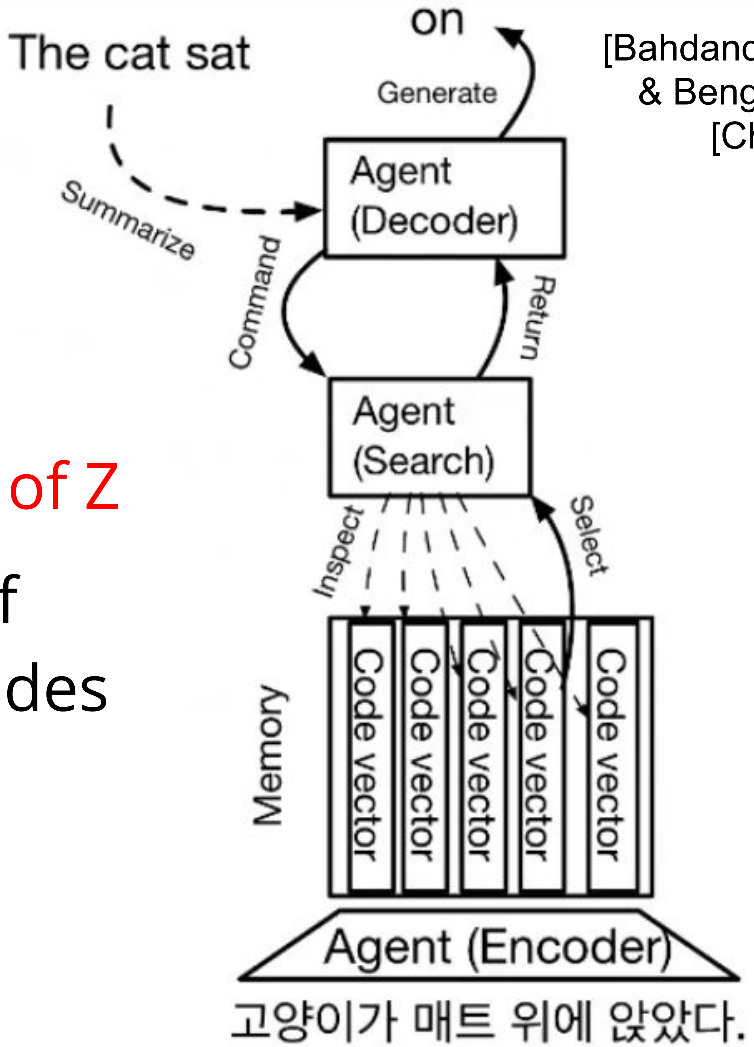
- Cooperation among three agents

1. **Agent 1** (Encoder): transforms the source sentence into a set of code vectors in a memory
2. **Agent 2** (Search): searches for relevant code vectors in the memory based on the command from the Agent 3 and returns them to the Agent 3.
3. **Agent 3** (Decoder): observes the current state (previously decoded symbols), commands the Agent 2 to find relevant code vectors and generates the next symbol based on them.



Re-thinking sequence-to-sequence learning

1. Don't generate from the ball of Z
2. Generate from a sequence of source token ids, which encodes the semantics of the target sentence

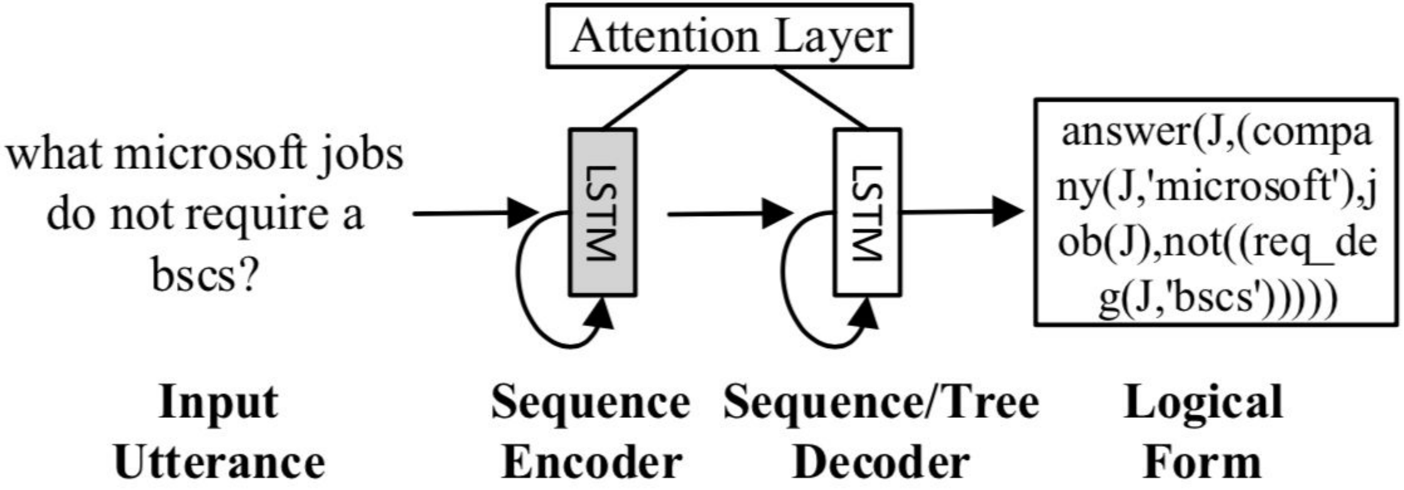


[Bahdandau, Cho & Bengio 2015]
[Cho 2017]

[Dong & Lapata, 2016]

Language to Logical Form with Neural Attention

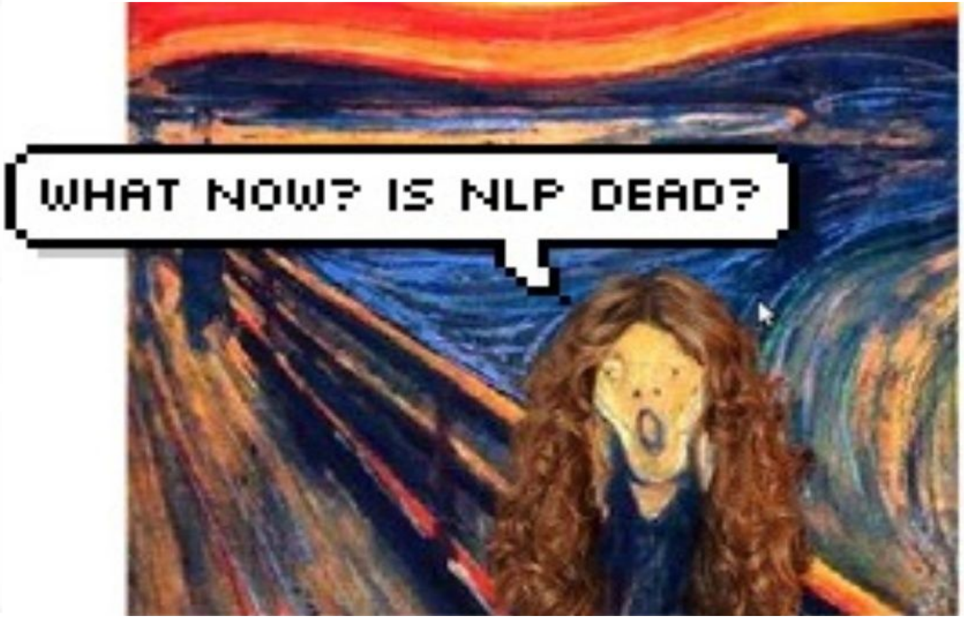
- "compared to previous methods our model achieves similar or better performance .. with no hand-engineered .. features."
- Relies on full supervision (labeled logical forms)



[Lapata ACL 2017]

LSTM & Lapata's scream

- LSTM has been applied to all kinds of NLP tasks and has greatly simplified system designs



What now? Is NLP DEAD?

- The need for **symbolic operations** and **effective model optimization**

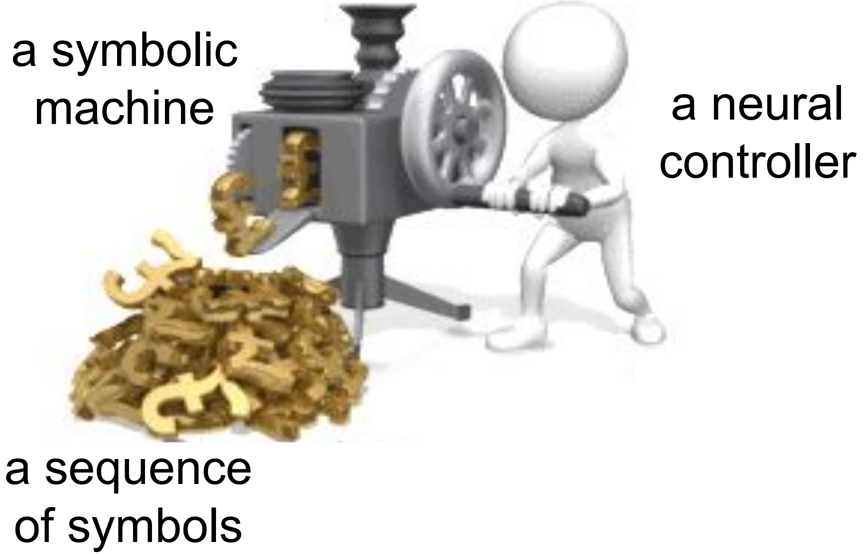
	SEMPRE/STAGG	**This talk**
Paraphrase (semi-supervised)	Co-occurrences collected from 1B docs (ClueWeb) and Freebase	Embeddings trained from 840B text tokens (GloVe)
Compositionality	Relies on syntactic structure (text spans) and domain specific rules to constrain the generation of logic forms	A (LISP like) language specify computations on KG
Search	priority queue & heuristic rules	RL & a program interpreter with syntactical & semantic checks
Modeling	Expert designed features	Deep learning

Connectionism vs Symbolism

The symbolic models represents elegant solutions to problems, and have been dominating AI for a very long time

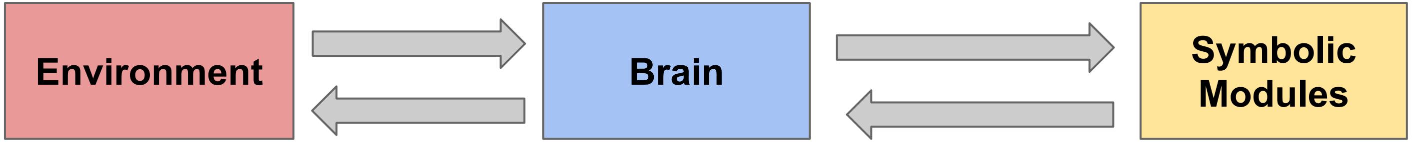
VS.

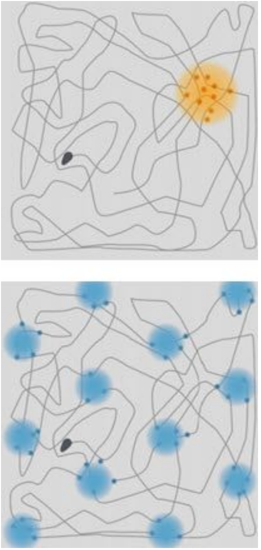
Once we have figured out how to train them, the connectionism approaches starts to win



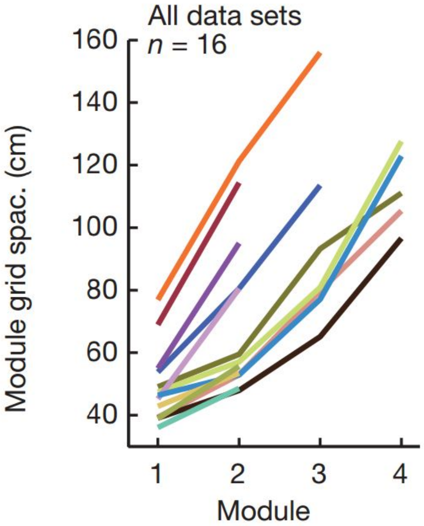
[Stensola+ 2012]

Symbolic Machines in Brains



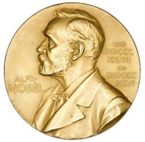


Location cells & grid cells



Grid spacing for all modules (M1–M4) in different animals

- 2014 Nobel Prize in Physiology or Medicine awarded for ‘inner GPS’ research



- Positions are represented as discrete representations in animals' brains, which enable accurate and autonomous calculations

Neural Symbolic Machines

ACL [Liang+ 2017]

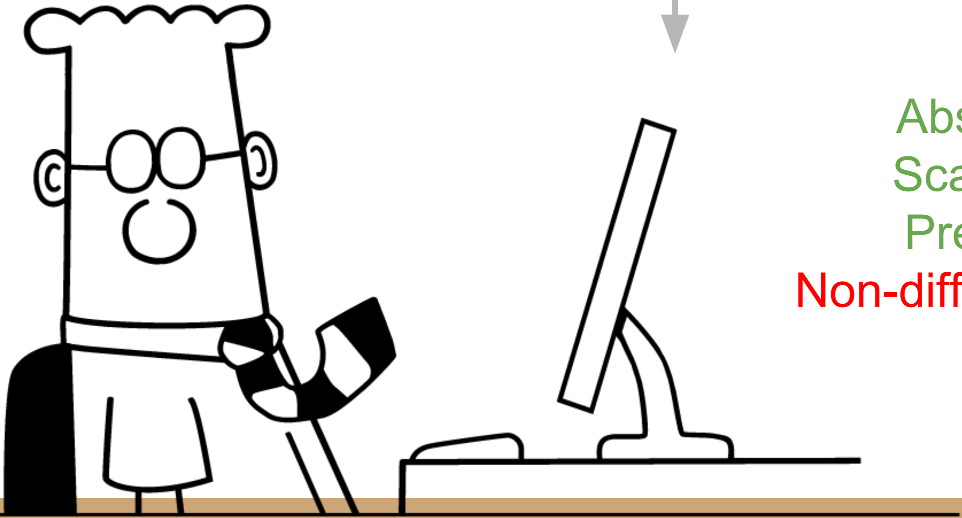
Weak supervision — **Neural** — **Symbolic**

Manager $\xrightarrow{\text{Question}}$ **Programmer** $\xrightarrow{\text{Program}}$ **Computer**

Manager $\xleftarrow{\text{Answer}}$ **Programmer** $\xleftarrow{\text{Results}}$ **Computer**

Knowledge Base → **Computer**

Predefined Functions → **Computer**



Abstract

Scalable

Precise

Non-differentiable

Computer with Domain Specific Languages

ACL [Liang+ 2017]

- Lisp Interpreter
  - Program => exp_1 exp_2 … exp_n <END>
  - Exp => (f arg_1 arg_2 … arg_n)

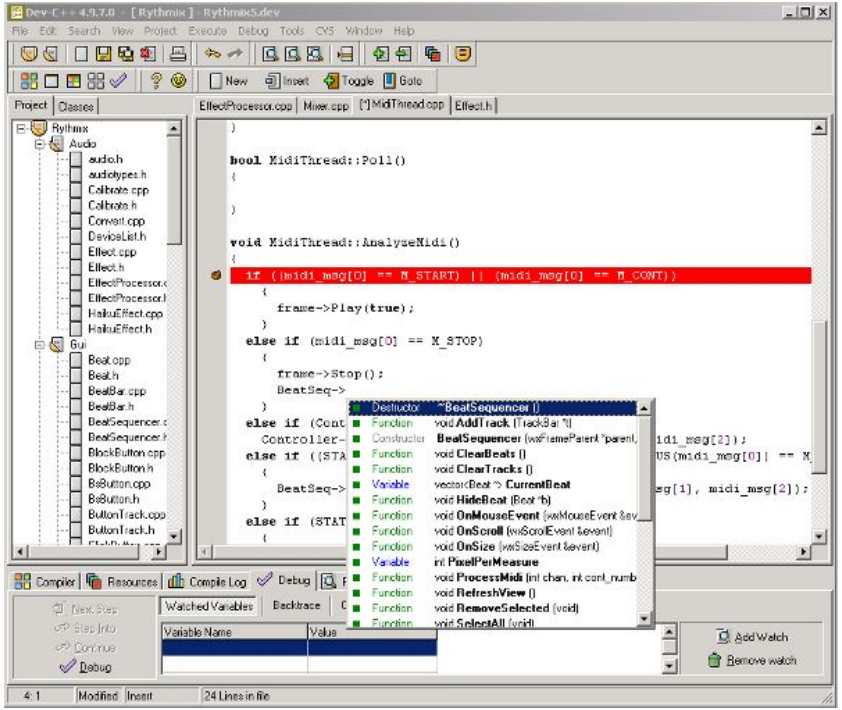
```
(hop m.russell_wilon /education)
(hop v0 /institution)
(filter= v1 m.univeristy
          /notable_types)
<END>
```

- What functions will be useful for the given task?
  - 10 operations for WebQuesitons
  - 22 different operations for WikiTableQuestions
    - hop
    - argmax, argmin
    - $filter_{=}$, $filter_{!=}$, $filter_{>}$, $filter_{<}$, $filter_{>=}$, $filter_{<=}$, $filter_{in}$, $filter_{!in}$,
    - first, last, previous, next
    - max, min, average, sum, mode, diff, same

Code Assistance to Prune Search Space

ACL [Liang+ 2017]

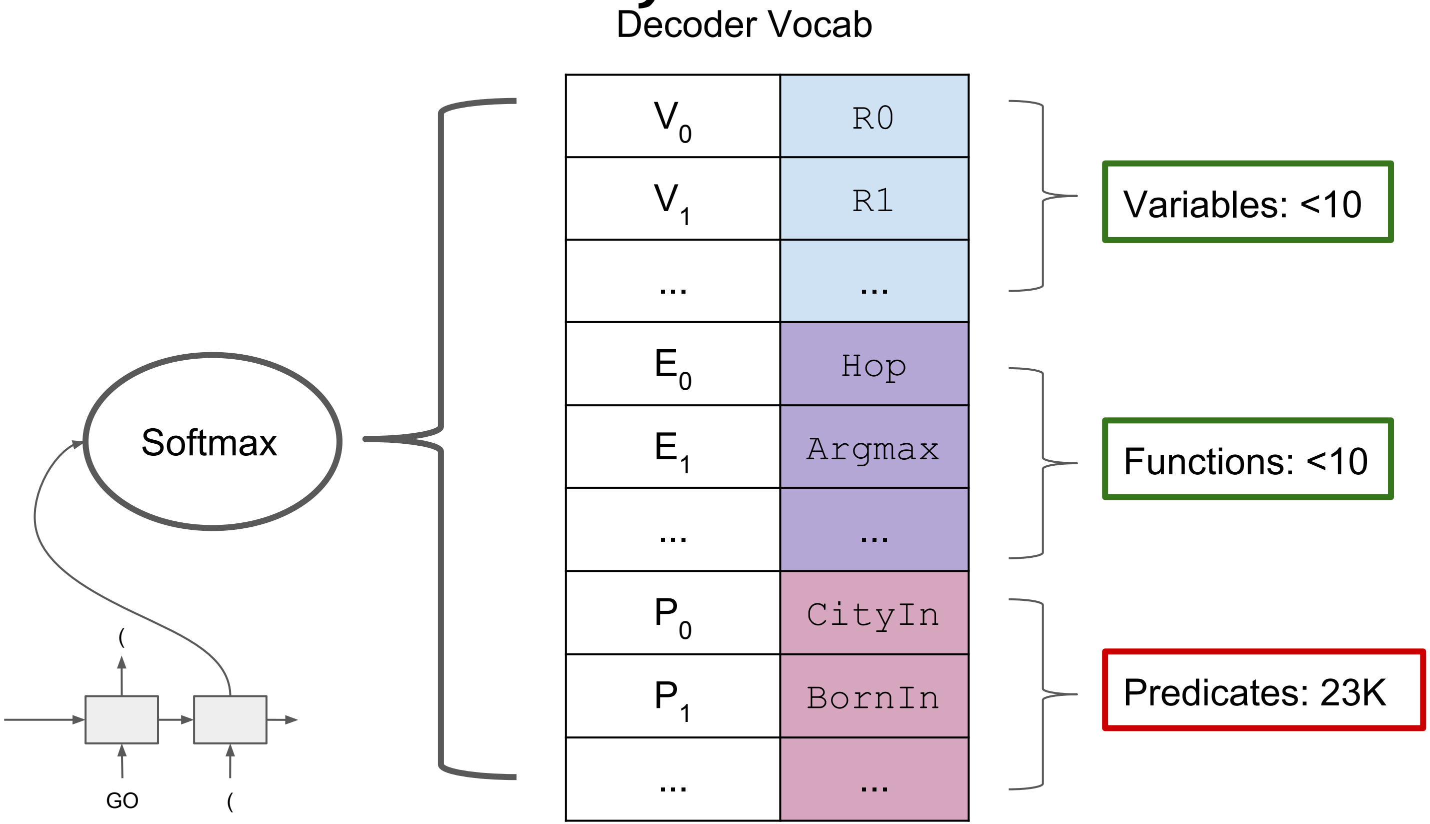
Pen and paper



IDE

Code Assistance: Syntactic Constraint

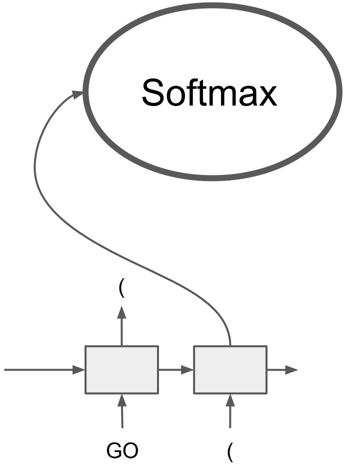
ACL [Liang+ 2017]

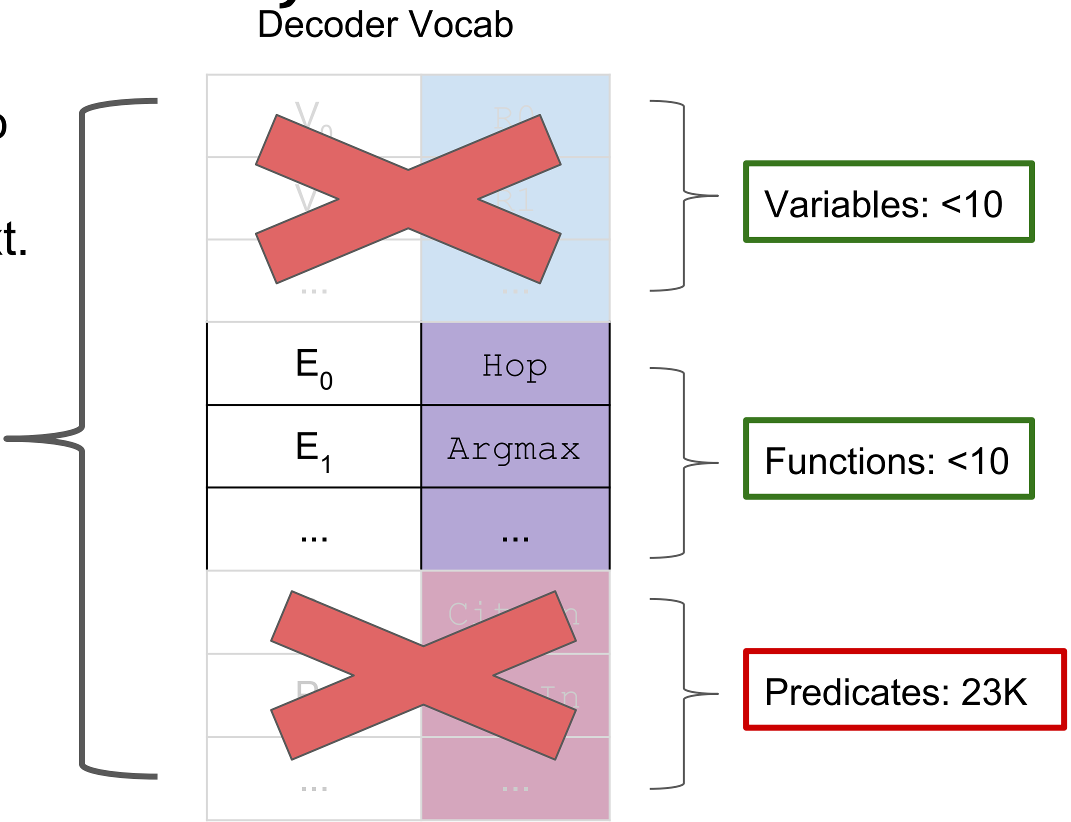


ACL [Liang+ 2017]

Code Assistance: Syntactic Constraint

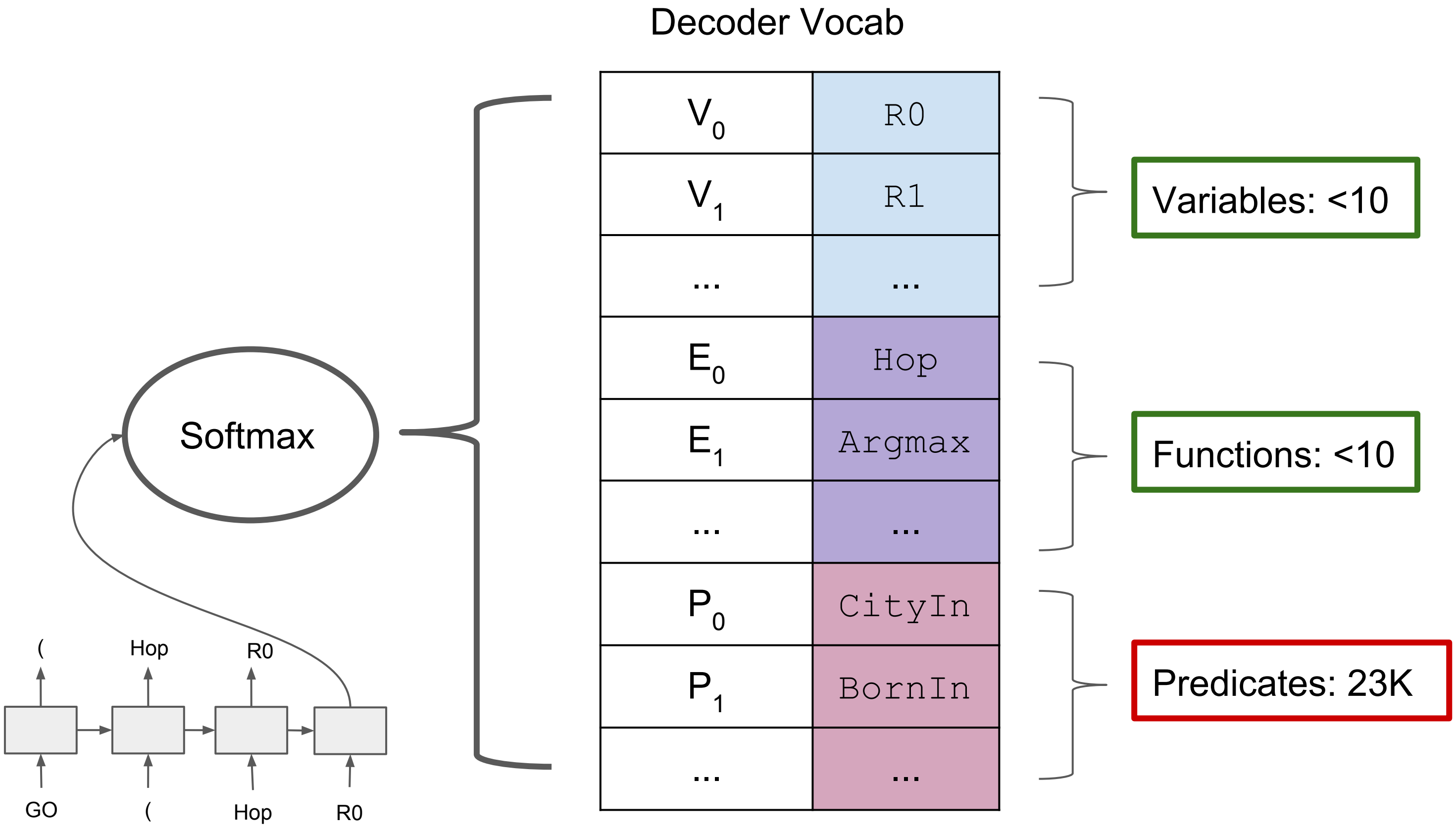
Last token is ‘(’, so has to output a function name next.





Code Assistance: Semantic Constraint

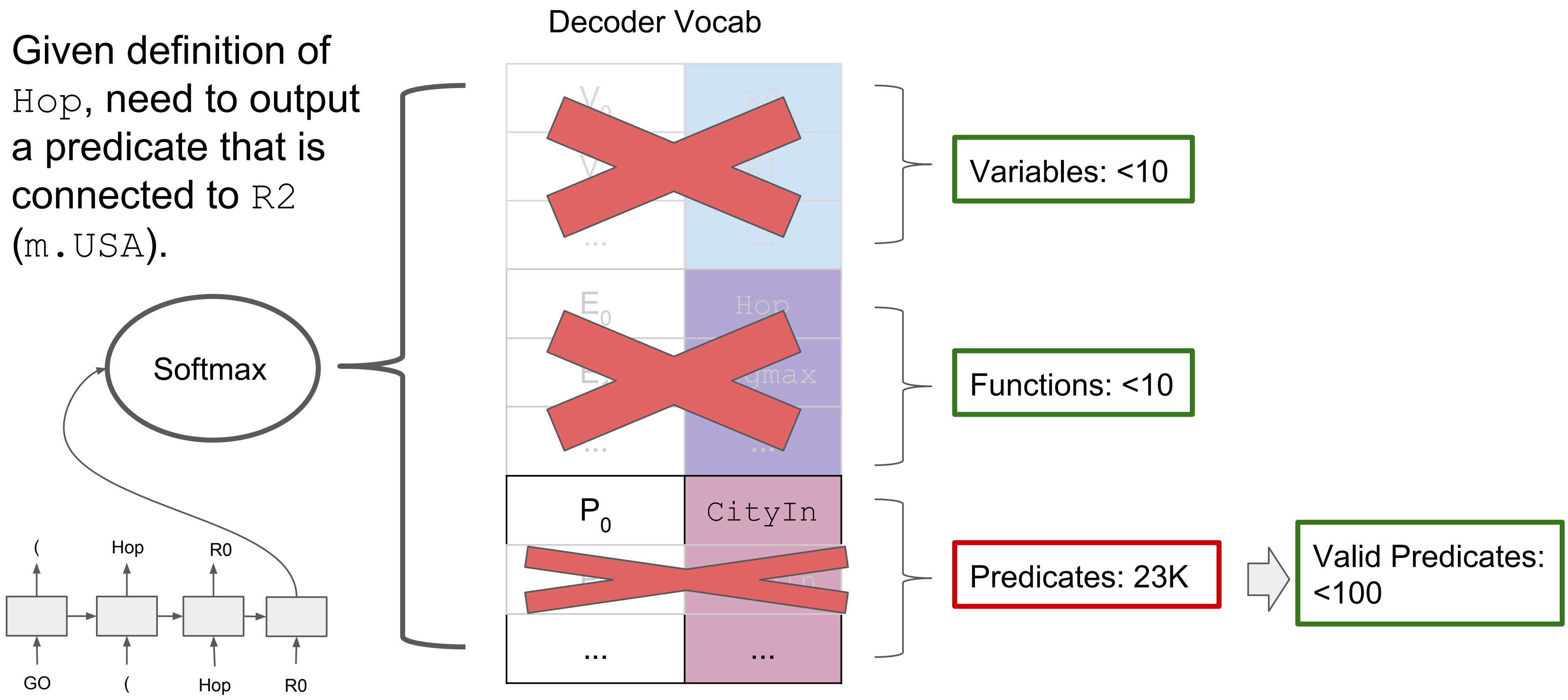
ACL [Liang+ 2017]



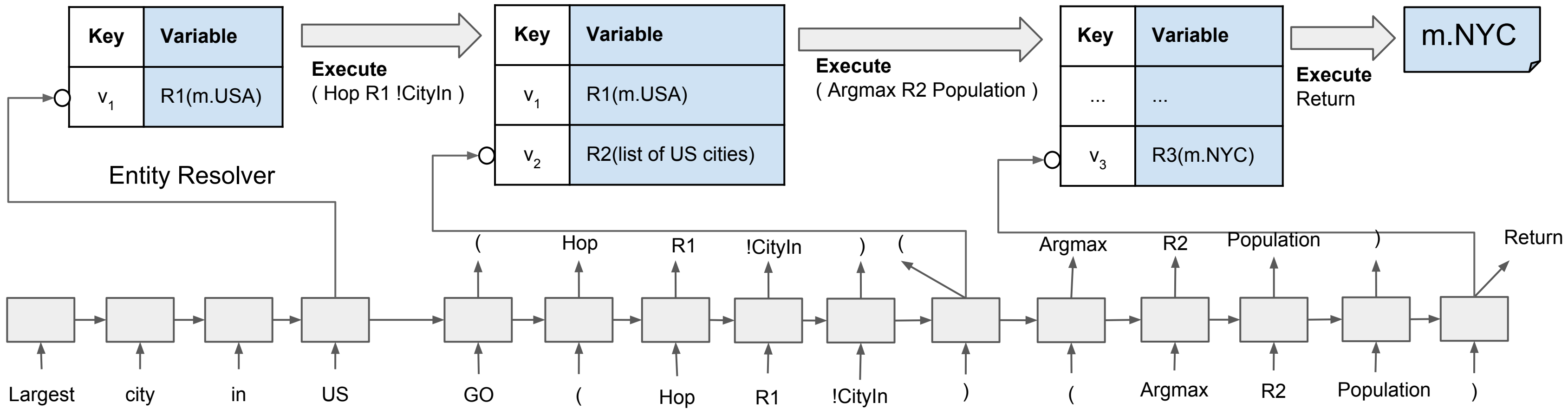
Code Assistance: Semantic Constraint

ACL [Liang+ 2017]

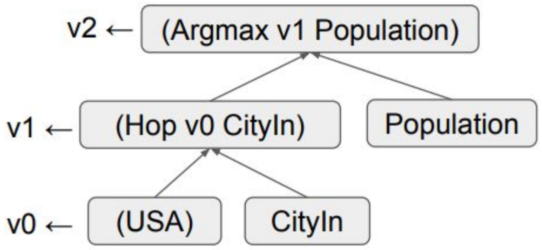
Given definition of `Hop`, need to output a predicate that is connected to `R2` (`m.USA`).



Key-Variable Memory for Semantic Compositionality



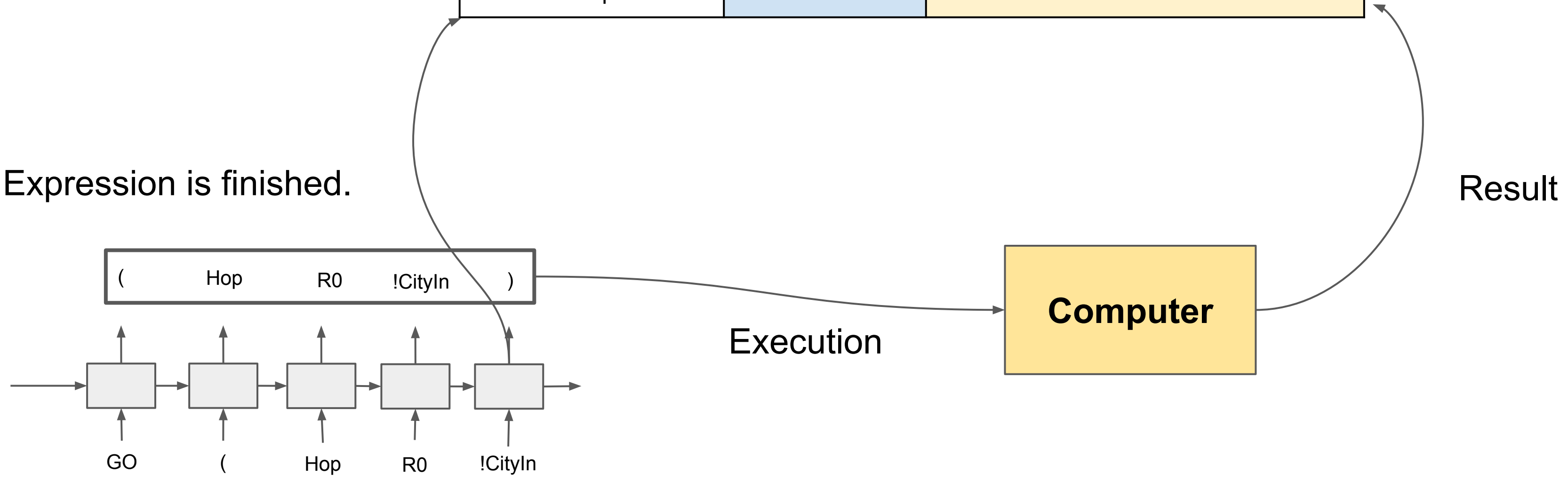
- Equivalent to a linearised bottom-up derivation of the recursive program



Save Intermediate Values

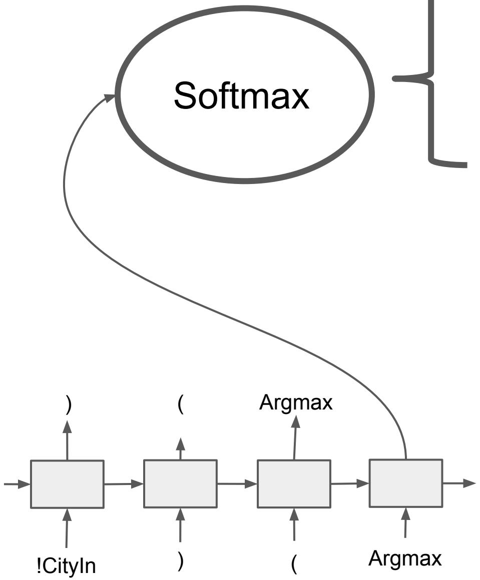
ACL [Liang+ 2017]

Key (Embedding)	**Variable** (Symbol)	**Value** (Data in Computer)
V_0	`R0`	`m.USA`
V_1	`R1`	`[m.SF, m.NYC, ...]`



Reuse Intermediate Values

ACL [Liang+ 2017]



Key (Embedding)	**Variable** (Symbol)	**Value** (Data in Computer)
V_0	`R0`	`m.USA`
V_1	`R1`	`[m.SF, m.NYC, ...]`

Augmented REINFORCE

- REINFORCE get stuck at local maxima
- Iterative ML training is not directly optimizing the F1 score
- Augmented REINFORCE obtains better performances

Settings	Train Avg. F1@1	Valid Avg. F1@1
iterative ML only	68.6	60.1
REINFORCE only	55.1	47.8
Augmented REINFORCE	83.0	**67.2**

ACL [Liang+ 2017]

State-of-the-Art on WebQuestionsSP

- First end-to-end neural network to achieve SOTA on semantic parsing with weak supervision over large knowledge base
- The performance is approaching SOTA with full supervision

Model	Avg. Prec.@1	Avg. Rec.@1	Avg. F1@1
STAGG	67.3	73.1	66.8
NSM – our model	70.8	76.0	**69.0**
STAGG (full supervision)	70.9	80.3	71.7

Plan

Access slides and join discussions at
weakly-supervised-nlu google group

Mobile

Desktop

- ***Weak Supervision NLP***
  - NLP, AI, software 2.0
  - Semantics as a foreign language
  - Unsupervised learning
  - Knowledge representation (symbolism)
- ***Semantic Parsing Tasks***
  - *WebQuestionsSP, WikiTableQuestions*
- ***Neural Symbolic Machines*** (ACL 2017)
  - Compositionality (short term memory)
  - Scalable KB inference (symbolism)
  - RL vs MLE
- ***Memory Augmented Policy Optimization*** (NIPS 2018)
  - Experience replay (long term memory & optimal updating strategy)
  - Systematic exploration
  - Memory Weight Clipping (unbiased cold start strategy)

What is RL?

Directly Optimizing The Expected Reward

- MLE optimizes the log likelihood of target sequences

$$J^{ML}(\theta) = \sum_q \log P(a_{0:T}^{best}(q) | q, \theta)$$

- RL optimizes the expected reward under a stochastic policy

$$J^{RL}(\theta) = \sum_q \mathbb{E}_{P(a_{0:T} | q, \theta)} [R(q, a_{0:T})]$$

[Williams 1992]
[Sutton & Barto 1998]

Book [Sutton & Barto 1998]
NIPS [Abbeel & Schulman 2016]

Challenges of applying RL

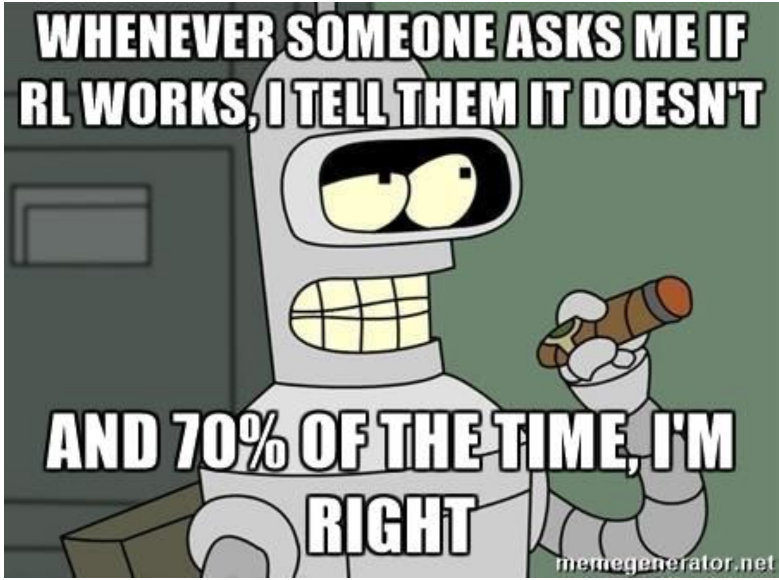
- Large search space (sparse rewards)
  - Supervised pretraining (MLE)
  - Systematic exploration [Houthooft+ 2017]
  - Curiosity
- Credit assignment (delayed reward)
  - Bootstrapping (E.g., train a value function to estimate the future reward)
  - Rollout n-steps
- Train speed & stability
  - Trust region approaches (e.g., PPO)
  - **Experience replay** ⇦ Our focus today

[Irpan 2018]

Efficiency challenge

- RL is still far from data efficient
    - E.g. the best learning algorithm (DeepMind RainbowDQN) “passes median human performance on 57 Atari games at about 18 million frames (around 90 hours) of gameplay, while most humans can pick up a game within a few minutes.”

- How to improve its efficiency?



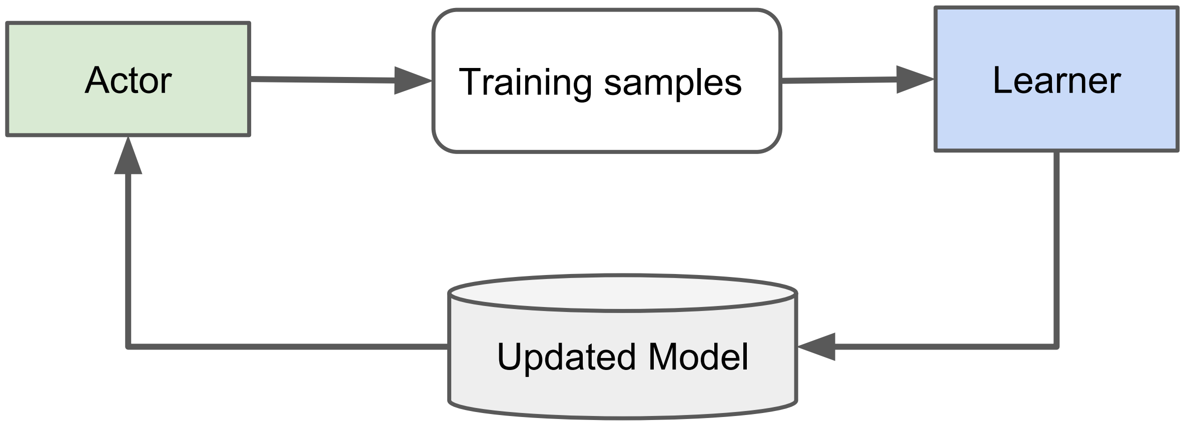
Credit to Alex Irpan from Google Brain Robotics

Applying RL to NLP

- Benefits of RL
  - Weak supervision (e.g., expected answer, user clicks)
  - Directly optimizing the metric (e.g., F1, accuracy, BLEU etc.)
  - Work with structured hidden variables (e.g., logical forms/programs)

- Challenges with existing solutions
  - Large search space sparse reward often leads to slow and unstable training
  - Spurious reward often lead to biased solutions

RL models generate their own training data



- Training sample management issue
  - Too many low quality examples ⇒ slow training
  - Boosting high reward experience ⇒ biased training

Complementary Learning Theory

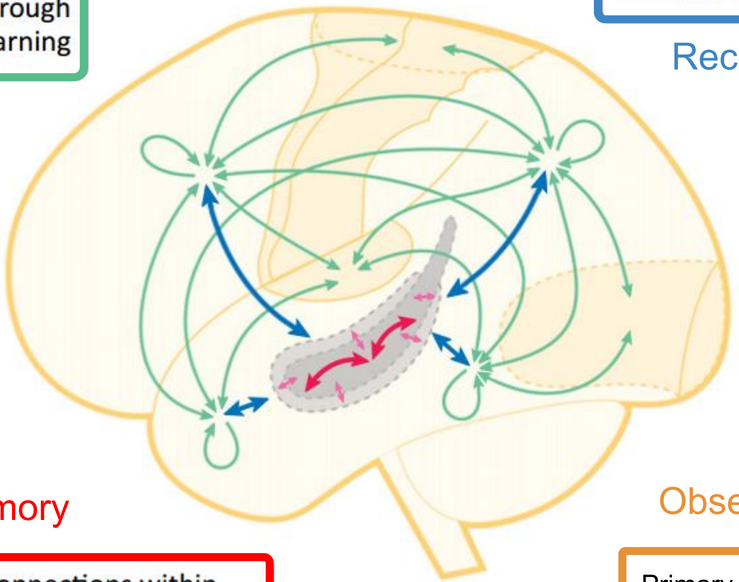
[McClelland+ 1995]
[Kumaran+ 2016]

Connections within and among neocortical areas (green) support gradual acquisition of structured knowledge through interleaved learning

Encoder

Bidirectional connections (blue) link neocortical representations to the hippocampus/MTL for storage, retrieval, and replay

Record & Replay



Episodic memory

Rapid learning in connections within hippocampus (red) supports initial learning of arbitrary new information

Observations

Primary sensory and motor cortices

Most of the past experience are not helpful for improving the current model

http://insideout.wikia.com/wiki/Long_Term_Memory

Mammals learn from “interesting” dreams

- In early 2000s, scientists discovered that animals have complex dreams and are able to retain and recall long sequences of events while they are asleep.

- Recent studies indicate that by consolidating memory traces with **high emotional / motivational value**, "sleep and dreaming may offer a neurobehavioral substrate for the offline … learning”

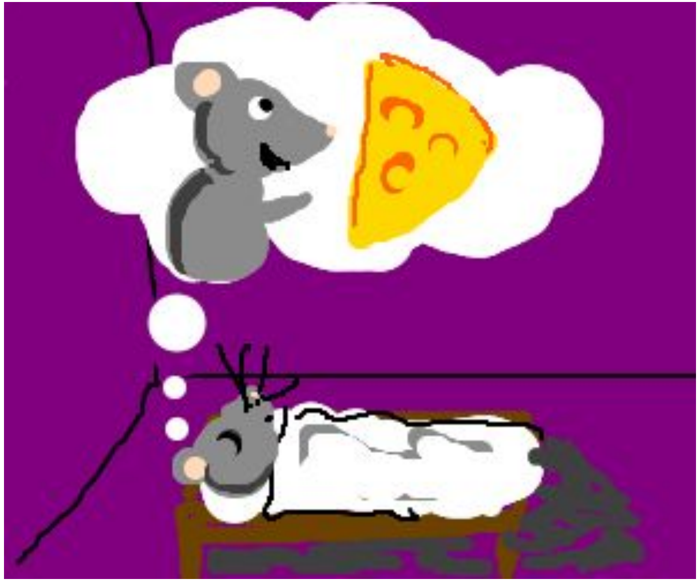


Image courtesy of Kote on Drawception.com, 2012

[Liang+ 2016]
[Abolafia+ 2017]

Augment REINFORCE with Memory

Linear combination of maximum likelihood objective and expected return:

λ log-likelihood on a top-k buffer + (1 - λ) expected return

$$\lambda \sum_{y \in \mathrm{TopK}} \log p(y \mid x) \;+\; (1-\lambda)\mathbb{E}_{\tilde{y} \sim p(y|x)} R(\tilde{y})$$

- Not robust against spurious programs
- The composite objective is ad-hoc, and the gradient is biased

Spurious programs: right answer, wrong reason

Rank	Nation	Gold	Silver	Bronze	Total
1	Nigeria	14	12	9	35
2	Algeria	9	4	4	17
3	Kenya	8	11	4	23
4	Ethiopia	2	4	7	13
5	Ghana	2	2	2	6
6	Ivory Coast	2	1	3	6
7	Egypt	2	1	0	3
8	Senegal	1	1	5	7
9	Morocco	1	1	1	3
10	Tunisia	0	3	1	4
11	Madagascar	0	1	1	2
12	Rwanda	0	0	1	1
12	Zimbabwe	0	0	1	1
12	Seychelles	0	0	1	1

Which nation won the most silver medal?

- **Correct program**:
  (argmax rows “Silver”)
  (hop v1 “Nation”)

- **Many spurious programs**:
  (argmax rows “Gold”)
  (hop v1 “Nation”)

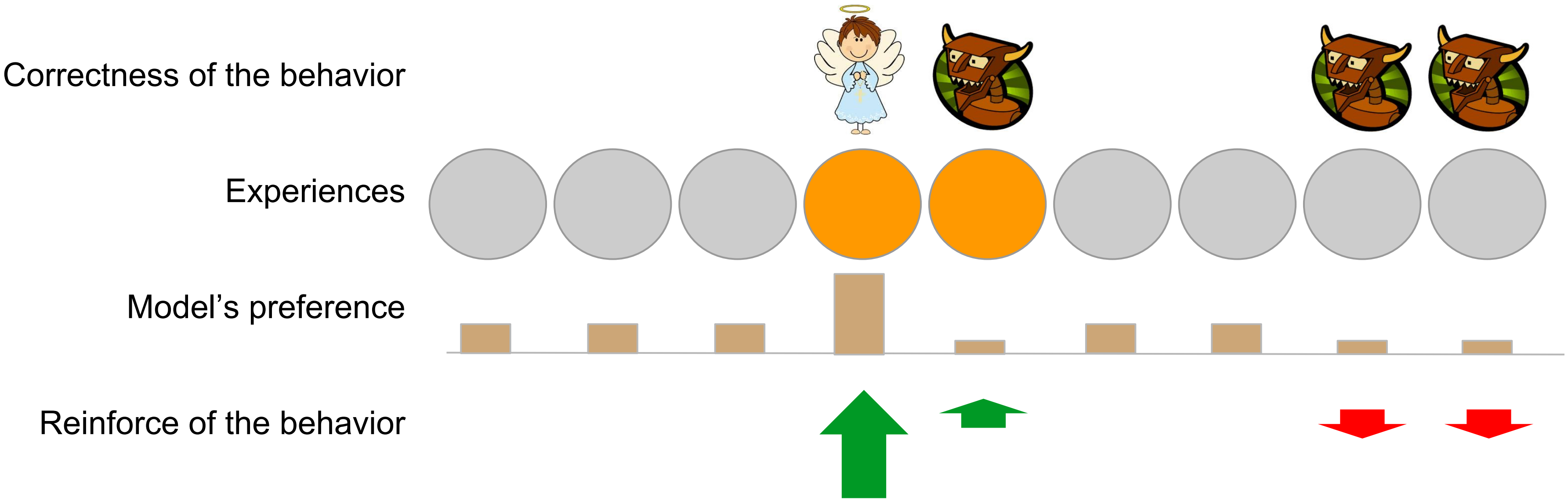
  (argmax rows “Bronze”)
  (hop v1 “Nation”)

  (argmin rows “Rank”)
  (hop v1 “Nation”)

  ...

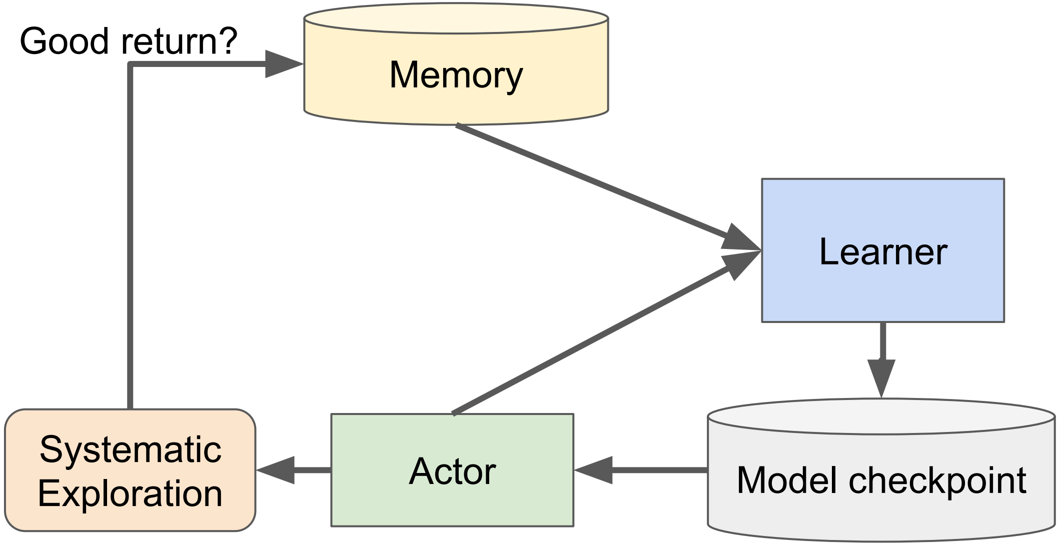
Combating spurious rewards

- **Reinforce** a **rewarded experience** only if the model (current policy) also thinks that it is the right thing to do



NIPS [Liang+ 2018]

Memory Augmented Policy Optimization (MAPO)



NIPS [Liang+ 2018]

Lower the gradient variance without introducing bias

Given a memory buffer of (sequence, return) pairs: $\mathcal{B} \equiv \left\{ \left(y^{(i)}, r^{(i)} \right) \right\}_{i=1}^{n}$,

re-express expected return as,

$$p(\mathcal{B}) \underbrace{\mathbb{E}_{p(\tilde{y}) | \tilde{y} \in \mathcal{B}} R(\tilde{y})}_{\text{inside the buffer}} + (1 - p(\mathcal{B})) \underbrace{\mathbb{E}_{p(\tilde{y}) | \tilde{y} \notin \mathcal{B}} R(\tilde{y})}_{\text{outside the buffer}}$$

- **Importance sampling**
  - Sample more frequently inside the buffer
  - Rejection sampling for samples outside the buffer.

Details

NIPS [Liang+ 2018]

Optimal Sample Allocation

Given that we want to apply stratified sampling to estimate the gradient of REINFORCE with baseline under 1/0 rewards. It can be shown that the optimal strategy is to allocate the **same number of samples** to reward vs no reward experiences

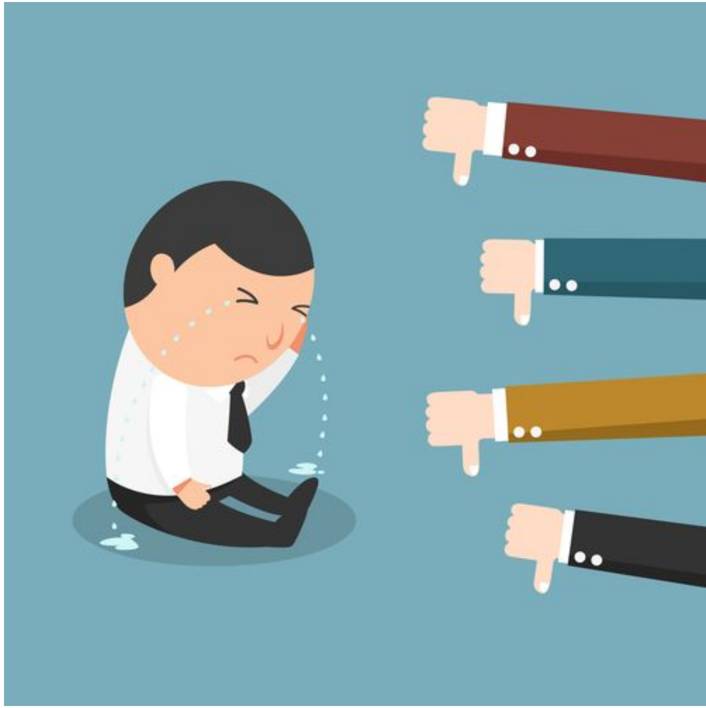
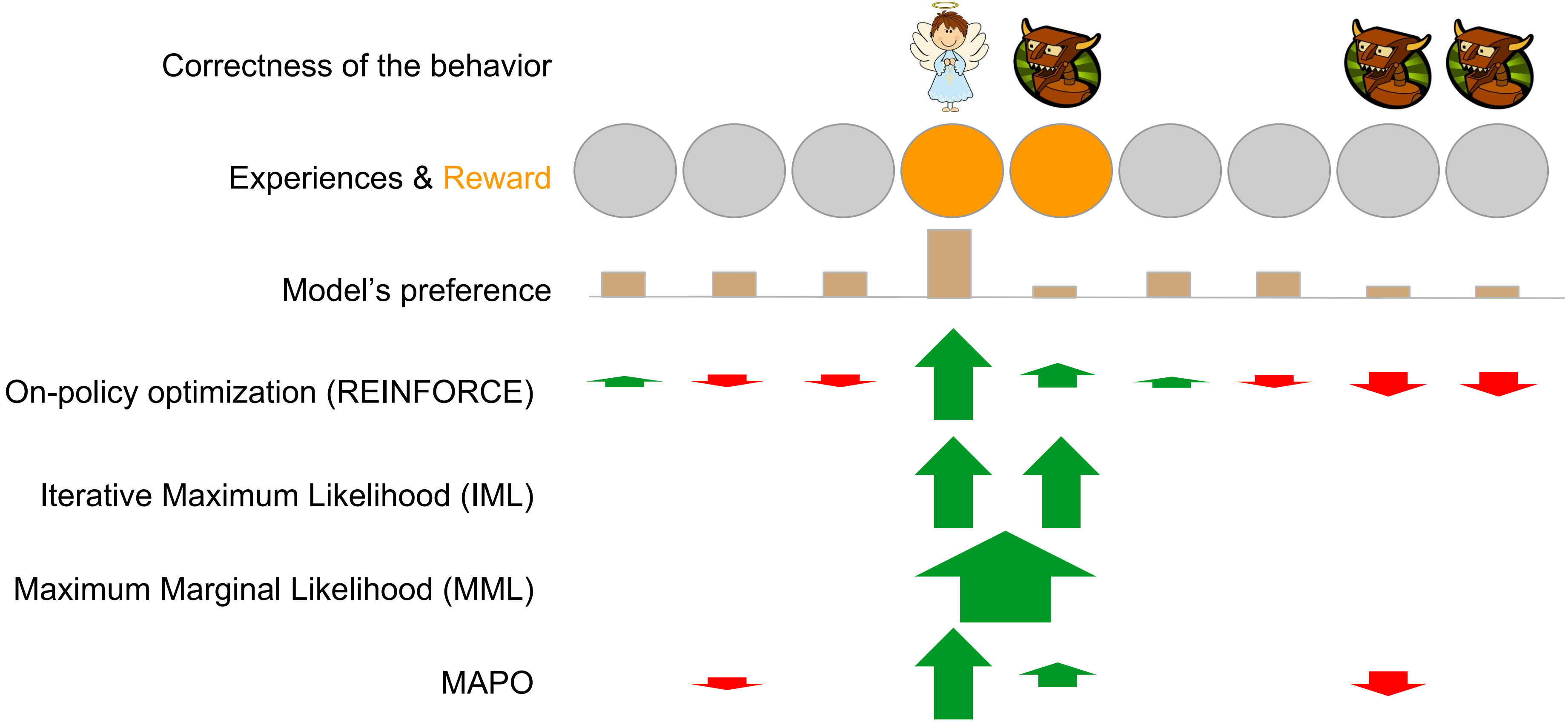


Image source:
Guy Harris, 2018
How to Give Feedback in a Non-Threatening Way

Comparison of model update strategies

NIPS [Liang+ 2018]



NIPS [Liang+ 2018]

Memory Weight Clipping

- Policy gradient methods usually suffer from a cold start problem, because P is very small initially

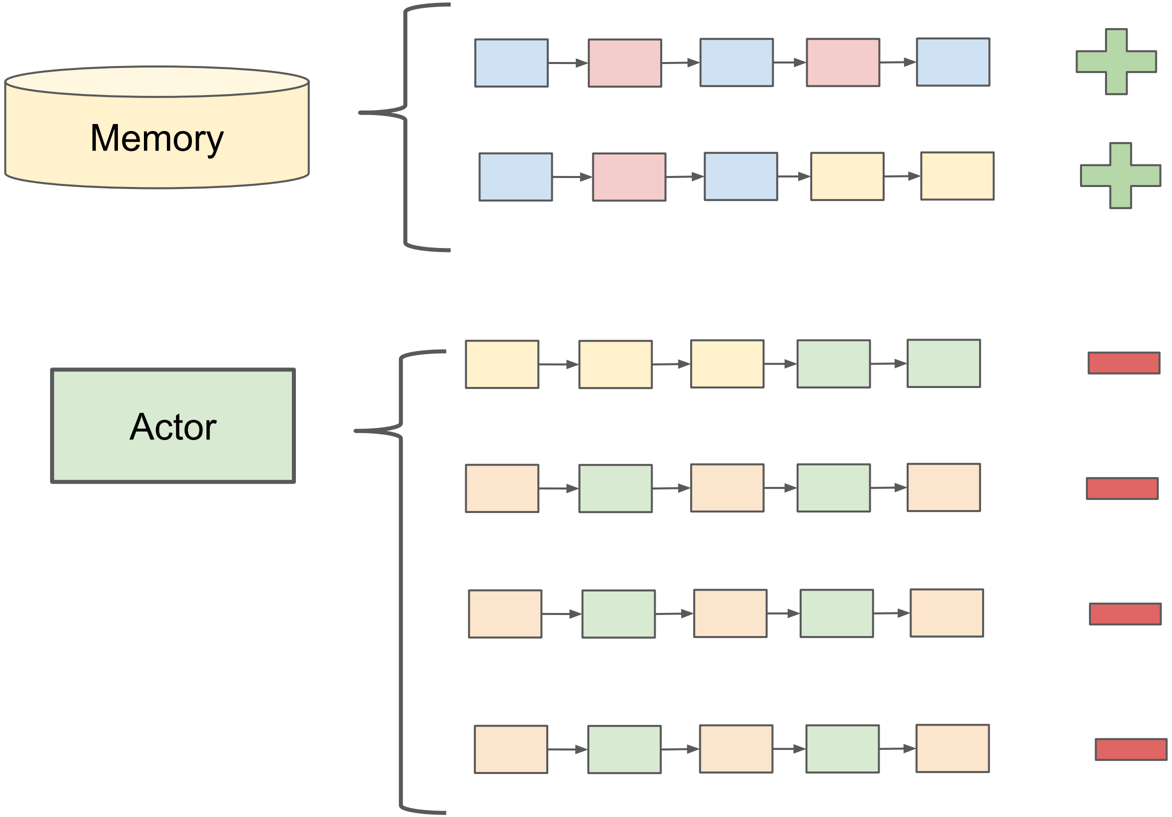
$$\nabla_\theta J^{RL}(\theta) = \sum_q \sum_{a_{0:T}} \boxed{P(a_{0:T}|q,\theta)} [R(q, a_{0:T}) - B(q)] \nabla_\theta \log P(a_{0:T}|q,\theta)$$

- We adopt a clipping mechanism to ensure that the buffer probability is larger than a threshold a

$$\begin{aligned} &\text{Sample } \mathbf{a}_i^+ \sim \pi_\theta^{old} \text{ over } \mathcal{B}_i \\ &w_i^+ \leftarrow \max(\pi_\theta^{old}(\mathcal{B}_i), \alpha) \\ &D \leftarrow D \cup (\mathbf{a}_i^+, R(\mathbf{a}_i^+), w_i^+) \end{aligned}$$

Memory Weight Clipping

NIPS [Liang+ 2018]



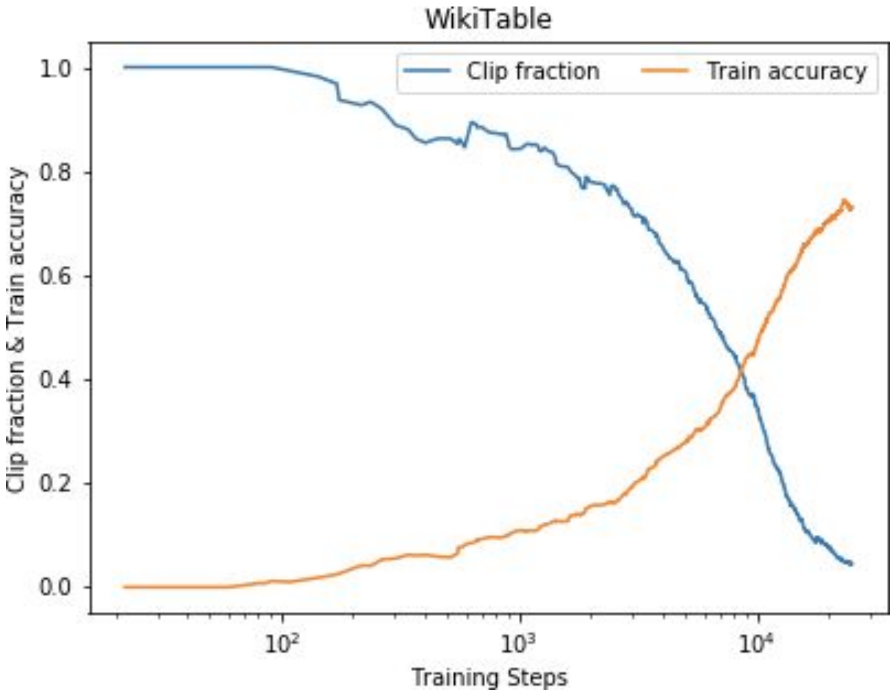
$\max(P(B), \alpha)$

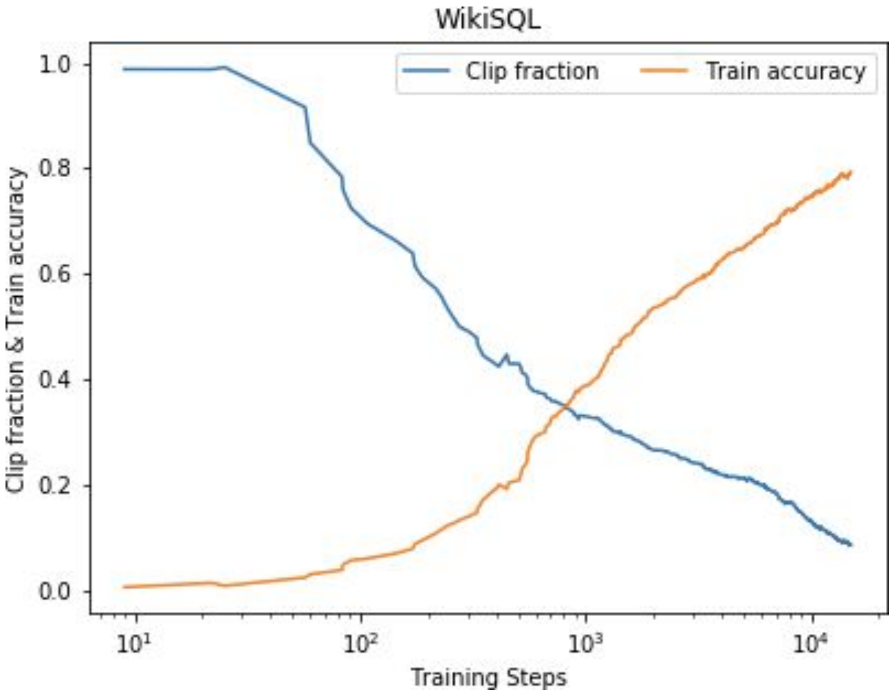
Trade off bias in the initial stage for faster training

$1 - \max(P(B), \alpha)$

NIPS [Liang+ 2018]

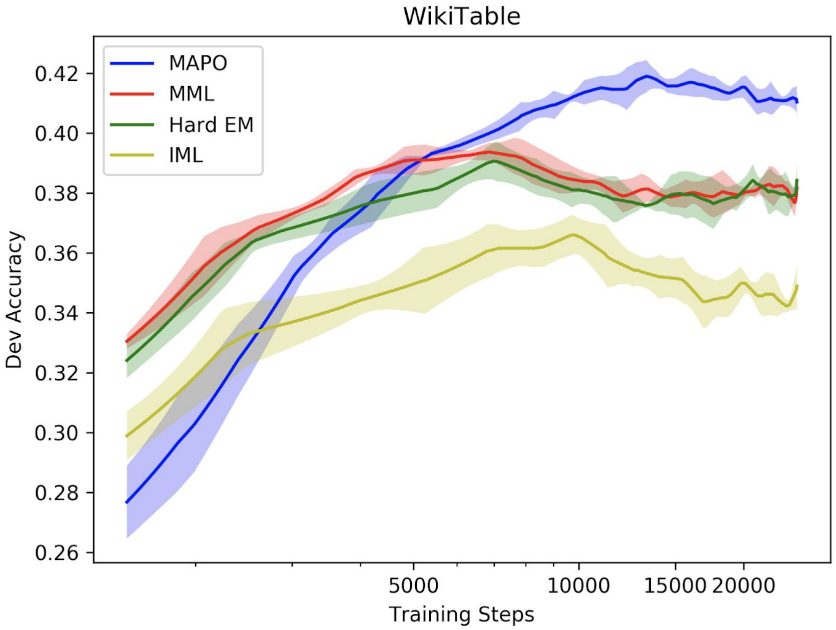
Training becomes less biased over time

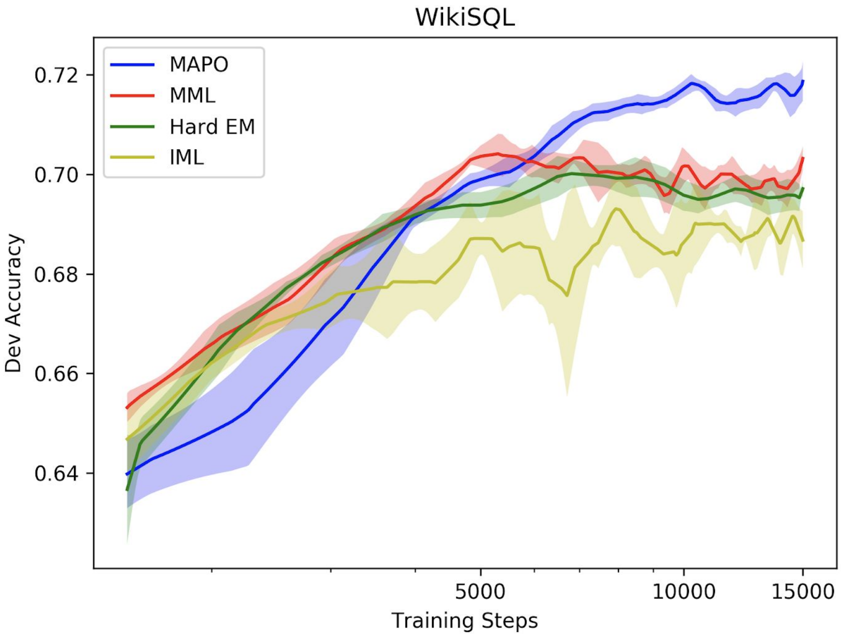




Comparison

- REINFORCE does not work at all
- MAPO is slower but less biased





- The shaded area represents the standard deviation of the dev accuracy

NIPS [Liang+ 2018]

SOTA results with weak supervision

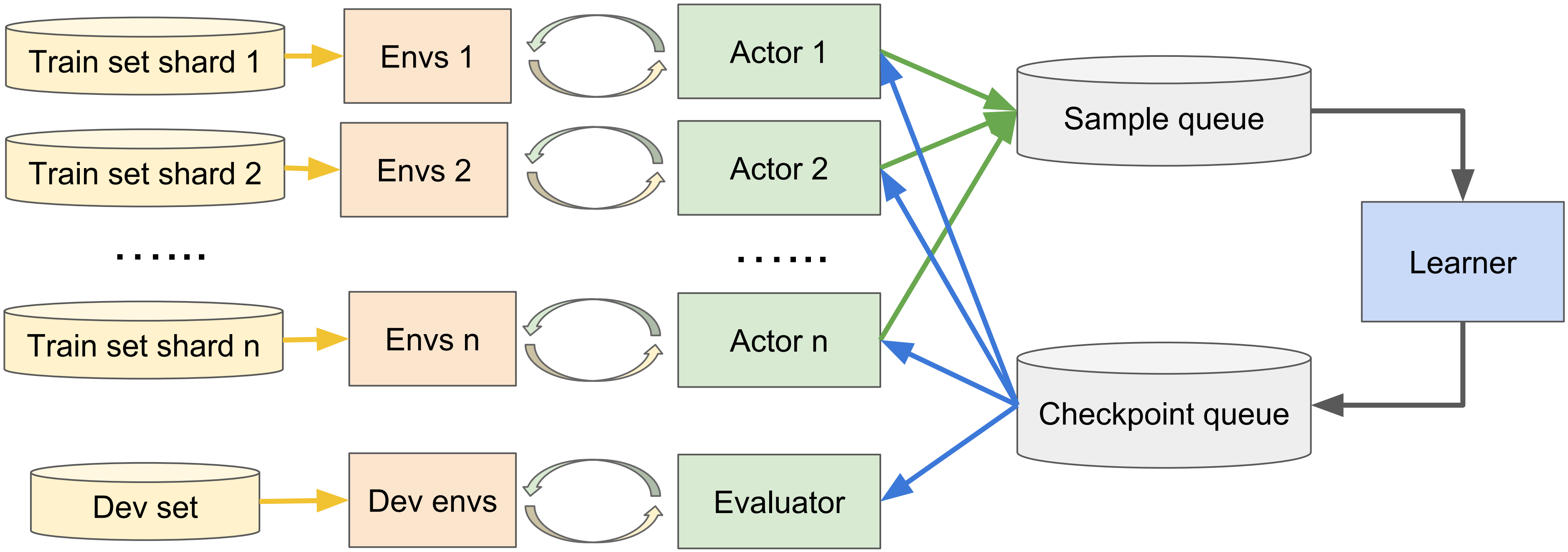
Model	E.S.	Dev.	Test
Pasupat & Liang (2015)[28]	-	37.0	37.1
Neelakantan et al. (2017)[26]	1	34.1	34.2
Neelakantan et al. (2017)[26]	15	37.5	37.7
Haug et al. (2017)[15]	1	-	34.8
Haug et al. (2017)[15]	15	-	38.7
Zhang et al. (2017)[51]	-	40.4	43.7
MAPO	1	42.7	43.8
MAPO (ensembled)	5	-	46.2

Table 3: Results on WIKITABLEQUESTIONS. E.S. is the number of ensembles (if applicable).

Model	Dev.	Test
Zhong et al. (2017)[52]*	60.8	59.4
Wang et al. (2017)[40]*	67.1	66.8
Xu et al. (2017)[46]*	69.8	68.0
Huang et al. (2018)[18]*	68.3	68.0
Yu et al. (2018)[48]*	74.5	73.5
Sun et al. (2018)[38]*	75.1	74.6
Dong & Lapata (2018)[12]*	79.0	78.5
MAPO	72.4	72.6
MAPO (ensemble of 5)	-	74.9

Table 4: Results on WIKISQL. *All other methods use question-program pairs as strong supervision, while MAPO only uses question-answer pairs as weak supervision.

Scale up: Distributed Actor-Learner architecture



[Liang et al, 2017; Espeholt et al, 2018; Liang et al, 2018]

Thanks!

Access slides and join discussions at
weakly-supervised-nlu google group

- ***Weak Supervision NLP***
  - NLP, AI, software 2.0
  - Semantics as a foreign language
  - Unsupervised learning
  - Knowledge representation (symbolism)

- ***Semantic Parsing Tasks***
  - *WebQuestionsSP, WikiTableQuestions*

- ***Neural Symbolic Machines*** (ACL 2017)
  - Compositionality (short term memory)
  - Scalable KB inference (symbolism)
  - RL vs MLE

- ***Memory Augmented Policy Optimization*** (NIPS 2018)
  - Experience replay (long term memory & optimal updating strategy)
  - Systematic exploration
  - Memory Weight Clipping (unbiased cold start strategy)

Mobile

Desktop