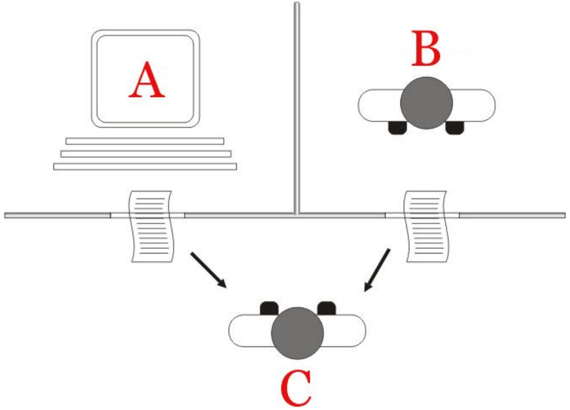
Weakly Supervised Natural Language Understanding

Ni Lao
Chief Scientist @mosaix.ai
5.14.2019

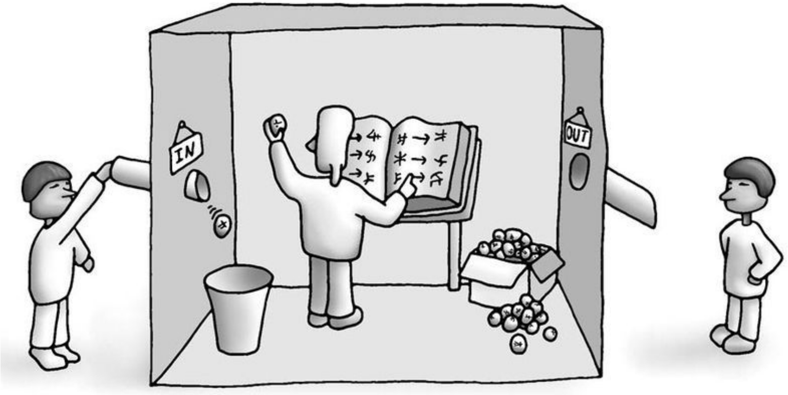
What is understanding?

"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

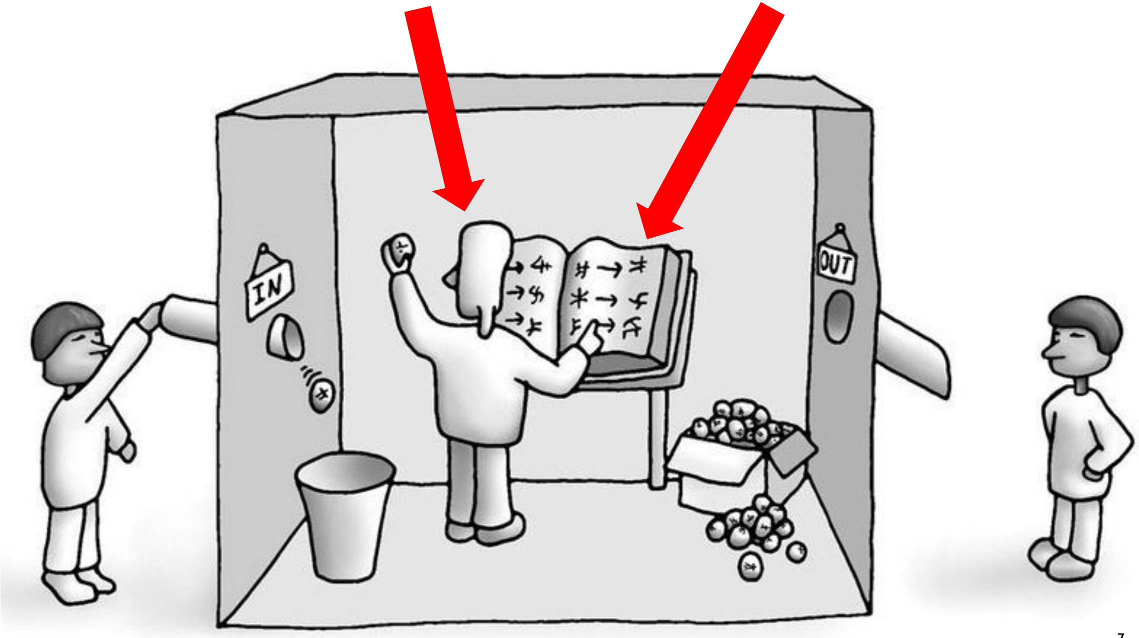
Full Supervision NLP

- Traditionally NLP is a labor-intensive business

Applications	
Discourse Processing	
Semantic Parsing	
Syntactic analysis	
Morphological analysis	
ASR	OCR
speech	text

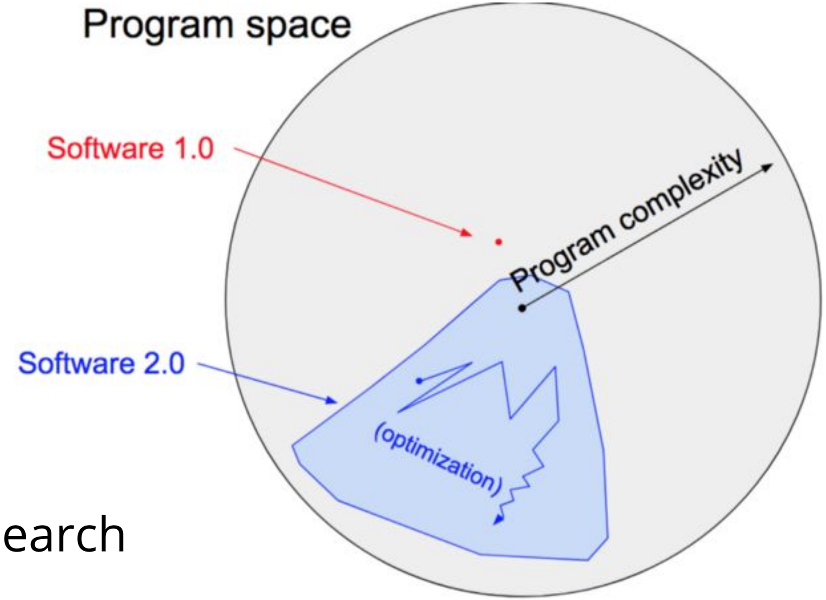
NLP algorithms simply lookup

Rules written by 1000 PhDs



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 disposal to search this space for a program that works.



[Karpathy 2017; Watson 2017; Ratner+ 2018]

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.

Where does knowledge come from?

- Can only cover the most popular semantics used by human

the world

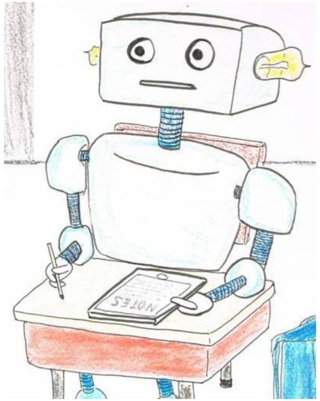
domain experts
(bottlenecks)

expert systems with
knowledge bases

Weak Supervision NLP

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

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



machine learning

intelligent systems
with knowledge

[Berant+ 2013]
[Liang 2013]

Semantic Parsing

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

Natural language

Semantic parsing

Full supervision (hard to collect)

Logical form / Program

Knowledge Base

Execution

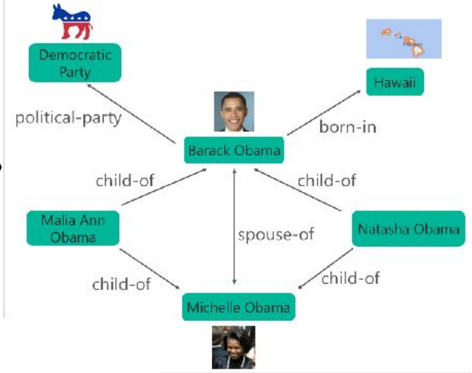
Weak supervision (easy to collect)

Desired behavior / Answer

Question Answering with Knowledge Base

Largest city in US?

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



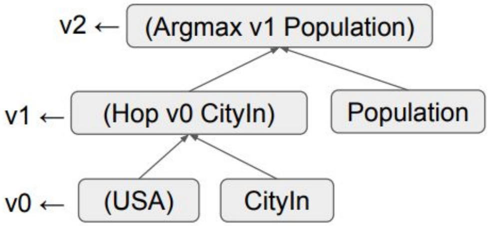
Freebase, DBpedia, YAGO, NELL

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

[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

[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: $\{2004\}$

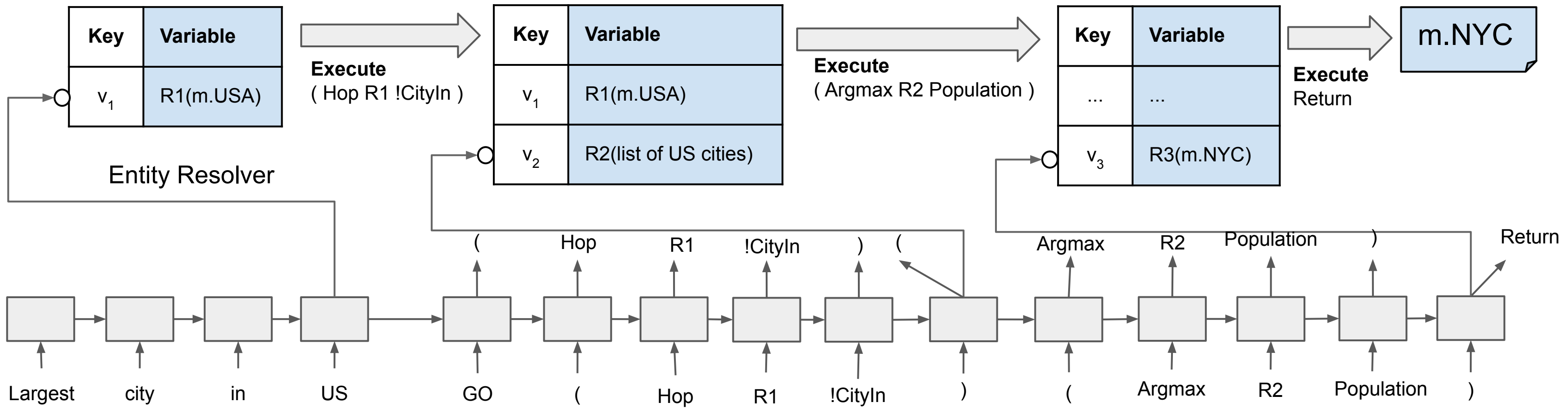
x_2: *"In which city's the first time with at least 20 nations?"*
y_2: $\{\text{Paris}\}$

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

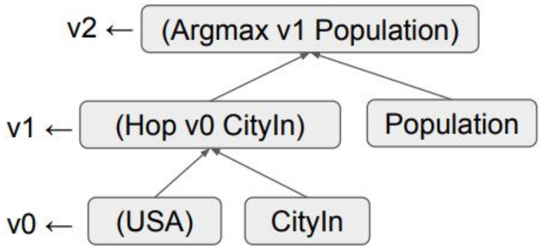
x_4: *"How many events were in Athens, Greece?"*
y_4: $\{2\}$

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

Key-Variable Memory for Semantic Compositionality



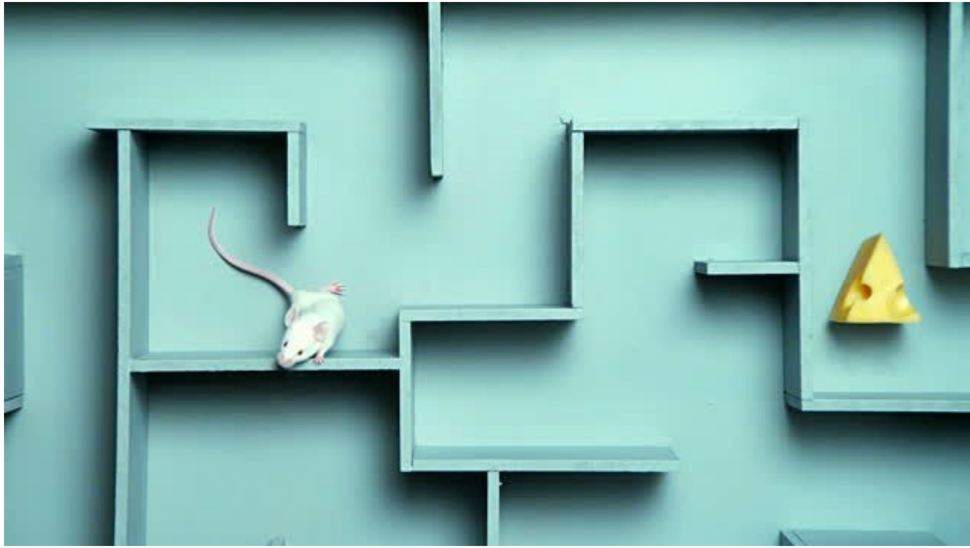
- Equivalent to a linearised bottom-up derivation of the recursive program



Structured Prediction Often Needs to Do Search

Which fits well with RL's ability to form good search policy

- dialogue
- semantic parsing
- program synthesis
- architecture search
- machine translation
- summarization
- image caption
- knowledge graph reasoning
- information extraction
- ...



RL is attractive:

Directly Optimizing The Expected Reward
Which can be very useful for structured predictions

- ML optimizes the log likelihood of target sequences

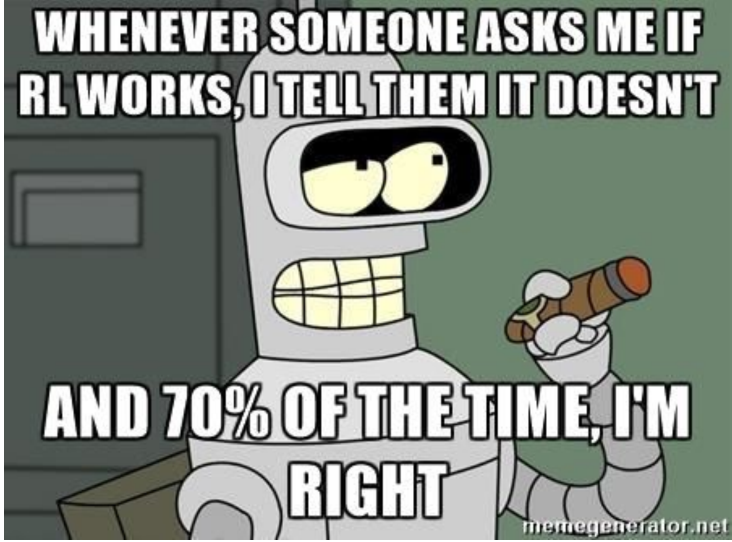
$$J^{ML}(\theta) = \sum_q \log P(a_{0:T}^{best}(q) | q, \theta)$$

- 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]
[Liang+ 2017]

RL has challenges:

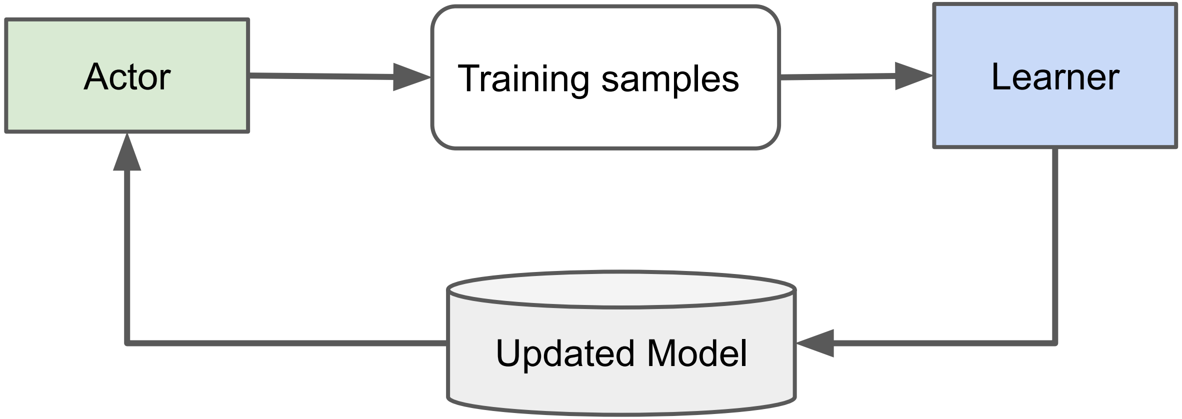


[Sutton & Barto 1998]
[Abbeel & Schulman 2016]

Which we need to be aware of

- Large search space (sparse rewards)
  - Supervised pretraining (MLE)
  - Systematic exploration [Houthooft+ 2017]
  - Curiosity [Schmidhuber 1991][Pathak2017]
- Credit assignment (delayed reward)
  - Bootstrapping
    - E.g., AlphaGo uses a value function to estimate the future reward
  - Rollout n-steps
- Train speed & stability (optimization)
  - Trust region approaches (e.g., PPO)
  - Experience replay

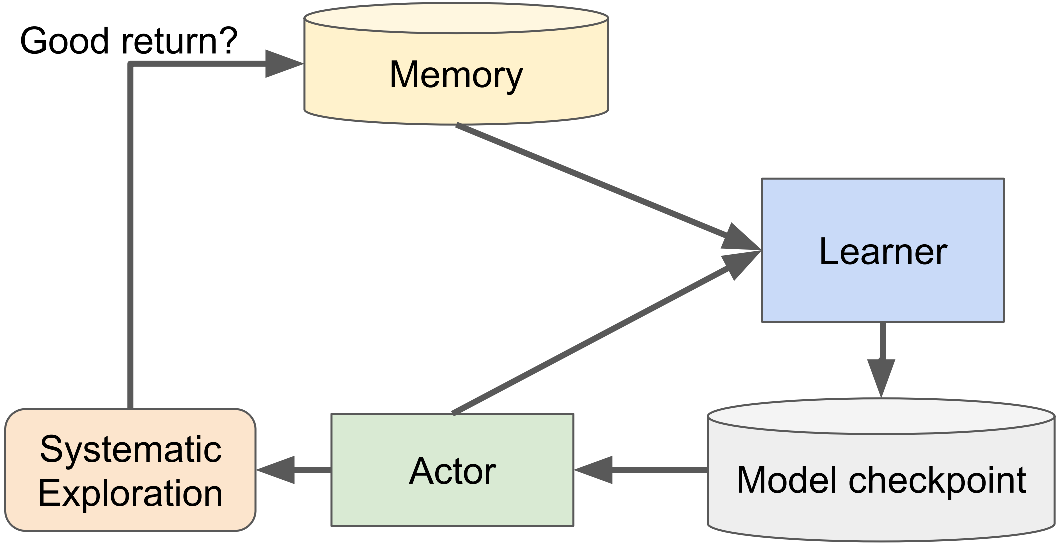
RL models generate their own training data



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

NIPS [Liang+ 2018]

Memory Augmented Policy Optimization (MAPO)

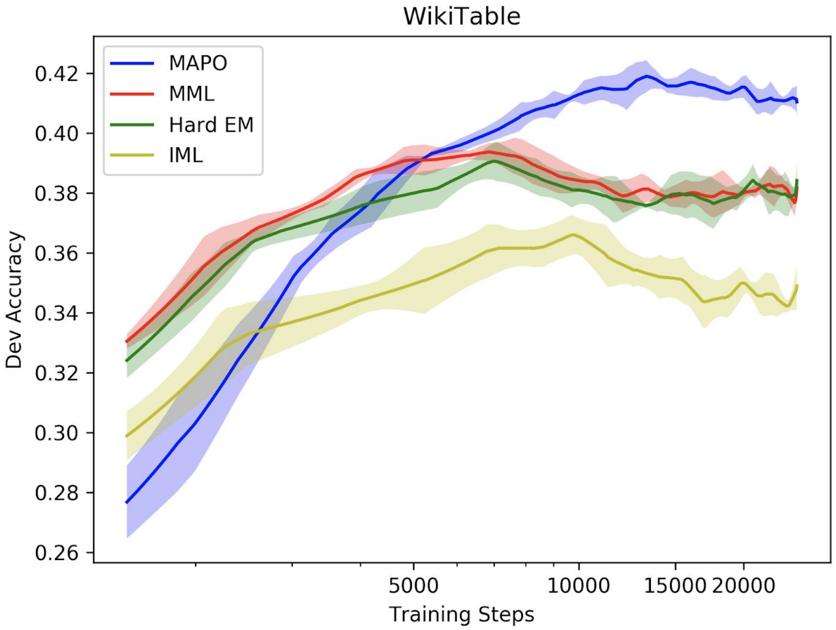


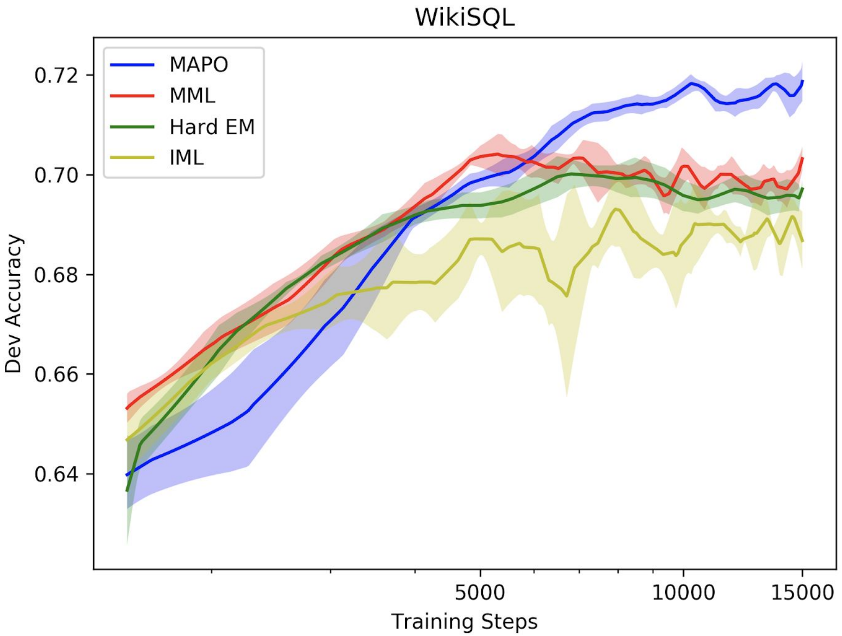
Most of the past experience are not helpful for improving the current model

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

Comparison

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





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