DeepLearn2017
Bilbao

An Introduction to Deep Learning for Natural Language Processing

Jianfeng Gao

Thanks for the slides by **Bill Dolan, Michel Galley**, **Xiaodong He, Lihong Li, Rangan Majumder, Scott Yih et al.**

Joint work with many Microsoft colleagues and interns (see the list of collaborators)

Microsoft AI & Research

International Summer School on Deep Learning 2017

July 20-21, 2017, Bilbao, Spain

Contact Information:

www.microsoft.com/en-us/research/people/jfgao/

Collaborators:

Faisal Ahmed, Chris Brockett, Asli Celikyilmaz, Ming-Wei Chang, Weizhu Chen, Yun-Nung Chen, Li Deng, Bhuwan Dhingra, Bill Dolan, Michel Galley, Marjan Ghazvininejad, Xiaodong He, Po-Sen Huang, Sungjin Lee, Jiwei Li, Lihong Li, Xiujun Li, Zachary Lipton, Xiaodong Liu, Rangan Majumder, Nasrin Mostafazadeh, Baolin Peng, Mir Rosenberg, Yelong Shen, Alessandro Sordoni, Saurabh Tiwary, Lucy Vanderwende, Luan Yi, Scott Yih et al.

Tutorial Outline

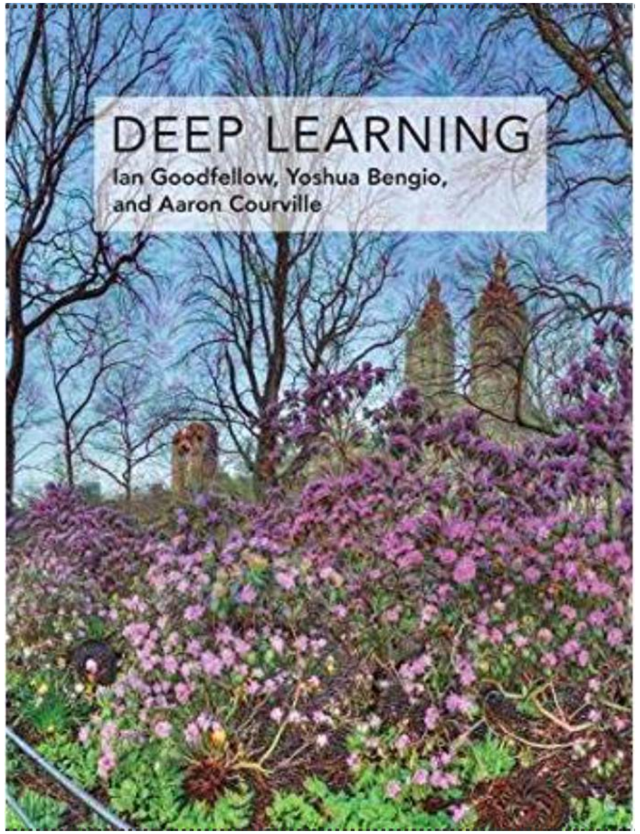
- Part 1: Background
- Part 2: Deep semantic similarity models for text processing
- Part 3: Recurrent neural networks for text generation
- Part 4: Neural machine reading models for question answering
- Part 5: Deep reinforcement learning for task-completion dialogue

Tutorial Outline

- **Part 1: Background**
  - **A brief history of deep learning**
  - **Transition of NLP to neural methods**
  - **An example of neural models for query classification**
- Part 2: Deep semantic similarity models for text processing
- Part 3: Recurrent neural networks for text generation
- Part 4: Neural machine reading models for question answering
- Part 5: Deep reinforcement learning for task-completion dialogue

Historical trends in deep learning (DL)

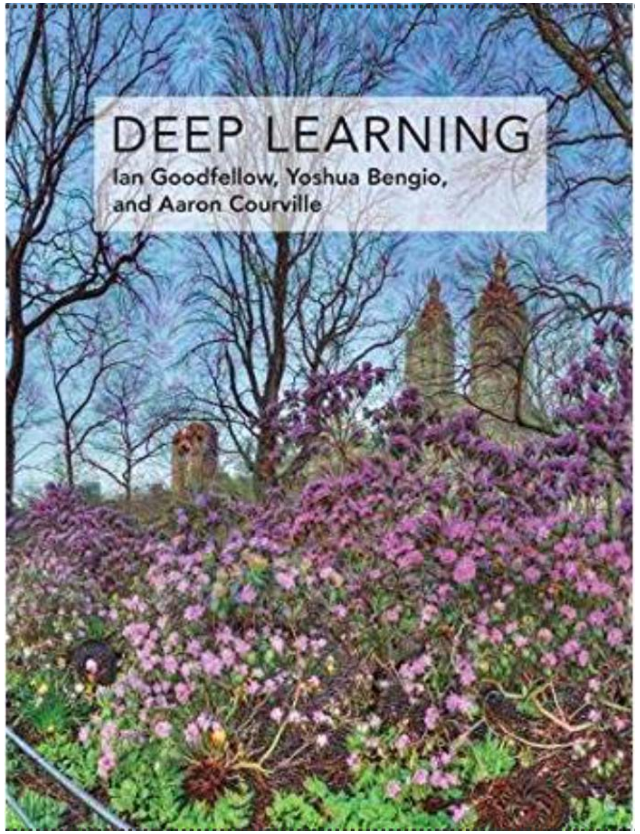
- DL has had a long and rich history, with different names (cybernetics, connectionism, neural nets)
- Become more useful as the amount of available training data has increased
- DL models have grown in size over time as computer infrastructure (both hardware and software) for DL has improved
- DL has solved increasingly complicated tasks with increasing accuracy over time.



[Goodfellow+ 16]

Three waves of DL

- Wave 1: cybernetics
  - Started in 40s to 60s [McCulloch & Pitts 43; Hebb 49]
  - Development of perceptron [Rosenblatt 58]
- Wave 2: connectionism or neural networks
  - Started in 80s to 90s
  - Development of back-propagation [Rumelhart+ 86]
- Wave 3 (current wave): deep learning
  - Started at 2006 [Hinton+ 06; Bengio+ 07; Ranzato+ 07]



[Goodfellow+ 16]

On the Origin of Deep Learning

Haohan Wang HAOHANW@CS.CMU.EDU

Bhiksha Raj BHIKSHA@CS.CMU.EDU

Language Technologies Institute
School of Computer Science
Carnegie Mellon University

Abstract

This paper is a review of the evolutionary history of deep learning models. It covers from the genesis of neural networks when associationism modeling of the brain is studied, to the models that dominate the last decade of research in deep learning like convolutional neural networks, deep belief networks, and recurrent neural networks. In addition to a review of these models, this paper primarily focuses on the precedents of the models above, examining how the initial ideas are assembled to construct the early models and how these preliminary models are developed into their current forms. Many of these evolutionary paths last more than half a century and have a diversity of directions. For example, CNN is built on prior knowledge of biological vision system; DBN is evolved from a trade-off of modeling power and computation complexity of graphical models and many nowadays models are neural counterparts of ancient linear models. This paper reviews these evolutionary paths and offers a concise thought flow of how these models are developed, and aims to provide a thorough background for deep learning. More importantly, along with the path, this paper summarizes the gist behind these milestones and proposes many directions to guide the future research of deep learning.

300 BC	Aristotle	introduced Associationism, started the history of human's attempt to understand brain.
1873	Alexander Bain	introduced Neural Groupings as the earliest models of neural network, inspired Hebbian Learning Rule.
1943	McCulloch & Pitts	introduced MCP Model, which is considered as the ancestor of Artificial Neural Model.
1949	Donald Hebb	considered as the father of neural networks, introduced Hebbian Learning Rule, which lays the foundation of modern neural network.
1958	Frank Rosenblatt	introduced the first perceptron, which highly resembles modern perceptron.
1974	Paul Werbos	introduced Backpropagation
1980	Teuvo Kohonen	introduced Self Organizing Map
	Kunihiko Fukushima	introduced Neocogitron, which inspired Convolutional Neural Network
1982	John Hopfield	introduced Hopfield Network
1985	Hilton & Sejnowski	introduced Boltzmann Machine
1986	Paul Smolensky	introduced Harmonium, which is later known as Restricted Boltzmann Machine
	Michael I. Jordan	defined and introduced Recurrent Neural Network
1990	Yann LeCun	introduced LeNet, showed the possibility of deep neural networks in practice
1997	Schuster & Paliwal	introduced Bidirectional Recurrent Neural Network
	Hochreiter & Schmidhuber	introduced LSTM, solved the problem of vanishing gradient in recurrent neural networks
2006	Geoffrey Hinton	introduced Deep Belief Networks, also introduced layer-wise pretraining technique, opened current deep learning era.
2009	Salakhutdinov & Hinton	introduced Deep Boltzmann Machines
2012	Geoffrey Hinton	introduced Dropout, an efficient way of training neural networks

[Wang & Raj 17]

Geoff Hinton

The New York Times

Scientists See Promise in Deep-Learning Programs

John Markoff November 23, 2012

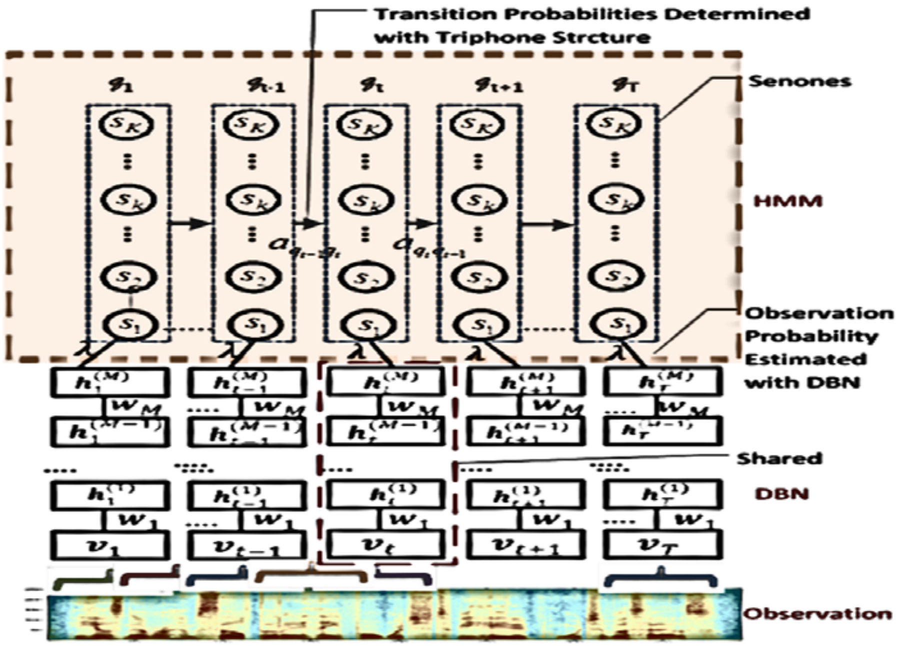
Rick Rashid in **Tianjin, China**, October, 25, 2012

The universal translator on "Star Trek" comes true…

A voice recognition program translated a speech given by Richard F. Rashid, Microsoft's top scientist, into Chinese.

Slide from Li Deng

CD-DNN-HMM

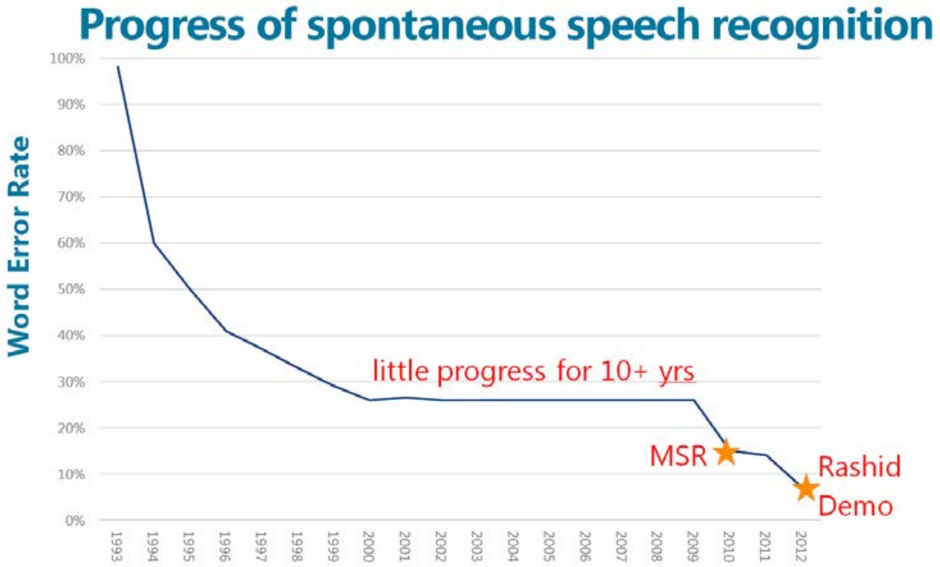


Dahl, Yu, Deng, and Acero, "Context-Dependent Pre-trained Deep Neural Networks for Large Vocabulary Speech Recognition," *IEEE Trans. ASLP*, Jan. 2012

Seide, Li, and Yu, "Conversational Speech Transcription using Context-Dependent Deep Neural Networks," *INTERSPEECH* 2011.

After no improvement for 10+ years by the research community…

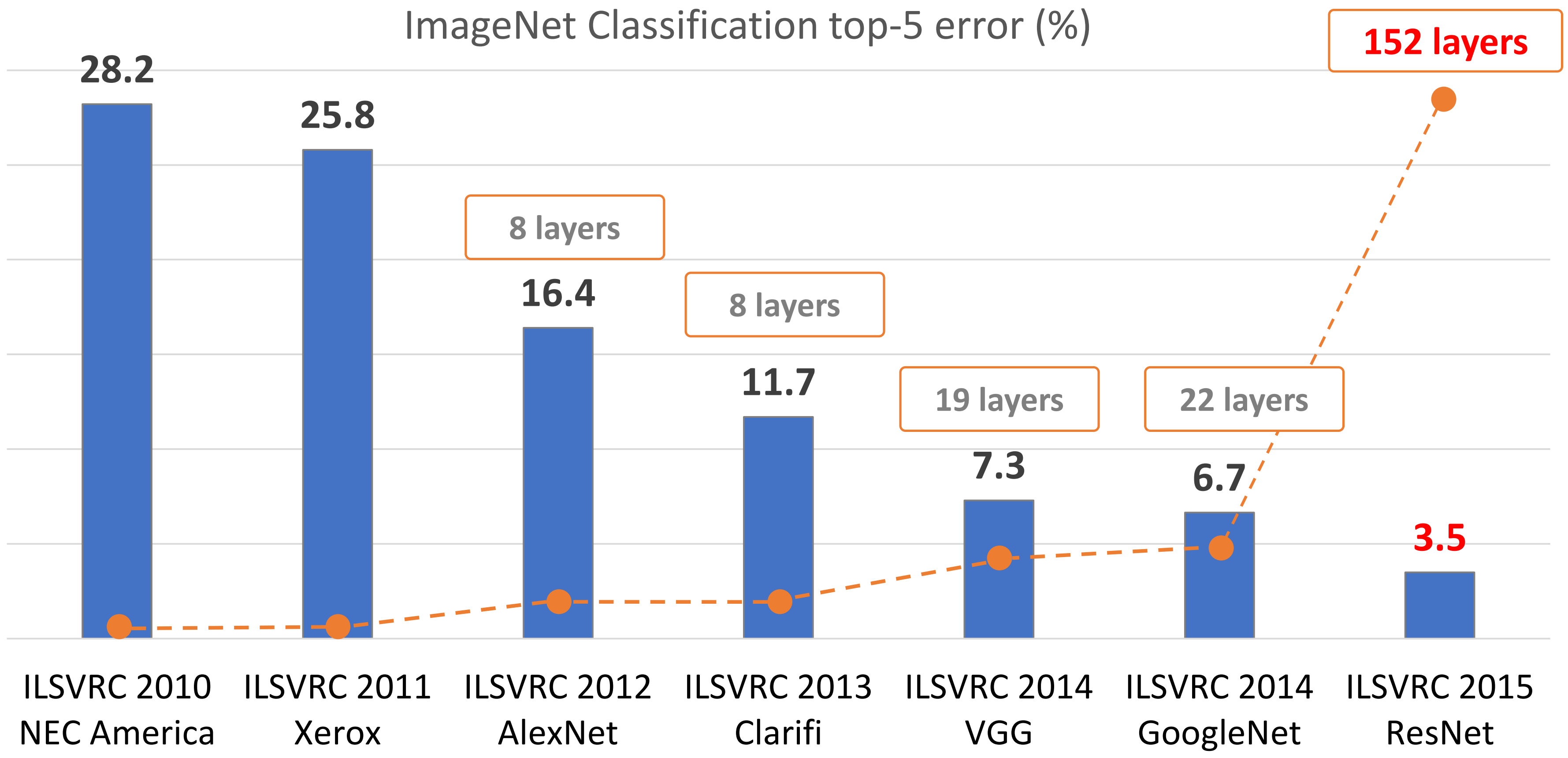
MSR reduced error from **~23%** to **<13%** (and under 7% for Rick Rashid's S2S demo)!



Slide from Li Deng

ImageNet Large Scale Visual Recognition Challenge (ILSVRC)

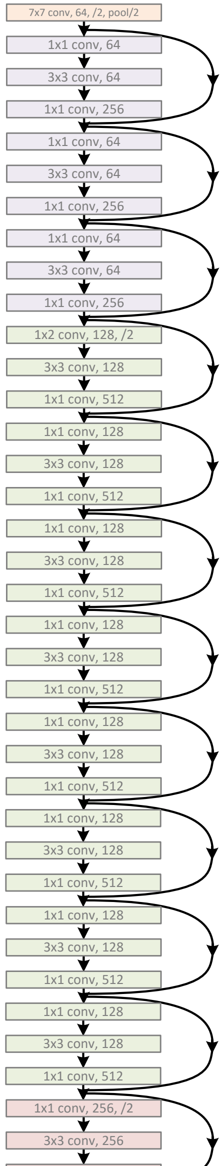
ImageNet Large Scale Visual Recognition Challenge (ILSVRC)



Microsoft

Slide from Jian Sun

Revolution of Depth: ResNet w. 152 layers



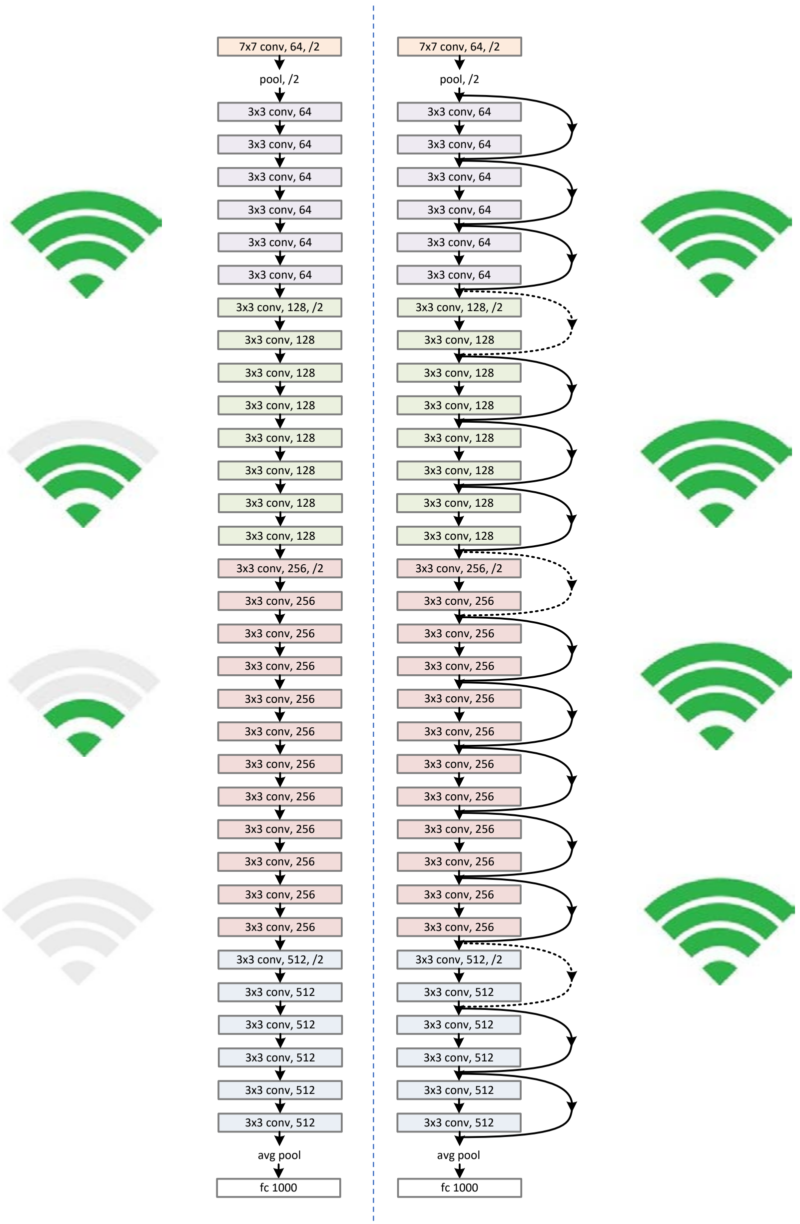
Slide from Jian Sun

Plain deep network

signals flow thru a single path

Deep residual network

signals flow thru many paths



[Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. “Deep Residual Learning for Image Recognition”. arXiv 2015.]

The focus of this tutorial

- is NOT on speech or image,
- but on natural language processing (NLP).

- What is NLP?
- The transition of NLP to neural methods

Traditional definition of NLP: the branch of AI

- Deal with analyzing, understanding and generating the languages that humans use naturally (natural language)
- Study *knowledge of language* at different levels
  - Phonetics and Phonology – the study of linguistic sounds
  - Morphology – the study of the meaning of components of words
  - Syntax – the study of the structural relationships between words
  - Semantics – the study of meaning
  - Discourse – they study of linguistic units larger than a single utterance

[Jurafsky & Martin 10]

Pragmatic definition: building computer systems

- Process large text corpora, turning information into knowledge
  - Text classification
  - Information retrieval and extraction
  - Machine reading comprehension and question answering
  - …
- Enable human-computer interactions, making knowledge accessible to humans in the most natural way
  - Dialogue and conversational agents
  - Machine translation
  - …

Challenge of NLP: the diversity of natural language

Many-to-many mapping btw *symbolic* language and *semantic* meaning

Ambiguity

Example: I made her duck.

- *I cooked waterfowl for her.*
- *I cooked waterfowl belonging to her.*
- *I created the plaster duck she owns.*
- *I caused her to quickly lower her head or body.*
- *I waved my magic wand and turned her into undifferentiated waterfowl.*

Paraphrase

Example: How long is the X river?

- *The Mississippi River is 3,734 km (2,320 mi) long.*
- *...is a short river, some 4.5 miles (7.2 km) in length*
- *The total length of the river is 2,145 kilometers.*
- *… at the estimated length of 5,464 km (3,395 mi)…*
- *… has a meander length of 444 miles (715 km)…*
- *… Bali's longest river, measuring approximately 75 kilometers from source to mouth.*
- *The … mainstem is 2.75 miles (4.43 km) long although total distance from headwater source tributaries to the sea is 14 miles (23 km).*

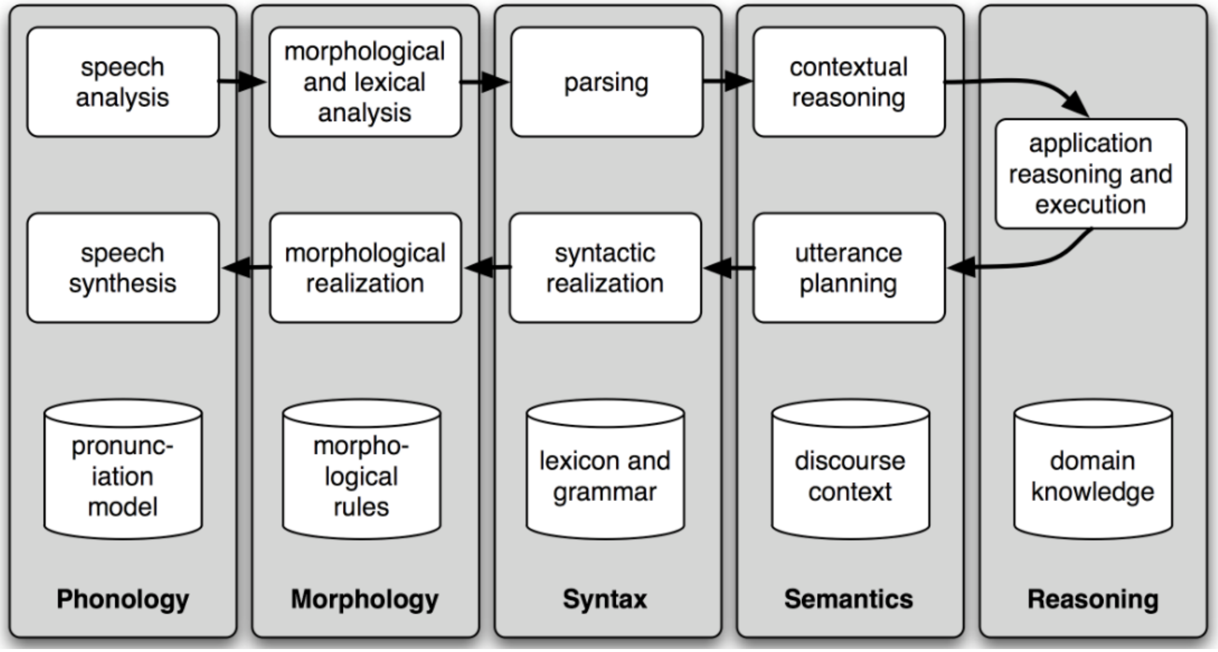
[Jurafsky & Martin 10; Dolan 17]

The transition of NLP to neural methods

- Paradigm shift in NLP: from symbolic to neural computation
- End-to-end learning simplifies systems, reduces effort for feature engineering and localization
- New state of the art results both at the component level and end-application
- Opens up new end applications and experience
- Large-scale (GPU) computing resources are critical
- Long-term success relies on BIG data

[Arul Menezes & Bill Dolan. 2017. The transition of natural language processing to neural methods. Unpublished report.]

Traditional NLP component stack



1. **Natural language understand (NLU):** parsing (speech) input to semantic meaning and update the system state
2. **Application reasoning and execution:** take the next action based on state
3. **Natural language generation (NLG):** generating (speech) response from action

[Menezes & Dolan 17]

DL leads to a paradigm shift in NLP

Traditional symbolic approaches

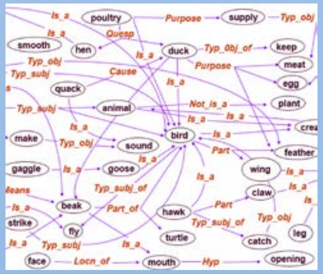
- Discrete, symbolic space
- Human comprehensible
  - easy to debug
- Computationally inefficient
  - Sensitive to ambiguity/paraphrase
  - Cascaded models prone to error propagation and require careful feature engineering


Squire Trelawney, Dr. Livesey, and the rest of these gentlemen having asked me to write down the whole particulars about Treasure Island, from the beginning to the end, keeping nothing back but the bearings of the island, and that only because there is still treasure not yet lifted, I take up my pen in the year of grace 17— and go back to the time when my father kept the Admiral Benbow inn and the brown old seaman with the sabre cut first took up his lodging under our roof.

I remember him as if it were yesterday, as he came plodding to the inn door, his sea-chest following behind him in a hand-barrow; a tall, strong, heavy, nut-brown man, his tarry pigtail falling over the shoulder of his soiled blue coat, his hands ragged and scarred, with black, broken nails, and the sabre cut across one cheek, a dirty, livid white. I remember him looking round the cover and whistling to himself as he did so, and then breaking out in that old sea-song that he sang so often afterwards:

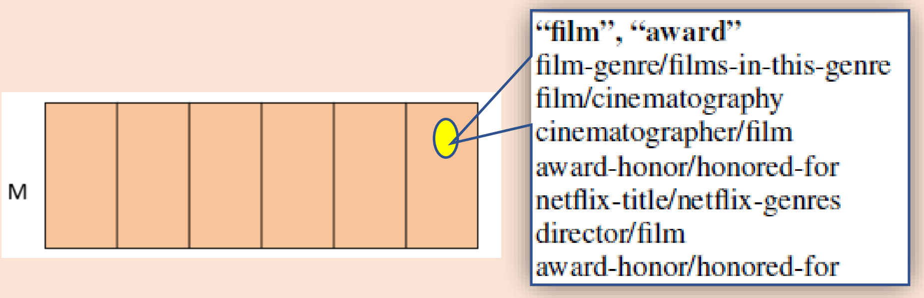
'Fifteen men on the dead man's chest-Yo-ho-ho, and a bottle of rum!' in the high, old tottering voice that seemed to have been tuned and broken at the capstan bars. Then he rapped on the door with a bit of stick like a handspike that he carried, and when my father appeared, called roughly for a glass of rum. This, when it was brought to him, he drank slowly, like a connoisseur, lingering on the taste and still looking about him at the cliffs and up at our signboard.

'This is a handy cove,' says he at length; 'and a pleasant sittyated



Deep Learning (DL) approaches

- Continuous, neural space
- Human incomprehensible
  - hard to debug
- Computationally efficient
  - Robust to ambiguity/paraphrase
  - E2E learning leads to better performance and simplified systems



E2E approaches based on DL

Discrete, symbolic space

- Human comprehensible
- Input: x
- Output: y

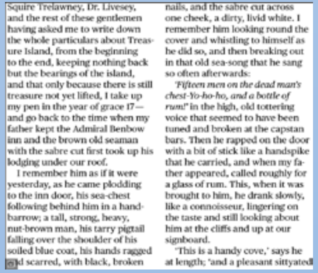
$\vec{\boldsymbol{x}} = \boldsymbol{f_e}(\boldsymbol{h_x}; \boldsymbol{\theta_e})$, Symbolic → Neural
by embedding models / encoder

$\frac{\partial \mathcal{L}}{\partial \theta_e}$ $\frac{\partial \mathcal{L}}{\partial \theta_r}$ $\frac{\partial \mathcal{L}}{\partial \theta_d}$

$\boldsymbol{h_y} = \boldsymbol{f_r}(\boldsymbol{h_x}; \boldsymbol{\theta_r})$,
Reasoning in neural space

Continuous, neural space

- Computationally efficient
- Input: h_x
- Output: h_y



Squire Trelawney, Dr. Livesey, and the rest of these gentlemen having asked me to write down the whole particulars about Treasure Island, from the beginning to the end, keeping nothing back but the bearings of the island, and that only because there is still treasure not yet lifted, I take up my pen in the year of grace 17— and go back to the time when my father kept the Admiral Benbow inn and the brown old seaman with the sabre cut first took up his lodging under our roof.

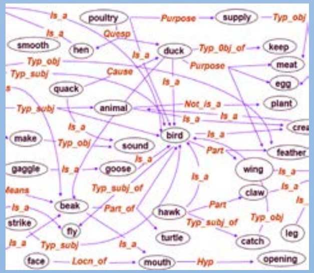
I remember him as if it were yesterday, as he came plodding to the inn door, his sea-chest following behind him in a hand-barrow; a tall, strong, heavy, nut-brown man, his tarry pigtail falling over the shoulder of his soiled blue coat, his hands ragged and scarred, with black, broken nails, and the sabre cut across one cheek, a dirty, livid white. I remember him looking round the cover and whistling to himself as he did so, and then breaking out in that old sea-song that he sang so often afterwards:

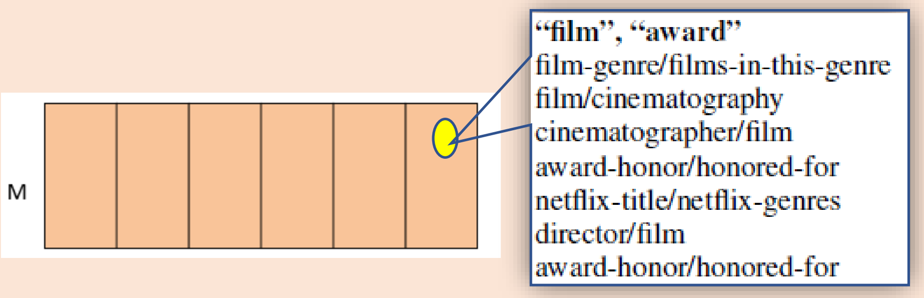
'Fifteen men on the dead man's chest-Yo-ho-ho, and a bottle of rum!' in the high, old tottering voice that seemed to have been tuned and broken at the capstan bars. Then he rapped on the door with a bit of stick like a handspike that he carried, and when my father appeared, called roughly for a glass of rum. This, when it was brought to him, he drank slowly, like a connoisseur, lingering on the taste and still looking about him at the cliffs and up at our signboard.

'This is a handy cove,' says he at length; 'and a pleasant sittyated

$\mathcal{L}(\theta) \propto Error(y, y^*)$

$\boldsymbol{y} = \boldsymbol{f_d}(\boldsymbol{h_y}; \boldsymbol{\theta_d})$, Neural → Symbolic
by generative models / decoder





State of the art results on NLP application-level tasks

Task	Test set	Metric	Best non-neural	Best neural	Source
Machine Translation	Enu-deu newstest16	BLEU	31.4	34.8	http://matrix.statmt.org
	Deu-enu newstest16	BLEU	35.9	39.9	http://matrix.statmt.org
Sentiment Analysis	Stanford sentiment bank	5-class Accuracy	71.0	80.7	Socher+ 13
Question Answering	WebQuestions test set	F1	39.9	52.5	Yih+ 15
Entity Linking	Bing Query Entity Linking set	AUC	72.3	78.2	Gao+ 14b
Image Captioning	COCO 2015 challenge	Turing test pass%	25.5	32.2	Fang+ 15
Sentence compression	Google 10K dataset	F1	0.75	0.82	Fillipova+ 15
Response Generation	Sordoni dataset	BLEU-4	3.98	5.82	Li+ 16a

[Menezes & Dolan 17]

State of the art results on NLP component tasks

Task	Test set	Metric	Best non-neural	Best neural	Source
POS tagging	PTB section 23	F1	97.17	97.78	Andor+ 16
Syntactic Parsing	PTB section 23	F1	90.1	93.3	Dyer+ 16
Dependency parsing	PTB section 23	F1	93.22	94.61	Andor+ 16
CCG parsing	CCGBank test	F1	85.2	88.7	Lee+ 16
Inference (NLI)	Stanford NLI corpus	Accuracy	78.2	88.3	Chen+ 16

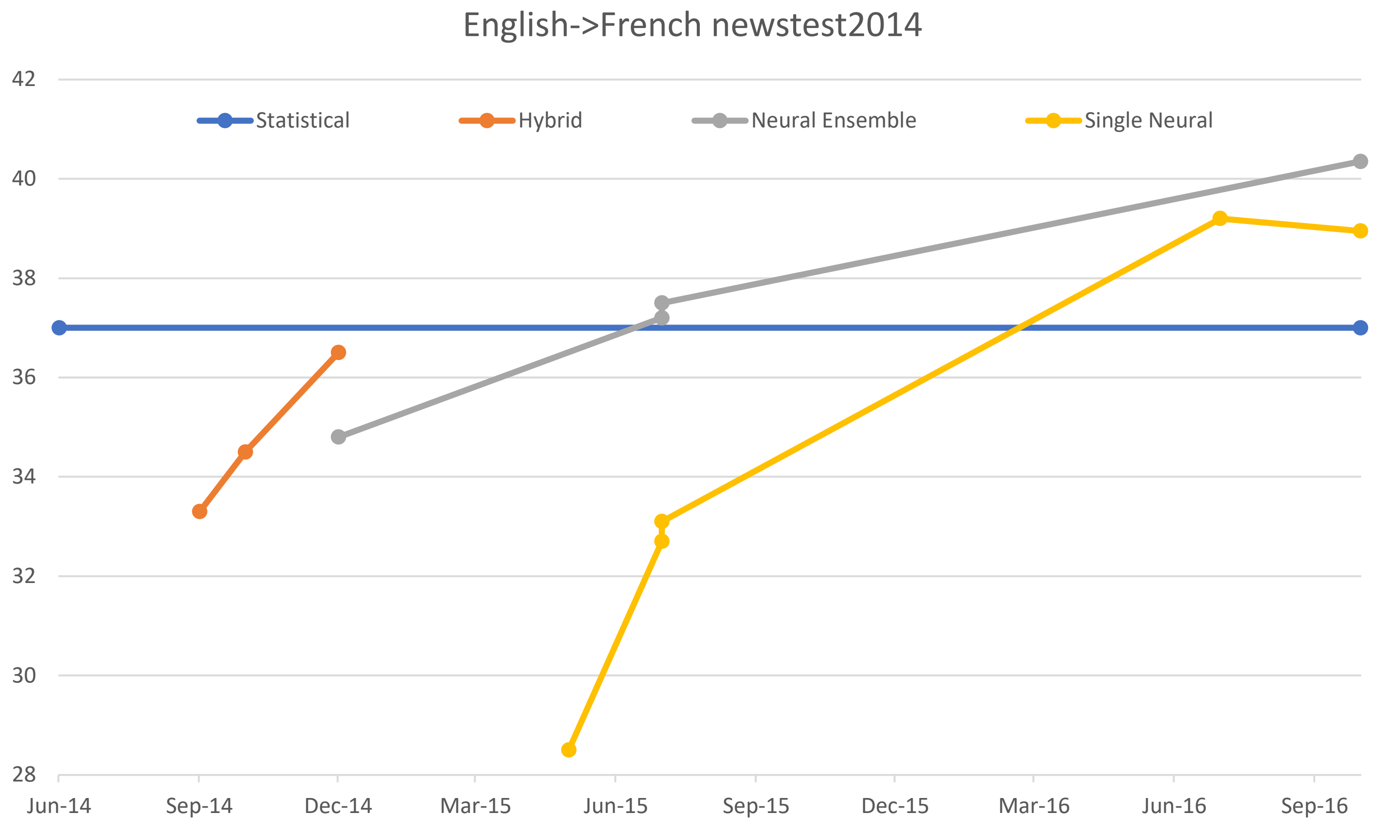
Also see a summary by [Goldberg 15]

[Menezes & Dolan 17]

Recent progress on machine translation (MT): BLEU score of state of the art systems

- Statistical MT state of the art is highly engineered and has made little progress in over two years

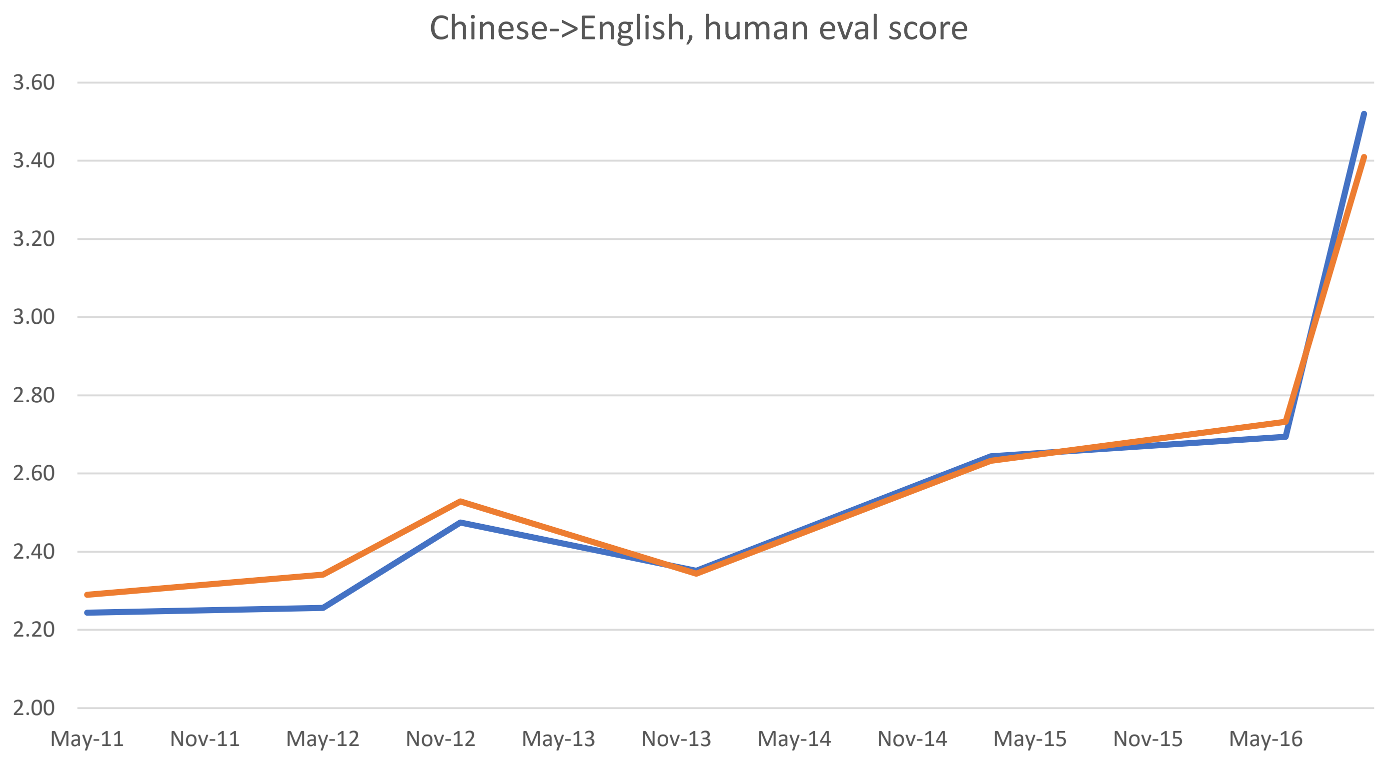
- Chart shows progress in three classes of neural systems
  - Hybrid: Add neural models to existing statistical MT system
  - Single: Single pure-neural system (to be discussed in Part 3)
  - Ensemble: Large ensemble of pure-neural systems



[Menezes & Dolan 17]

Human evaluation of MT quality

- Two popular commercial MT systems
- Human evaluation on a scale of 1 to 4
- Human preference for neural MT is much greater than the already large BLEU score gap
- Primarily because neural MT is much more fluent
- Current neural MT systems are getting close to “human quality”



[Menezes & Dolan 17]

DL opens up new end tasks and experience

Neural approaches allow language models to be grounded in the world, i.e., link language to real-world signals such as images, machine state, sensor data from biomedical devices.

photo + tweet: *i forgot to take a pic before i took a bite*
generated reply: ***is that ice cream*** (-0.63088092890878522)

photo + tweet: who wants some paleo donuts ?
generated reply: **oh my gosh oh my gosh can i have one** (-0.62831618986092508)

Output of a neural conversation model trained on 250K Twitter conversations sparked by a tweeted photo

[Menezes & Dolan 17; Sordoni+ 15; Li+ 16a]

A text (query) classification problem

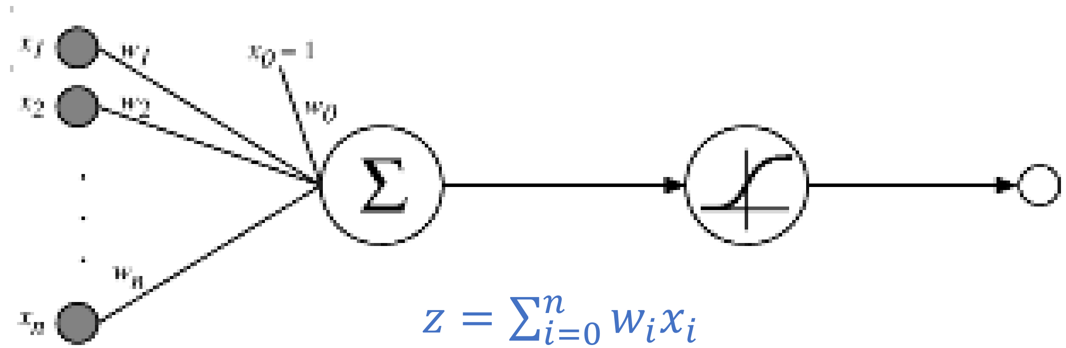
- Given a search query q, e.g., “denver sushi downtown”
- Identify its domain c e.g.,
  - Restaurant
  - Hotel
  - Nightlife
  - Flight
  - etc.
- So that a search engine can tailor the interface and result to provide a richer personalized user experience

A single neuron model

- For each domain c, build a binary classifier
  - Input: represent a query q as a vector of features $x = [x_1, \dots x_n]^T$
  - Output: $y = P(c|q)$
  - q is labeled c if $P(c|q) > 0.5$
- Input feature vector, e.g., a bag of words vector
  - Regards words as atomic symbols: *denver, sushi, downtown*
  - Each word is represented as a one-hot vector: $[0, \dots, 0,1,0, \dots, 0]^T$
  - Bag of words vector = sum of one-hot vectors
  - We may use other features, such as n-grams, phrases, (hidden) topics

A single neuron model

Input features x

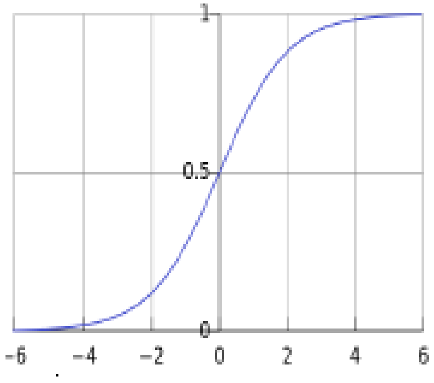


$z = \sum_{i=0}^{n} w_i x_i$

Output: $P(c|q)$

$y = \sigma(z) = \frac{1}{1+\exp(-z)}$

- w: weight vector to be learned
- z: weighted sum of input features
- σ: the logistic function
  - Turn a score to a probability
  - A sigmoid non-linearity (activation function), essential in multi-layer/deep neural network models



Model training: how to assign w

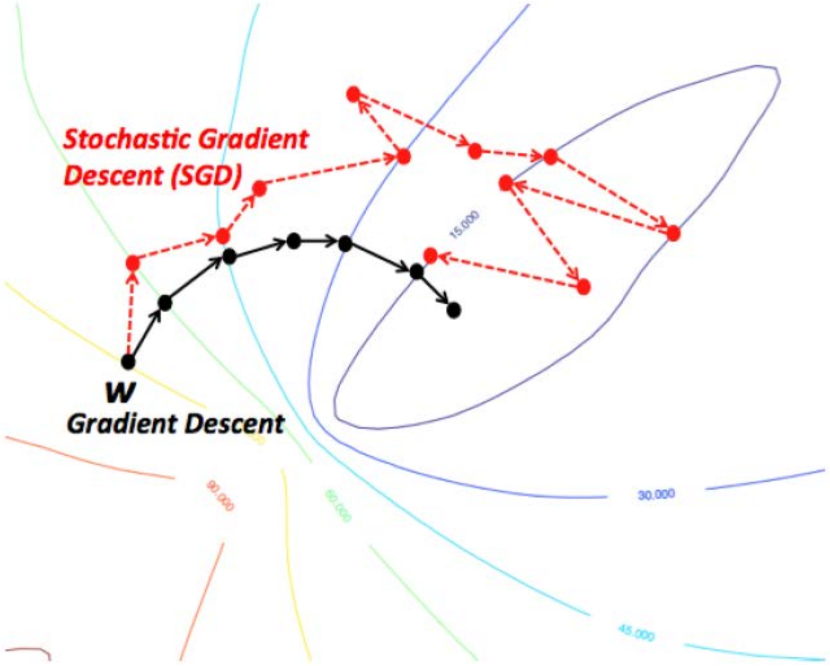
- Training data: a set of $\left(x^{(m)}, y^{(m)}\right)_{m=\{1,2,\ldots,M\}}$ pairs
  - Input $x^{(m)} \in R^n$
  - Output $y^{(m)} = \{0,1\}$
- Goal: learn function $f: x \to y$ to predict correctly on new input x
  - Step 1: choose a function family, e.g.,
    - neural networks, logistic regression, support vector machine, in our case
    - $f_w(x) = \sigma(\sum_{i=0}^{n} w_i x_i) = \sigma(w^T x)$
  - Step 2: optimize parameters w on training data, e.g.,
    - minimize a loss function (mean square error loss)
    - $\min\limits_{w} \sum_{m=1}^{M} L^{(m)}$
    - where $L^{(m)} = \frac{1}{2}\left(f_w\left(x^{(m)}\right) - y^{(m)}\right)^2$

Training the single neuron model, w

- Stochastic gradient descent (SGD) algorithm
  - Initialize w randomly
  - Update for each training sample until convergence: $w^{new} = w^{old} - \eta \frac{\partial L}{\partial w}$
- Mean square error loss: $L = \frac{1}{2}(\sigma(w^T x) - y)^2$
- Gradient: $\frac{\partial L}{\partial w} = \delta \sigma'(z) x$
  - $z = w^T x$
  - Error: $\delta = \sigma(z) - y$
  - Derivative of sigmoid $\sigma'(z) = \sigma(z)\big(1 - \sigma(z)\big)$

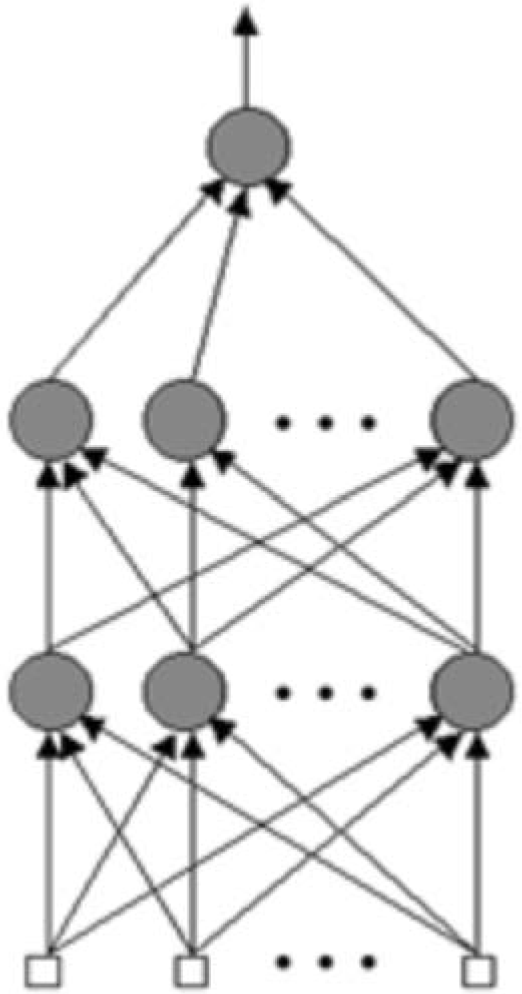
SGD vs. gradient descent

- Gradient descent is a batch training algorithm
  - update w per batch of training samples
  - goes in steepest descent direction
- SGD is noisy descent (but faster per iteration)
- Loss function contour plot
  - $\sum_{m=1}^{M} \frac{1}{2}(\sigma(w^T x) - y)^2 + ||w||$



[Duh, K. 2014. Deep learning for natural language processing and machine translation. Tutorial.]

Multi-layer (deep) neural networks



Output layer $y^o = \sigma(w^T y^2)$

Vector w

2st hidden layer $y^2 = \sigma(\mathbf{W}_2 y^1)$

Projection matrix $\mathbf{W}_2$

1st hidden layer $y^1 = \sigma(\mathbf{W}_1 x)$

Projection matrix $\mathbf{W}_1$

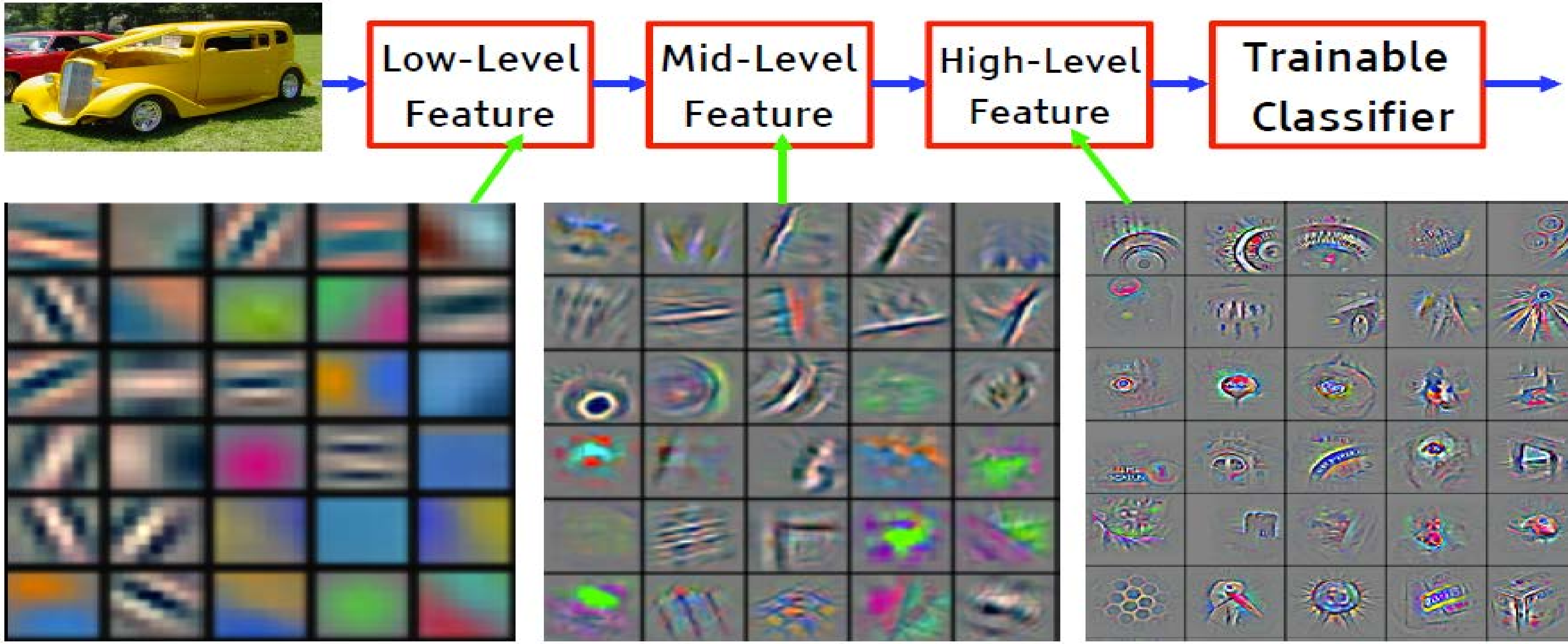
Input features x

This is exactly the **single neuron model** with **hidden** features.

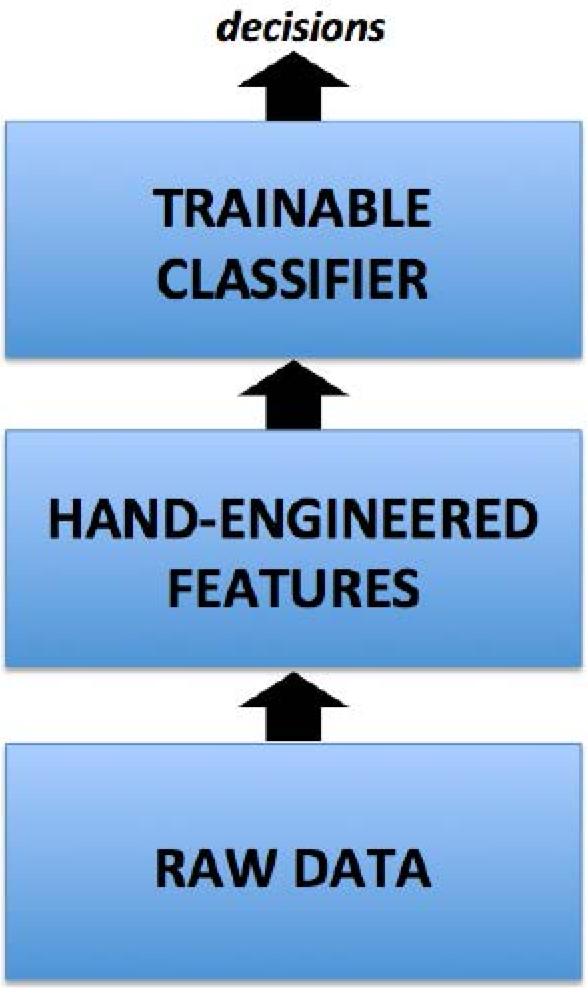
Feature generation: project raw input features (bag of words) to **hidden** features (topics).

Why Multiple Layers? [DL tutorial at NIPS'2015]

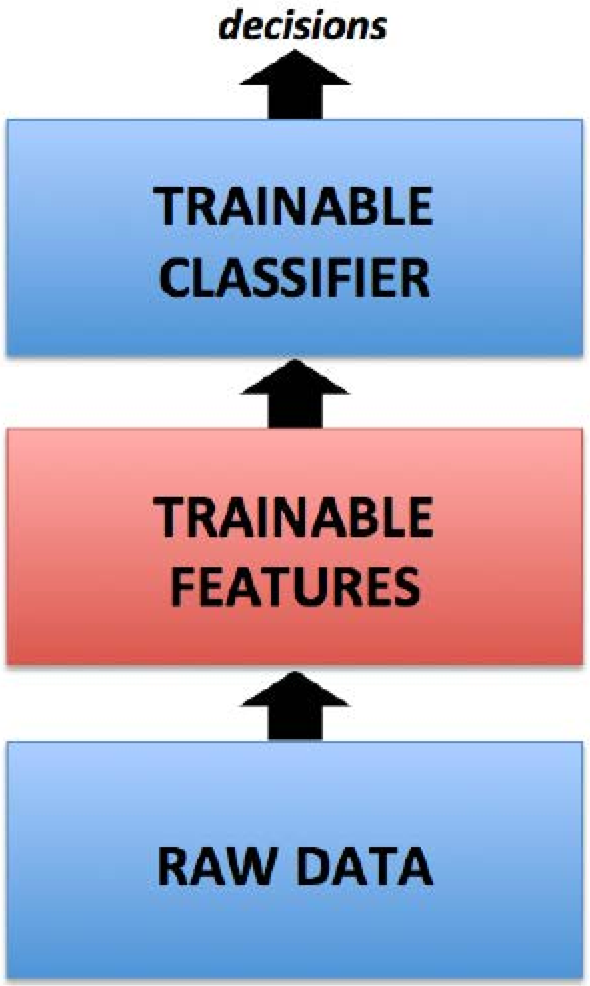
- Hierarchy of representations with increasing level of abstraction
- Each layer is a trainable feature transform
- Image recognition: pixel → edge → texton → motif → part → object
- ?? Text: character → word → word group → clause → sentence → story



Standard Machine Learning Process



Deep Learning



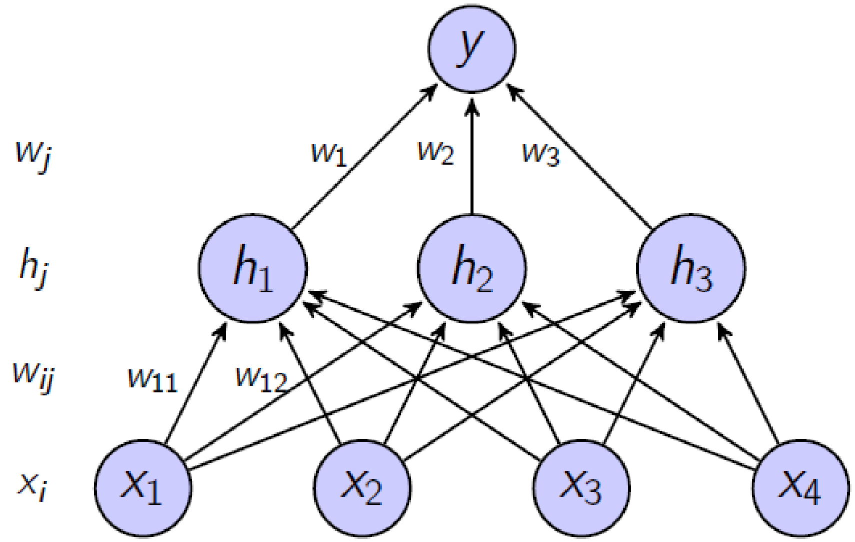
Adapted from [Duh 14]

Revisit the activation function: σ

- Assuming a L-layer neural network
  - $y = \mathbf{W}_L \sigma\left(\ldots \sigma\left(\mathbf{W}_2 \sigma(\mathbf{W}_1 x)\right)\right)$, where y is the output vector
- If σ is a linear function, then L-layer neural network is compiled down into a single linear transform
- σ: map scores to probabilities
  - Useful in prediction as it transforms the neuron weighted sum into the interval [0..1]
  - Unnecessary for model training except in the Boltzman machine or graphical models

Training a two-layer neural net

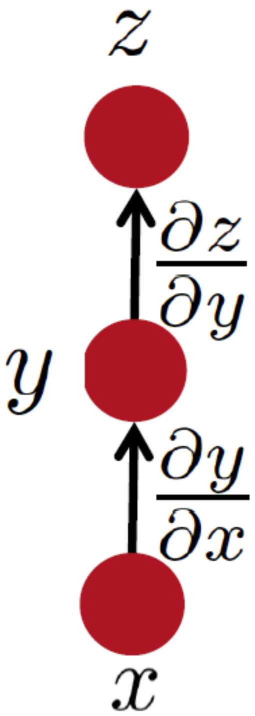
- Training data: a set of $\left(x^{(m)}, y^{(m)}\right)_{m=\{1,2,\ldots,M\}}$ pairs
  - Input $x^{(m)} \in R^n$
  - Output $y^{(m)} = \{0,1\}$
- Goal: learn function $f: x \rightarrow y$ to predict correctly on new input x
  - $f_w(x) = \sigma\left(\sum_j w_j \cdot \sigma(\sum_i w_{ij} x_i)\right)$
  - Optimize parameters w on training data via
    - minimize a loss function: $\min_w \sum_{m=1}^{M} L^{(m)}$
    - where $L^{(m)} = \frac{1}{2}\left(f_w\left(x^{(m)}\right) - y^{(m)}\right)^2$



Training neural nets: back-propagation

- Stochastic gradient descent (SGD) algorithm
  - $w^{new} = w^{old} - \eta \frac{\partial L}{\partial w}$
- $\frac{\partial L}{\partial w}$: sample-wise loss w.r.t. parameters
- Need to apply the derivative chain rule correctly
  - $z = f(y)$
  - $y = g(x)$
  - $\frac{\partial z}{\partial x} = \frac{\partial z}{\partial y} \frac{\partial y}{\partial x}$
- See a detailed discussion in [Socher & Manning 13; Goodfellow+ 16]

Simple chain rule



$$\Delta z = \frac{\partial z}{\partial y} \Delta y$$

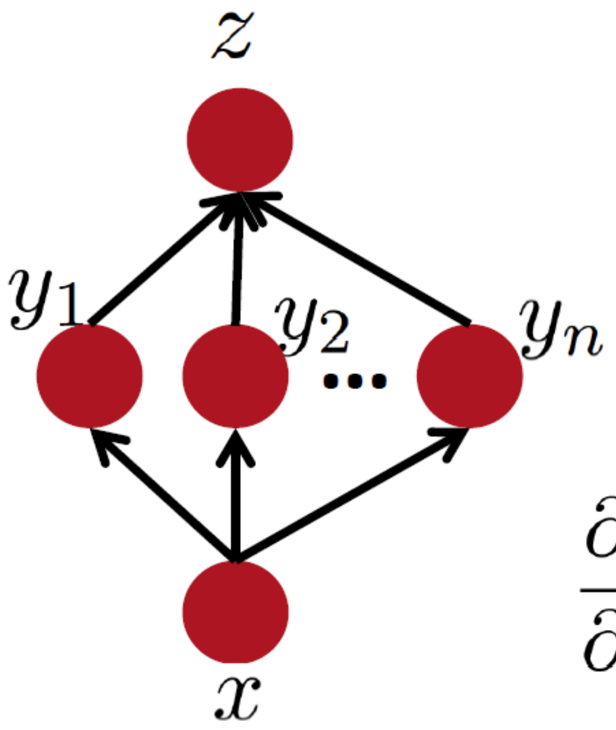
$$\Delta y = \frac{\partial y}{\partial x} \Delta x$$

$$\Delta z = \frac{\partial z}{\partial y} \frac{\partial y}{\partial x} \Delta x$$

$$\frac{\partial z}{\partial x} = \frac{\partial z}{\partial y} \frac{\partial y}{\partial x}$$

[Socher & Manning NAACL 2013 Tutorial]

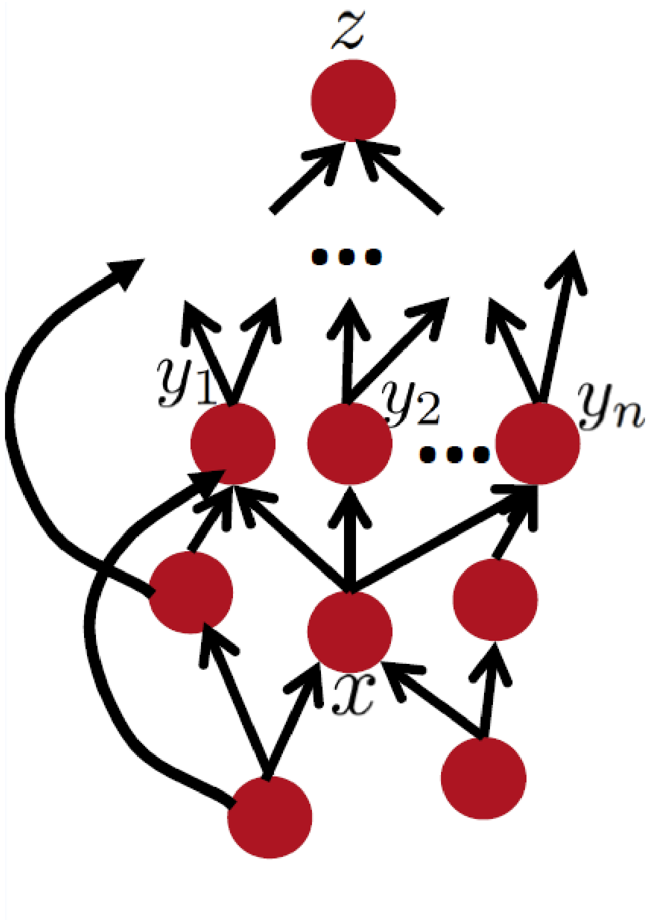
Multiple paths chain rule



$$\frac{\partial z}{\partial x} = \sum_{i=1}^{n} \frac{\partial z}{\partial y_i} \frac{\partial y_i}{\partial x}$$

[Socher & Manning 13]

Chain rule in flow graph



Flow graph: any directed acyclic graph

node = computation result

arc = computation dependency

$\{y_1,\ y_2,\ \ldots\ y_n\}$ = successors of x

$$\frac{\partial z}{\partial x} = \sum_{i=1}^{n} \frac{\partial z}{\partial y_i} \frac{\partial y_i}{\partial x}$$

[Socher & Manning 13]

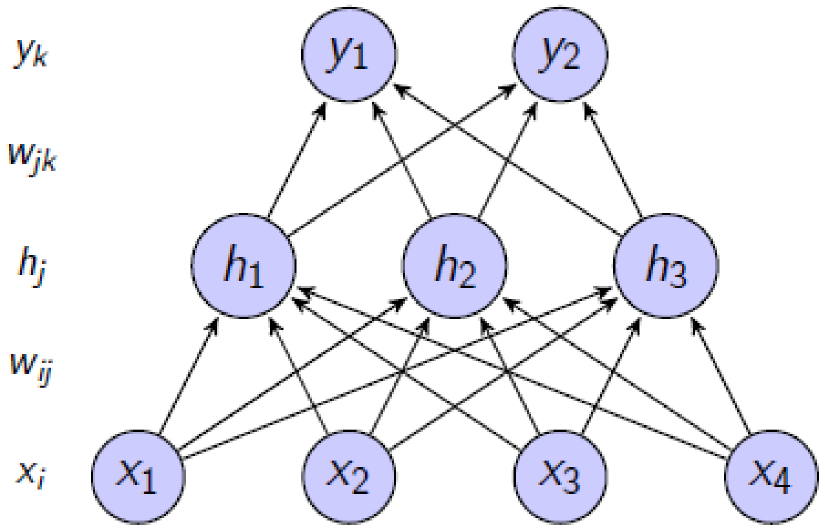
Training neural nets: back-propagation

Assume two outputs (y_1, y_2) per input x, and

Loss per sample: $L = \sum_k \frac{1}{2}(\sigma(z_k) - y_k)^2$

Forward pass:

$$y_k = \sigma(z_k),\ z_k = \sum_j w_{jk} h_j$$
$$h_j = \sigma(z_j),\ z_j = \sum_i w_{ij} x_i$$



Derivatives of the weights

$$\frac{\partial L}{\partial w_{jk}} = \frac{\partial L}{\partial z_k}\frac{\partial z_k}{\partial w_{jk}} = \delta_k \frac{\partial(\sum_j w_{jk} h_j)}{\partial w_{jk}} = \delta_k h_j$$
$$\frac{\partial L}{\partial w_{ij}} = \frac{\partial L}{\partial z_j}\frac{\partial z_j}{\partial w_{ij}} = \delta_j \frac{\partial(\sum_i w_{ij} x_i)}{\partial w_{ij}} = \delta_j x_i$$
$$\delta_k = \frac{\partial L}{\partial z_k} = (\sigma(z_k) - y_k)\sigma'(z_k)$$
$$\delta_j = \sum_k \frac{\partial L}{\partial z_k}\frac{\partial z_k}{\partial z_j} = \sum_k \delta_k \frac{\partial}{\partial z_j}\left(\sum_j w_{jk}\sigma(z_j)\right) = \left(\sum_k \delta_k w_{jk}\right)\sigma'(z_j)$$

Adapted from [Duh 14]

Training neural nets: back-propagation

- All updates involve some scaled error from output × input feature:
  - $\frac{\partial L}{\partial w_{jk}} = \delta_k h_j$ where $\delta_k = (\sigma(z_k) - y_k)\sigma'(z_k)$
  - $\frac{\partial L}{\partial w_{ij}} = \delta_j x_i$ where $\delta_j = \left(\sum_k \delta_k w_{jk}\right)\sigma'(z_j)$
- First compute δ_k from output layer, then δ_j for other layers and iterate.

- All updates involve some scaled error from output × input feature:
  - $\frac{\partial L}{\partial w_{jk}} = \delta_k h_j$ where $\delta_k = (\sigma(z_k) - y_k)\sigma'(z_k)$
  - $\frac{\partial L}{\partial w_{ij}} = \delta_j x_i$ where $\delta_j = \left(\sum_k \delta_k w_{jk}\right)\sigma'(z_j)$
- First compute δ_k from output layer, then δ_j for other layers and iterate.

$\delta_{k=y_1}$ $\delta_{k=y_2}$

w_{31} w_{32}

$$\delta_{j=h_3} = \left(\delta_{k=y_1} w_{31} + \delta_{k=y_2} w_{32}\right)\sigma'(z_{j=h_3})$$

Adapted from [Duh 14]

DNN forms for different language structures

- Text as a bag of words: Multi-Layer Perceptron (MLP)
- Text as a bag of chunks: Convolutional Neural Network (CNN)
- Text as a sequence of words: Recurrent Neural Network (RNN)
- *Text as a sequence of chunks: ???*

DNN models for the NLP tasks in this tutorial

- Classification task – label x by y
  - MLP/CNN/RNN as feature generator
- Ranking task – compute the semantic similarity btw x and y
  - Siamese neural network [Bromley et al. 1993]
  - Deep Semantic Similarity Model (DSSM)
- (Text) Generation task – generate y from x
  - Seq2Seq (RNN/LSTM)
  - Memory Network
- Question answering task
  - Neural machine reading models
- Task-completion dialogue
  - Deep reinforcement learning for dialogue agents

Tutorial Outline

- Part 1: Background
- **Part 2: Deep Semantic Similarity Models (DSSM) for text processing**
  - **Challenges of modeling semantic similarity**
  - **What is DSSM**
  - **DSSM for web search ranking**
  - **DSSM for recommendation**
  - **DSSM for automatic image captioning and other tasks**
- Part 3: Recurrent neural networks for text generation
- Part 4: Neural machine reading models for question answering
- Part 5: Deep reinforcement learning for task-completion dialogue

Computing Semantic Similarity

- Fundamental to almost all NLP tasks, e.g.,
  - Machine translation: similarity between sentences in different languages
  - Information retrieval: similarity between queries and documents

- Problems of the existing approaches
  - Lexical matching cannot handle language discrepancy.
  - Unsupervised word embedding or topic models are not optimal for the task of interest.

Deep Semantic Similarity Model (DSSM)

- Compute semantic similarity between two text strings X and Y
  - Map X and Y to feature vectors in a latent semantic space via deep neural net
  - Compute the cosine similarity between the feature vectors
  - Also called "Deep Structured Similarity Model" in [Huang+ 13]

Tasks	X	Y	Ref
Web search	*Search query*	*Web document*	Huang+ 13; Shen+ 14; Palangi+ 16
Entity linking	*Entity mention and context*	*Entity and its corresponding page*	Gao+ 14b
Online recommendation	*Doc in reading*	*Interesting things / other docs*	Gao+ 14b
Image captioning	*Image*	*Text*	Fang+ 15
Machine translation	*Sentence in language A*	*Translations in language B*	Gao+ 14a
Question answering	*Question*	*Answer*	Yih+ 15

Sent2Vec (DSSM) http://aka.ms/sent2vec

DSSM for web search ranking

- Task
- Model architecture
- Model training
- Evaluation
- Analysis

[Huang+ 13; Shen+ 14; Palangi+ 16]

An example of web search

Best Home Remedies for Cold and Flu

Wind Heat External Pathogens

By: Catherine Browne, L.Ac., MH, Dipl. Ac.

In Chinese medicine, colds and flu's are delineated into several different energetic classifications. Here we will outline the different types of cold and flu viruses that you will likely encounter, and then describe the best home remedies for these specific patterns that you can use to treat the cold or influenza virus.

Cold and Flu Basics

The basic pathogenic influences are:

- **Wind**
- **Cold**
- **Heat**
- **Damp**

Wind

Theoretically, wind enters the body through the back of the neck area or nose carrying the pathogen. It first attacks the Lung system (including the sinuses) because the Lung organ system is the most external Yin organ, a thus the most vulnerable to an external invasion. External Wind invasion is marked by acute conditions with a sudden onset of symptoms.

- cold home remedy
- cold remeedy
- flu treatment
- how to deal with stuffy nose

Semantic matching between Q and D

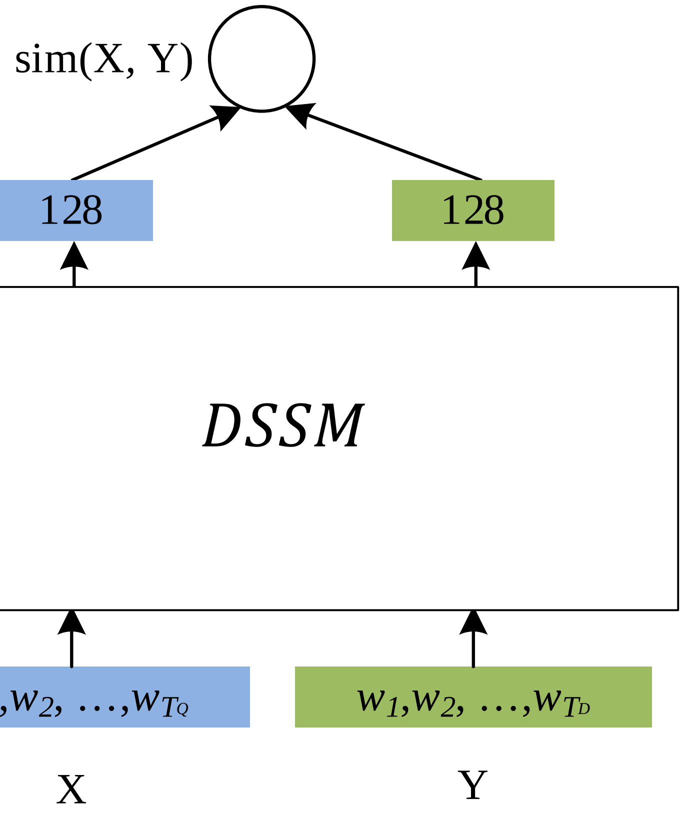
- Fuzzy keyword matching
  - Q: cold home remedy
  - D: best home remedies for cold and flu
- Spelling correction
  - Q: cold remeedies
  - D: best home remedies for cold and flu
- Query alteration/expansion
  - Q: flu treatment
  - D: best home remedies for cold and flu
- **Query/document semantic matching**
  - Q: how to deal with stuffy nose
  - D: best home remedies for cold and flu
  - Q: auto body repair cost calculator software
  - D: free online car body shop repair estimates

R&D progress

DSSM: Compute Similarity in Semantic Space

Relevance measured
by cosine similarity

Learning: maximize the similarity between X (source) and Y (target)

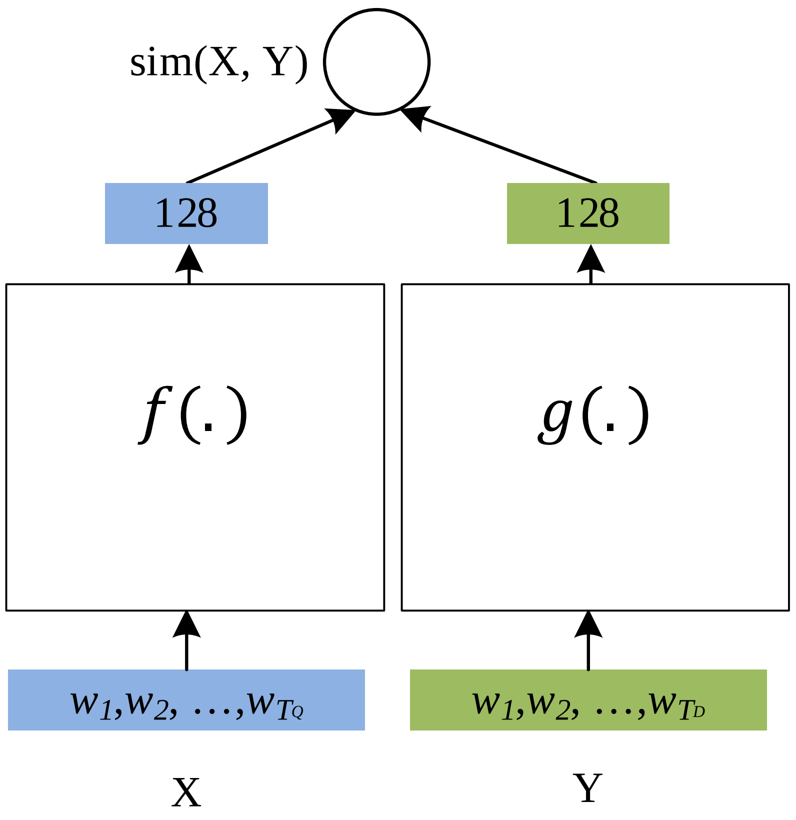


Word sequence x_t

DSSM: Compute Similarity in Semantic Space

Relevance measured by cosine similarity

Word sequence x_t



Learning: maximize the similarity between X (source) and Y (target)

Representation: use DNN to extract abstract semantic features, f or g is a

- Multi-Layer Perceptron (MLP) if text is a bag of words [Huang+ 13]
- **Convolutional Neural Network (CNN) if text is a bag of chunks** [Shen+ 14]
- Recurrent Neural Network (RNN) if text is a sequence of words [Palangi+ 16]

DSSM: Compute Similarity in Semantic Space

Relevance measured by cosine similarity

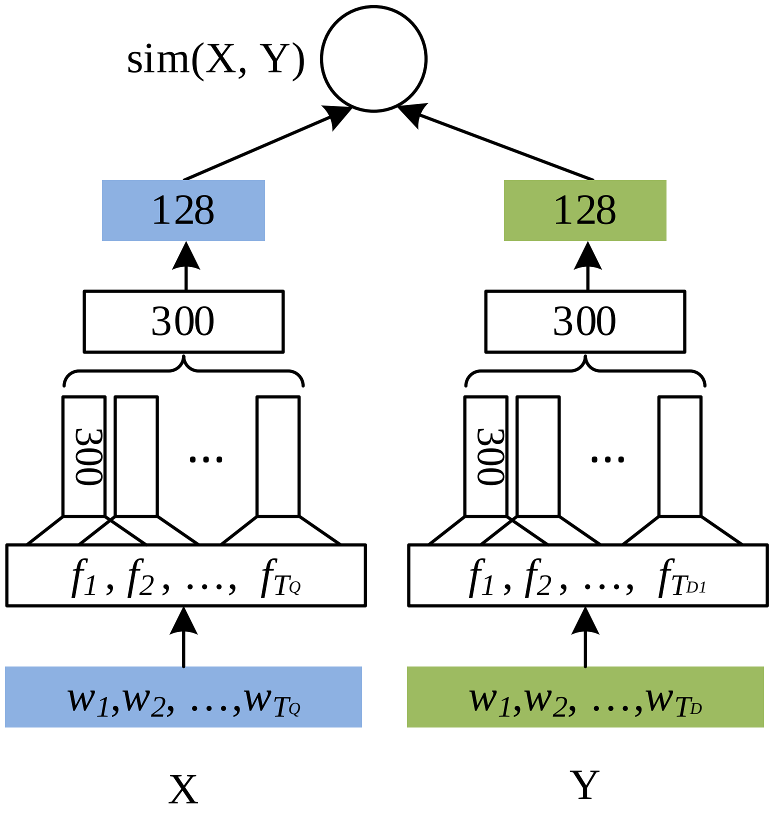
Semantic layer h

Max pooling layer v

Convolutional layer c_t

Word hashing layer f_t

Word sequence x_t



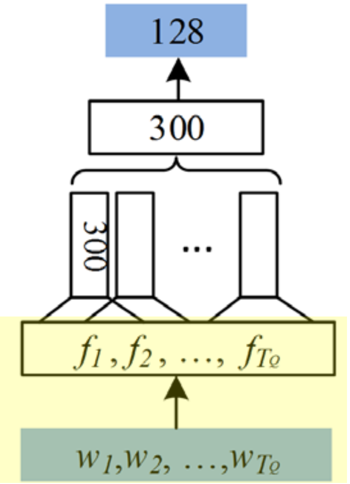
Learning: maximize the similarity between X (source) and Y (target)

Representation: use DNN to extract abstract semantic representations

Convolutional and Max-pooling layer: identify key words/concepts in X and Y

Word hashing: use sub-word unit (e.g., letter n-gram) as raw input to handle very large vocabulary

Letter-trigram Representation

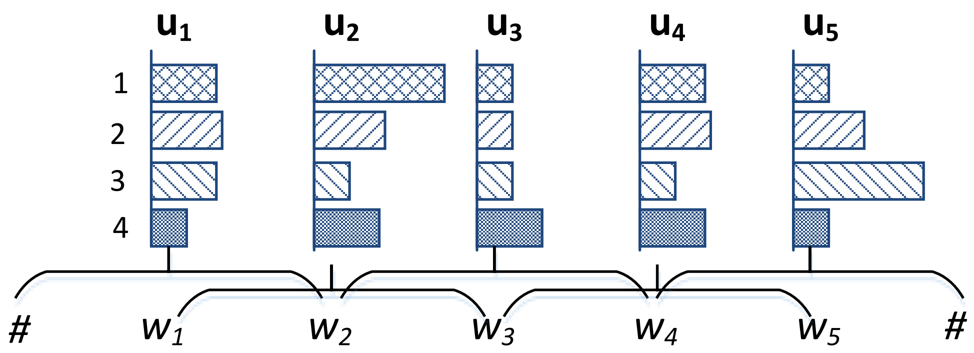


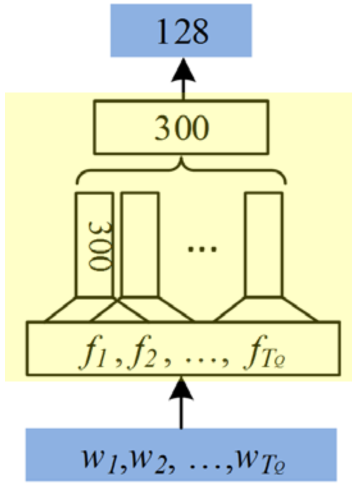
- Control the dimensionality of the input space
  - e.g., cat → #cat# → #-c-a, c-a-t, a-t-#
  - Only ~50K letter-trigrams in English; no OOV issue
- Capture sub-word semantics (e.g., prefix & suffix)
- Words with small typos have similar raw representations

- Collision: different words with same letter-trigram representation?

Vocabulary size	# of unique letter-trigrams	# of Collisions	Collision rate
40K	10,306	2	0.0050%
500K	30,621	22	0.0044%
5M	49,292	179	0.0036%

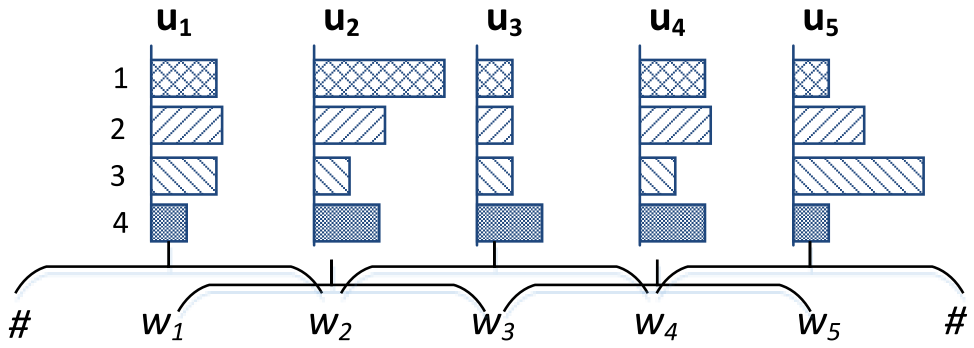
Convolutional Layer

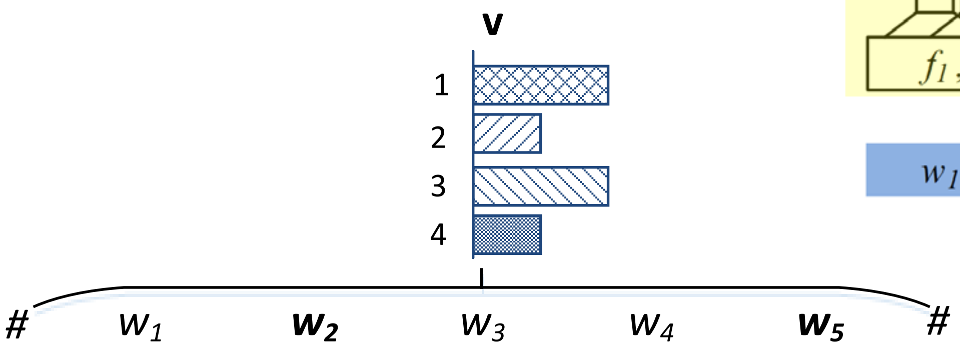


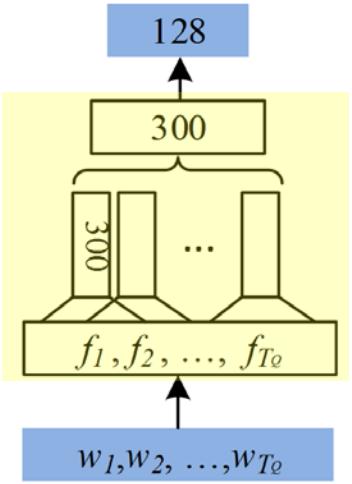


- Extract local features using convolutional layer
  - {w1, w2, w3} → topic 1
  - {w2, w3, w4} → topic 4

Max-pooling Layer

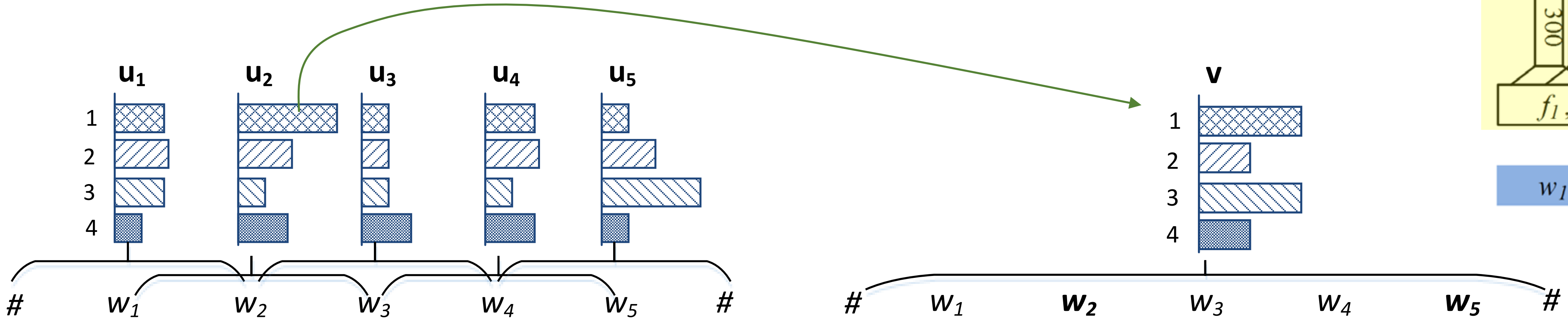


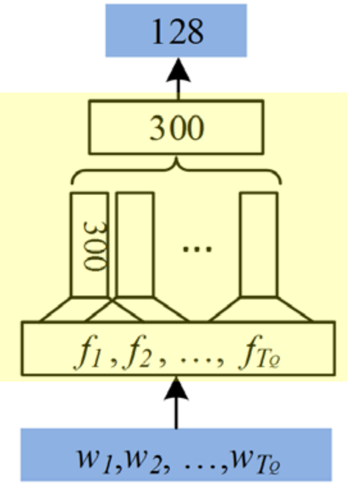




- Extract local features using convolutional layer
  - {w1, w2, w3} → topic 1
  - {w2, w3, w4} → topic 4
- Generate global features using max-pooling
  - Key topics of the text → topics 1 and 3
  - keywords of the text: w2 and w5

Max-pooling Layer



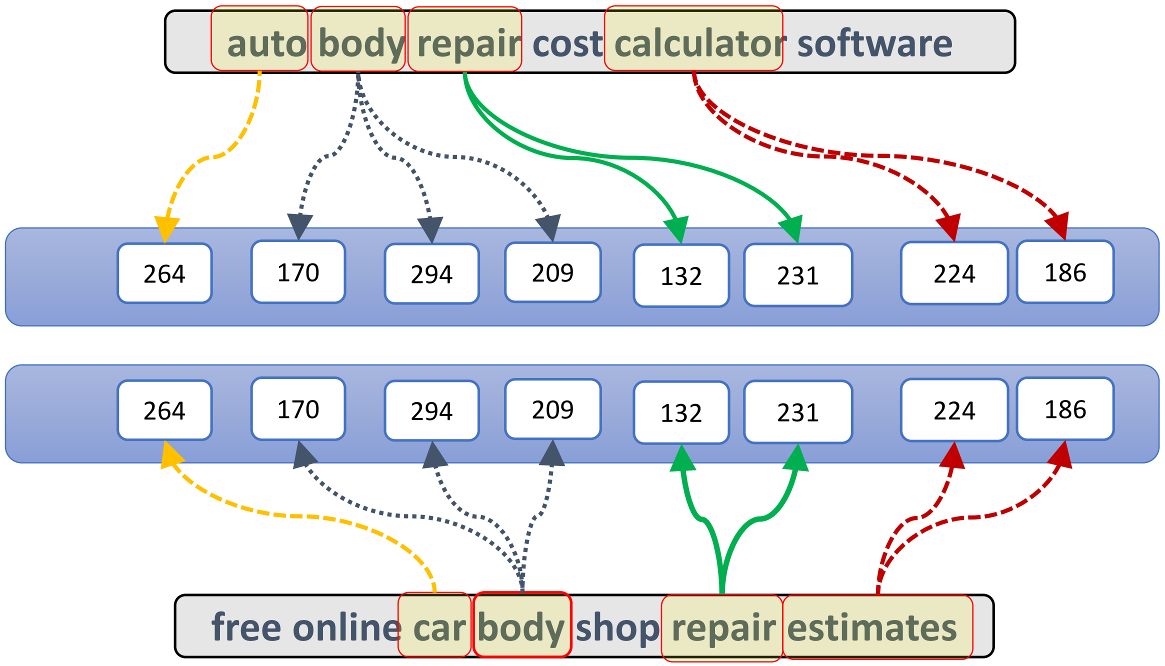


- Extract local features using convolutional layer
  - {w1, w2, w3} → topic 1
  - {w2, w3, w4} → topic 4
- Generate global features using max-pooling
  - Key topics of the text → topics 1 and 3
  - keywords of the text: w2 and w5

… the **comedy festival** formerly known as the us **comedy arts** festival is a comedy festival held each year in **las vegas nevada** from its 1985 inception to 2008 . it was held annually at the **wheeler opera house** and other venues in **aspen colorado** . the primary sponsor of the festival was hbo with co-sponsorship by caesars palace . the primary venue tbs **geico insurance** twix candy bars and **smirnoff vodka hbo** exited the festival business in 2007 …

Intent matching via convolutional-pooling

- Semantic matching of query and document



Most active neurons at the **max-pooling layers** of the query and document nets, respectively

More examples

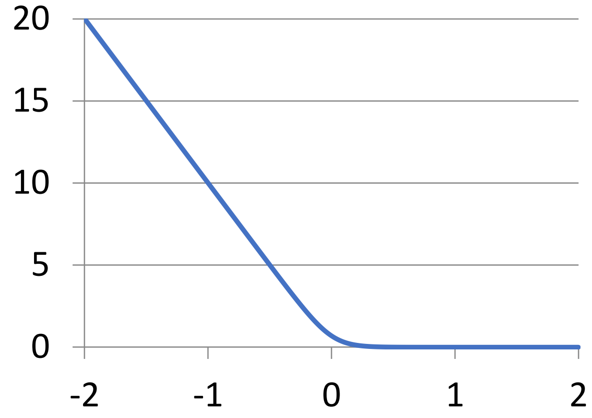
Query	Title of the top-1 returned document retrieved by CLSM
warm environment arterioles do what	**thermoregulation** wikipedia the free encyclopedia
auto body repair cost **calculator** software	free online **car body shop** repair **estimates**
what happens if our body **absorbs excessive** amount **vitamin d**	**calcium** supplements and **vitamin d** discussion stop **sarcoidosis**
how do **camera** use **ultrasound focus** automatically	**wikianswers how** does a **camera focus**
how to **change font excel office 2013**	**change font** default styles in **excel 2013**
where do i get my **federal tax return transcript**	how to get **trasncripts** of **federal** income **tax returns** fast **ehow**
12 **fishing boats trailers**	**trailer kits** and accessories **motorcycle** utility **boat snowmobile**
acp **ariakon combat pistol** 2.0	paintball acp **combat** pistol paintball **gun paintball pistol** package deal marker and gun

Learning DSSM from Labeled X-Y Pairs

- Consider a query X and two docs Y^+ and Y^-
  - Assume Y^+ is more relevant than Y^- with respect to X
- $\text{sim}_{\boldsymbol{\theta}}(X, Y)$ is the cosine similarity of X and Y in semantic space, mapped by DSSM parameterized by $\boldsymbol{\theta}$

Learning DSSM from Labeled X-Y Pairs

- Consider a query X and two docs Y^+ and Y^-
  - Assume Y^+ is more relevant than Y^- with respect to X
- $\text{sim}_{\boldsymbol{\theta}}(X, Y)$ is the cosine similarity of X and Y in semantic space, mapped by DSSM parameterized by $\boldsymbol{\theta}$
- $\Delta = \text{sim}_{\boldsymbol{\theta}}(X, Y^+) - \text{sim}_{\boldsymbol{\theta}}(X, Y^-)$
  - We want to maximize Δ
- $Loss(\Delta; \boldsymbol{\theta}) = \log(1 + \exp(-\gamma\Delta))$
- Optimize $\boldsymbol{\theta}$ using mini-batch SGD on GPU



Mine “labeled” X-Y pairs from search logs

how to deal with stuffy nose?	⟷	NO CLICK
stuffy nose treatment	⟷	NO CLICK
cold home remedies	⟷	http://www.agelessherbs.com/BestHomeRemediesColdFlu.html

[Gao+ 10]

Mine "labeled" X-Y pairs from search logs

how to deal with stuffy nose?

stuffy nose treatment

cold home remedies

Best Home Remedies for Cold and Flu

Wind Heat External Pathogens

By: Catherine Browne, L.Ac., MH, Dipl. Ac.

In Chinese medicine, colds and flu's are delineated into several different energetic classifications. Here we will outline the different types of cold and flu viruses that you will likely encounter, and then describe the best home remedies for these specific patterns that you can use to treat the cold or influenza virus.

Cold and Flu Basics

The basic pathogenic influences are:

- **Wind**
- **Cold**
- **Heat**
- **Damp**

Wind

Theoretically, wind enters the body through the back of the neck area or nose carrying the pathogen. It first attacks the Lung system (including the sinuses) because the Lung organ system is the most external Yin organ, a thus the most vulnerable to an external invasion. External Wind invasion is marked by acute conditions with a sudden onset of symptoms.

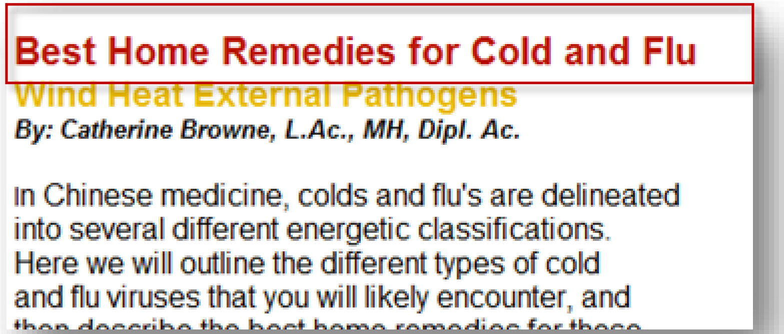
[Gao+ 10]

Mine "labeled" X-Y pairs from search logs

how to deal with stuffy nose?

stuffy nose treatment

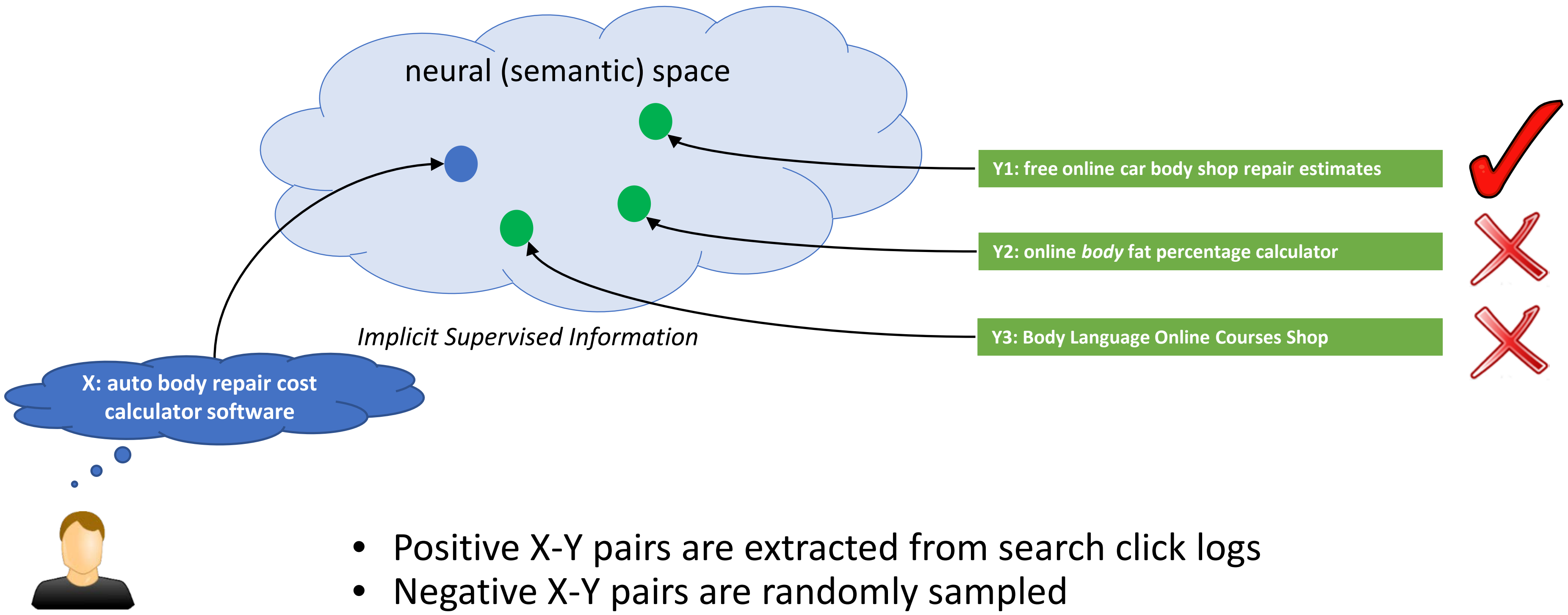
cold home remedies



QUERY (Q)	Title (T)
how to deal with stuffy nose	best home remedies for cold and flu
stuffy nose treatment	best home remedies for cold and flu
cold home remedies	best home remedies for cold and flu
… …	… …
go israel	forums goisrael community
skate at wholesale at pr	wholesale skates southeastern skate supply
breastfeeding nursing blister baby	clogged milk ducts babycenter
thank you teacher song	lyrics for teaching educational children s music
immigration canada lacolle	cbsa office detailed information

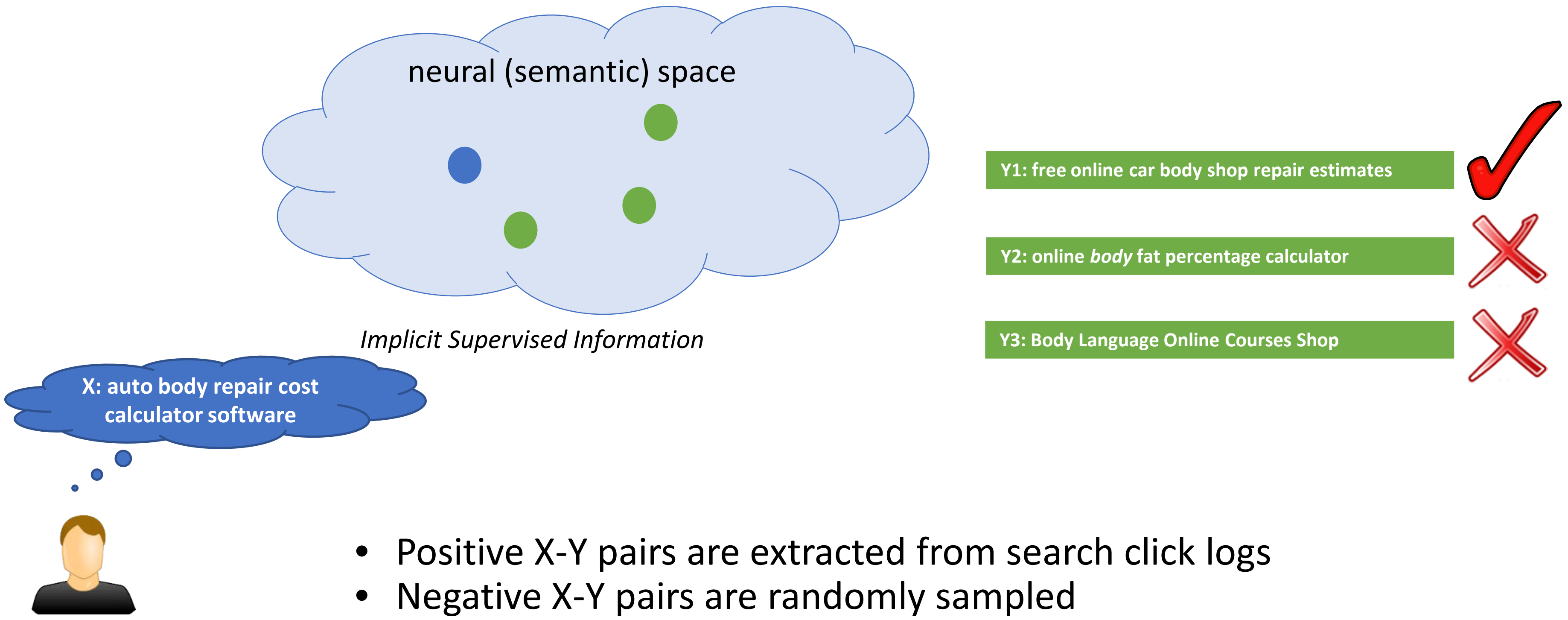
[Gao+ 10]

Learning DSSM from Labeled X-Y Pairs



- Positive X-Y pairs are extracted from search click logs
- Negative X-Y pairs are randomly sampled
- Map X and Y into the same semantic space via deep neural net

Learning DSSM from Labeled X-Y Pairs

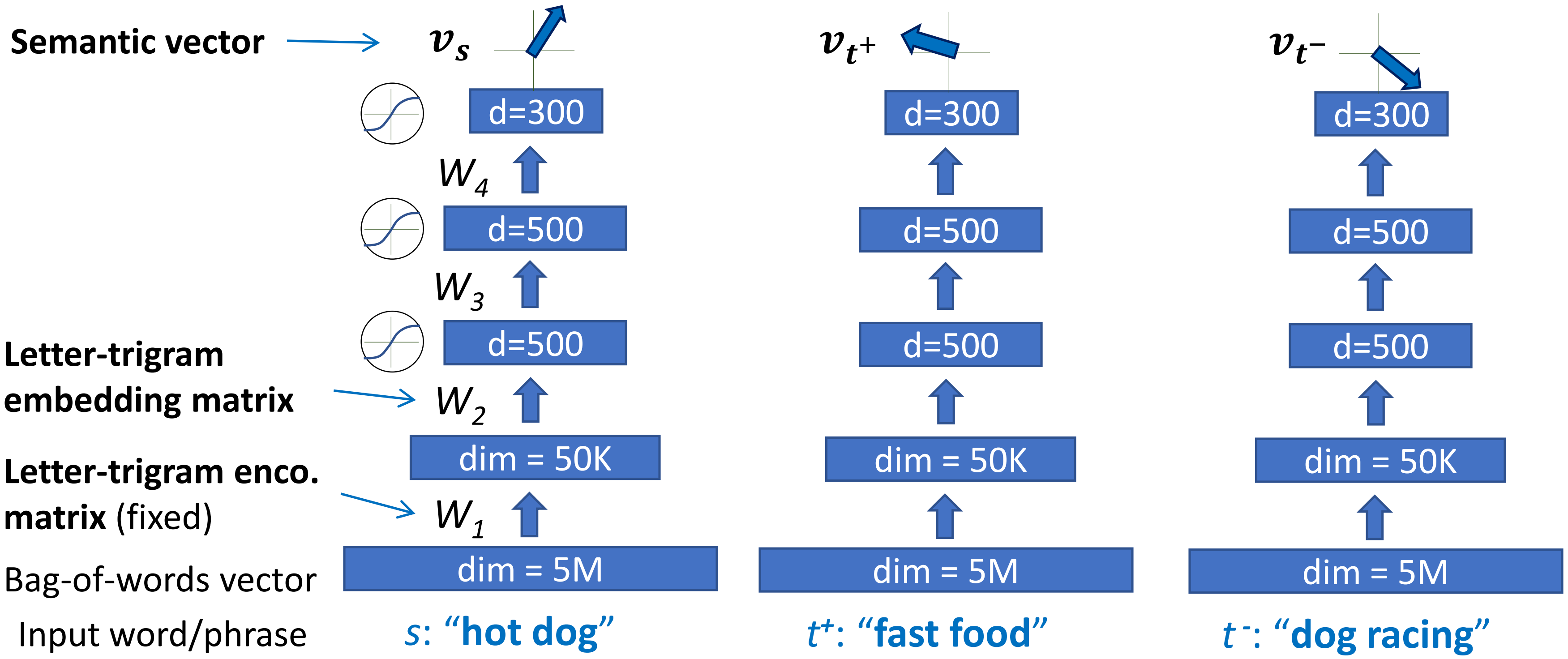


- Positive X-Y pairs are extracted from search click logs
- Negative X-Y pairs are randomly sampled
- Map X and Y into the same semantic space via deep neural net
- Positive Y are closer to X than negative Y in that space

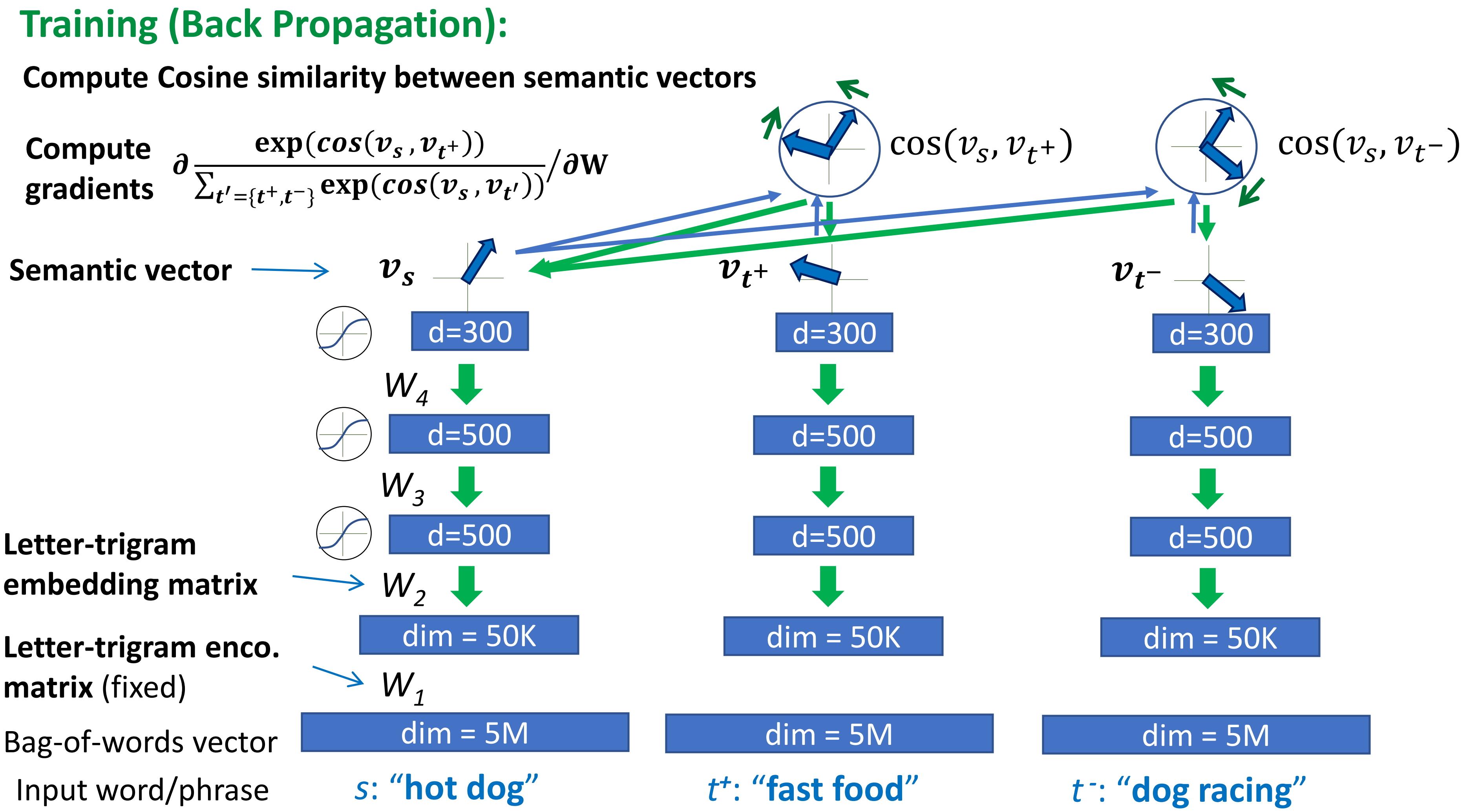
Learning DSSM on X-Y pairs via SGD

Initialization:

Neural networks are initialized with random weights

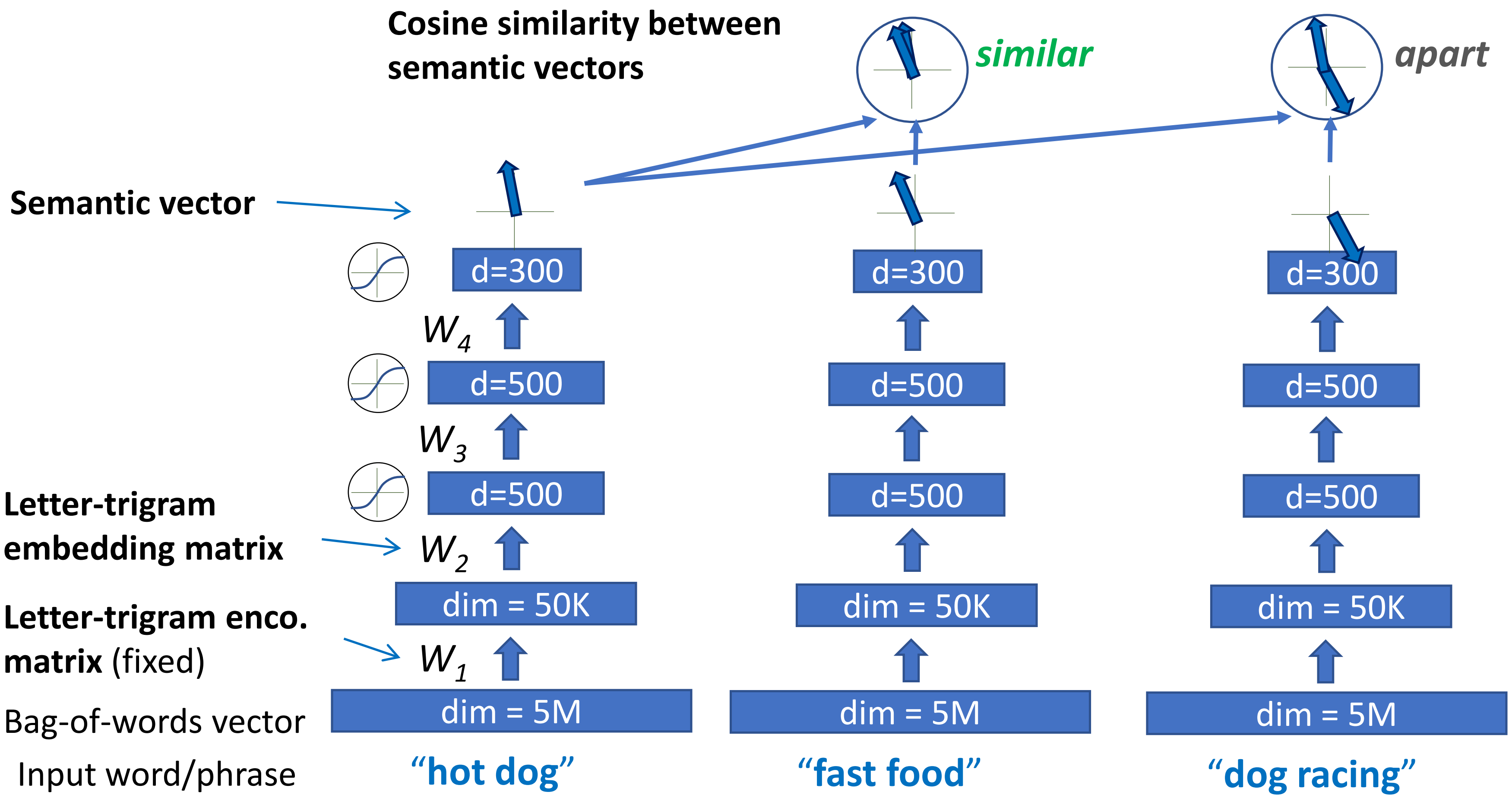


Learning DSSM on X-Y pairs via SGD



Learning DSSM on X-Y pairs via SGD

After training converged:



Evaluation Methodology

- Measurement: NDCG, t-test
- Test set:
  - 12,071 English queries sampled from 1-y log
  - 5-level relevance label for each query-doc pair
- Training data for translation models:
  - 82,834,648 query-title pairs
- Baselines
  - Lexicon matching models: BM25, ULM
  - Translation models [Gao+ 10]
  - Topic models [Hofmann 99; Blei+ 03; Gao+ 11]
  - Deep auto-encoder [Hinton & Salakhutdinov 10]

Translation models for web search

D: best home remedies for cold and flu

Q: how to deal with stuffy nose

- Leverage statistical machine translation (SMT) technologies and infrastructures to improve search relevance
- Model documents and queries as different languages, cast mapping queries to documents as bridging the language gap via translation
- Given a Q, D can be ranked by how likely it is that Q is “translated” from D, $P(\mathrm{Q}|\mathrm{D})$
  - Word translation model
  - Phrase translation model

[Gao+ 10]

Generative Topic Models

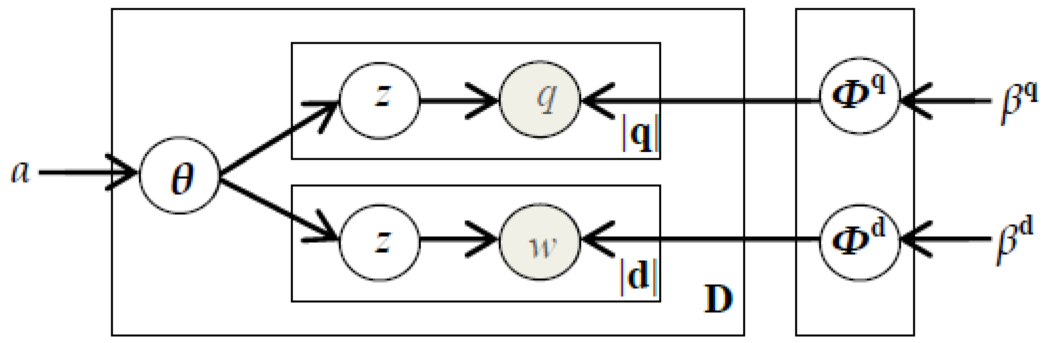
Q: stuffy nose treatment ← D: cold home remedies

Q: stuffy nose treatment ← Topic ← D: cold home remedies

- Probabilistic latent Semantic Analysis (PLSA)
  - $P(\mathrm{Q}|\mathrm{D}) = \prod_{q \in \mathrm{Q}} \sum_{z} P(q|\boldsymbol{\phi}_z) P(z|\mathrm{D}, \boldsymbol{\theta})$
  - D is assigned a single most likely topic vector
  - Q is generated from the topic vectors
- Latent Dirichlet Allocation (LDA) generalizes PLSA
  - a posterior distribution over topic vectors is used
  - PLSA = LDA with MAP inference

[Hofmann 99; Blei+ 03]

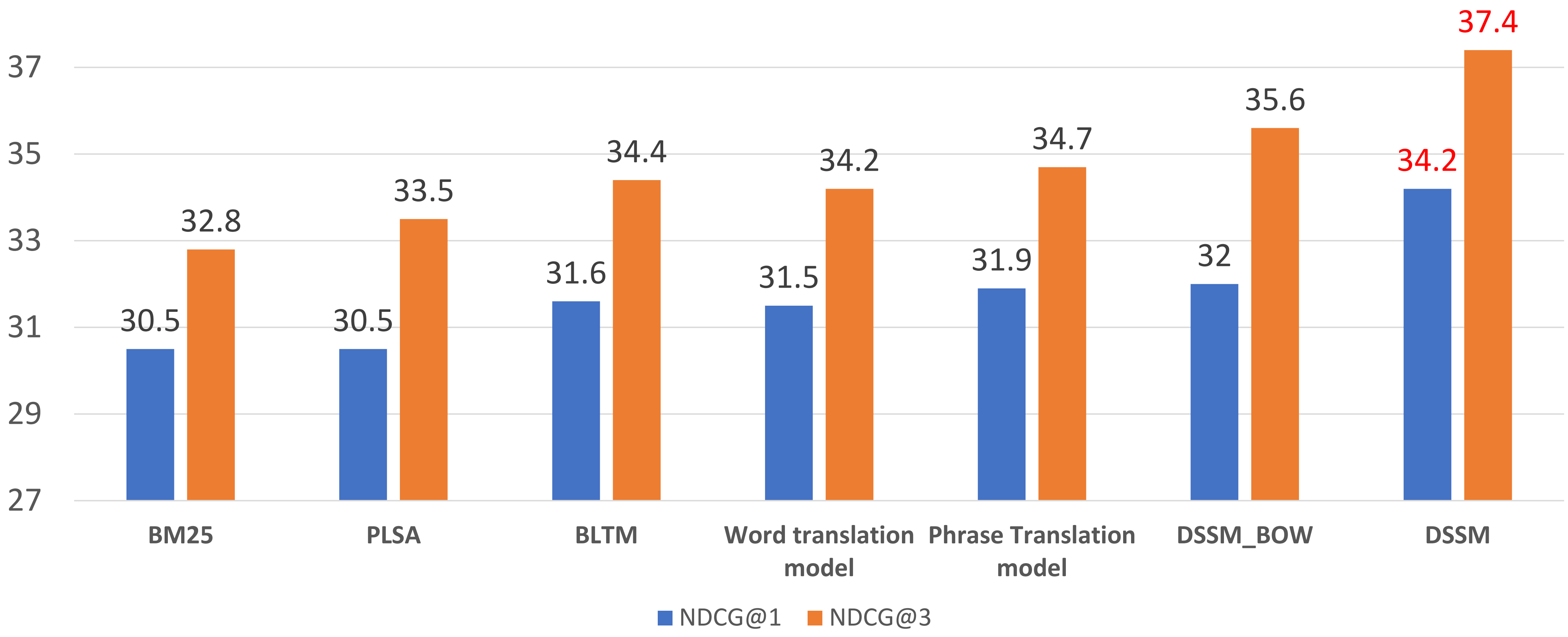
Bilingual topic model for web search



- For each topic *z*: $\left(\boldsymbol{\phi}_z^{\mathrm{Q}}, \boldsymbol{\phi}_z^{\mathrm{D}}\right) \sim \mathrm{Dir}(\boldsymbol{\beta})$
- For each Q-D pair: $\boldsymbol{\theta} \sim \mathrm{Dir}(\boldsymbol{\alpha})$
- Each q is generated by $z \sim \boldsymbol{\theta}$ and $q \sim \boldsymbol{\phi}_z^{\mathrm{Q}}$
- Each w is generated by $z \sim \boldsymbol{\theta}$ and $w \sim \boldsymbol{\phi}_z^{\mathrm{D}}$

[Gao+ 11]

Web doc ranking results



Summary

- Map the queries and documents into the same latent semantic space
- Doc ranking score is the cosine distance of Q/D vectors in that space
- DSSM outperforms all the competing models
- The learning DSSM vectors capture semantic similarities and relations btw words

DSSM for recommendation

- Two interestingness tasks for recommendation
- Modeling interestingness via DSSM
- Training data acquisition
- Evaluation
- Summary

[Gao+ 14b]

Two Tasks of Modeling Interestingness

- Automatic highlighting
  - Highlight the key phrases which represent the entities (person/loc/org) that interest a user when reading a document
  - Doc semantics influences what is perceived as interesting to the user
  - e.g., article about movie → articles about an actor/character

- Entity linking
  - Given the highlighted key phrases, recommend new, interesting documents by searching the Web for supplementary information about the entities
  - A key phrase may refer to different entities; need to use the contextual information to disambiguate

The Einstein Theory of Relativity

(1) The perihelion of Mercury shows a discrepancy which has long puzzled astronomers. This discrepancy is fully accounted for by Einstein. At the time when he published his theory, this was its only experimental verification.

(2) Modern physicists were willing to suppose that light might be subject to gravitation—i.e., that a ray of light passing near a great mass like the sun might be deflected to the extent to which a particle moving with the same velocity would be deflected according to the orthodox theory of gravitation. But Einstein's theory required that the light should be deflected just twice as much as this. The matter could only be tested during an eclipse among a number of bright stars. Fortunately a peculiarly favourable eclipse occurred last year. The results of the observations

The Einstein Theory of Relativity

(1) The perihelion of Mercury shows a discrepancy which has long puzzled astronomers. This discrepancy is fully accounted for by Einstein. At the time when he published his theory, this was its only experimental verification.

(2) Modern physicists were willing to suppose that light might be subject to gravitation—i.e., that a ray of light passing near a great mass like the sun might be deflected to the extent to which a particle moving with the same velocity would be deflected according to the orthodox theory of gravitation. But Einstein's theory required that the light should be deflected just twice as much as this. The matter could only be tested during an eclipse among a number of bright stars. Fortunately a peculiarly favourable eclipse occurred last year. The results of the observations

The Einstein Theory of Relativity

(1) The perihelion of Mercury shows a discrepancy which has long puzzled astronomers. This discrepancy is fully accounted for by Einstein. At the time when he published his theory, this was its only experimental verification.

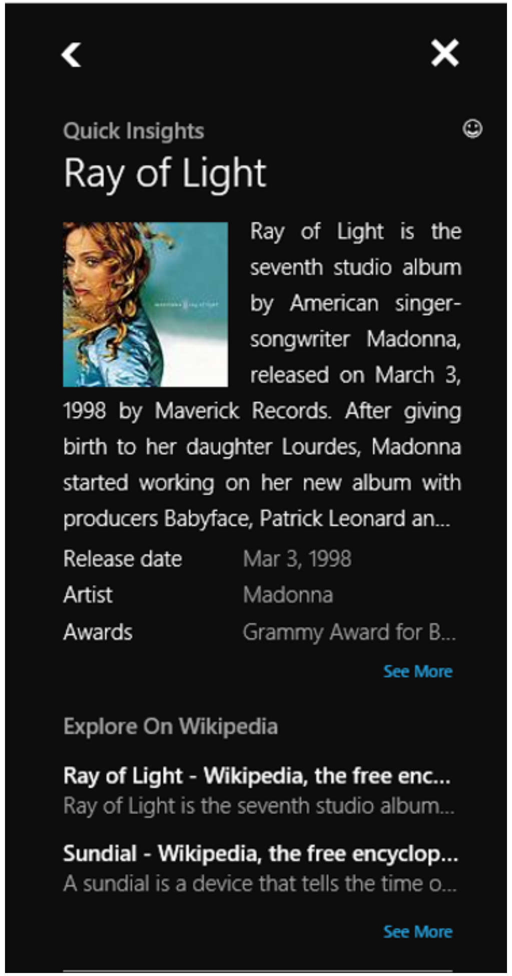
(2) Modern physicists were willing to suppose that light might be subject to gravitation—i.e., that a ray of light passing near a great mass like the sun might be deflected to the extent to which a particle moving with the same velocity would be deflected according to the orthodox theory of gravitation. But Einstein's theory required that the light should be deflected just twice as much as this. The matter could only be tested during an eclipse among a number of bright stars. Fortunately a peculiarly favourable eclipse occurred last year. The results of the observations

The Einstein Theory of Relativity

(1) The perihelion of Mercury shows a discrepancy which has long puzzled astronomers. This discrepancy is fully accounted for by Einstein. At the time when he published his theory, this was its only experimental verification.

(2) Modern physicists were willing to suppose that light might be subject to gravitation—i.e., that a ray of light passing near a great mass like the sun might be deflected to the extent to which a particle moving with the same velocity would be deflected according to the orthodox theory of gravitation. But Einstein's theory required that the light should be deflected just twice as much as this. The matter could only be tested during an eclipse among a number of bright stars. Fortunately a peculiarly favourable eclipse occurred last year. The results of the observations

Entity



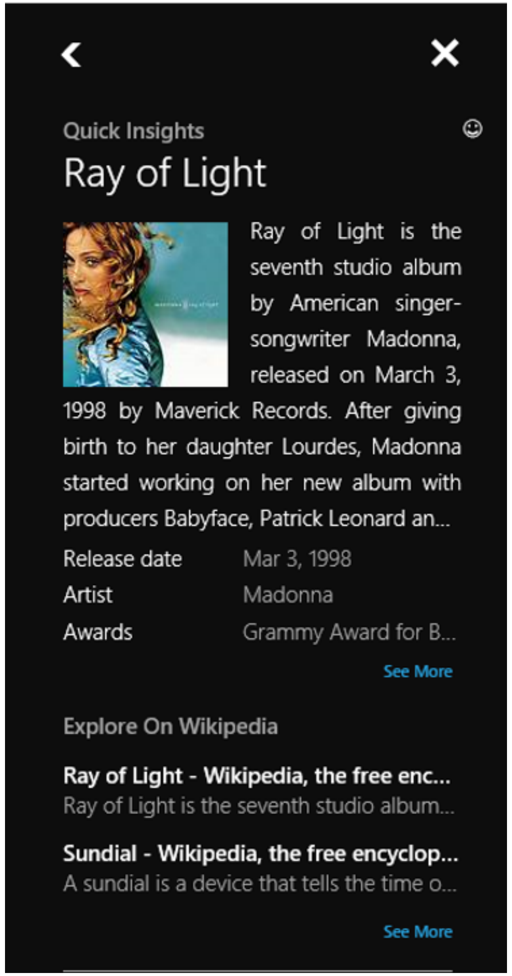
The Einstein Theory of Relativity

Context

(1) The perihelion of Mercury shows a discrepancy which has long puzzled astronomers. This discrepancy is fully accounted for by Einstein. At the time when he published his theory, this was its only experimental verification.

(2) Modern physicists were willing to suppose that light might be subject to gravitation—i.e., that a ray of light passing near a great mass like the sun might be deflected to the extent to which a particle moving with the same velocity would be deflected according to the orthodox theory of gravitation. But Einstein's theory required that the light should be deflected just twice as much as this. The matter could only be tested during an eclipse among a number of bright stars. Fortunately a peculiarly favourable eclipse occurred last year. The results of the observations

Entity



The Einstein Theory of Relativity

Context

(1) The perihelion of Mercury shows a discrepancy which has long puzzled astronomers. This discrepancy is fully accounted for by Einstein. At the time when he published his theory, this was its only experimental verification.

(2) Modern physicists were willing to suppose that light might be subject to gravitation—i.e., that a ray of light passing near a great mass like the sun might be deflected to the extent to which a particle moving with the same velocity would be deflected according to the orthodox theory of gravitation. But Einstein's theory required that the light should be deflected just twice as much as this. The matter could only be tested during an eclipse among a number of bright stars. Fortunately a peculiarly favourable eclipse occurred last year. The results of the observations

Entity

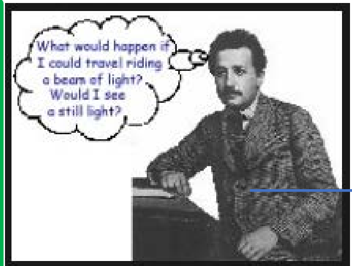
DSSM for Modeling Interestingness

Context

Key phrase

(1) The perihelion of Mercury shows a discrepancy which has long puzzled astronomers. This discrepancy is fully accounted for by Einstein. At the time when he published his theory, this was its only experimental verification.

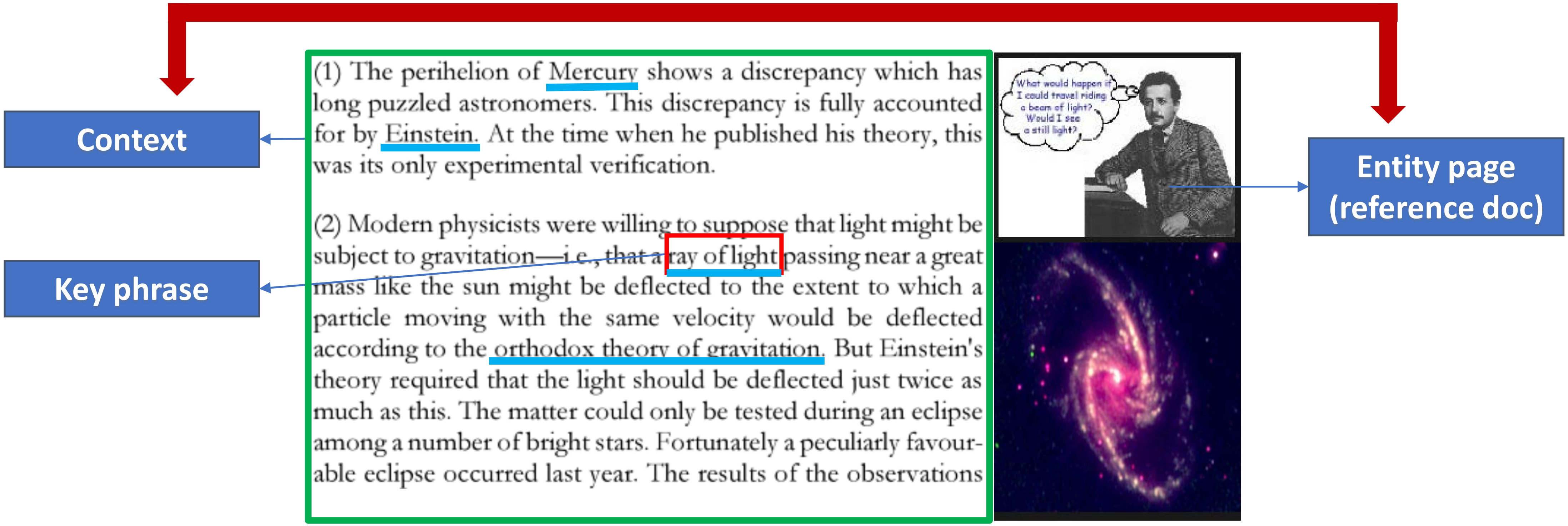
(2) Modern physicists were willing to suppose that light might be subject to gravitation—i.e., that a ray of light passing near a great mass like the sun might be deflected to the extent to which a particle moving with the same velocity would be deflected according to the orthodox theory of gravitation. But Einstein's theory required that the light should be deflected just twice as much as this. The matter could only be tested during an eclipse among a number of bright stars. Fortunately a peculiarly favourable eclipse occurred last year. The results of the observations



Entity page (reference doc)

Tasks	X (source text)	Y (target text)
Automatic highlighting	***Doc in reading***	***Key phrases to be highlighted***
Entity linking	*Entity mention*	*Entity and its corresponding (wiki) page*

DSSM for Modeling Interestingness

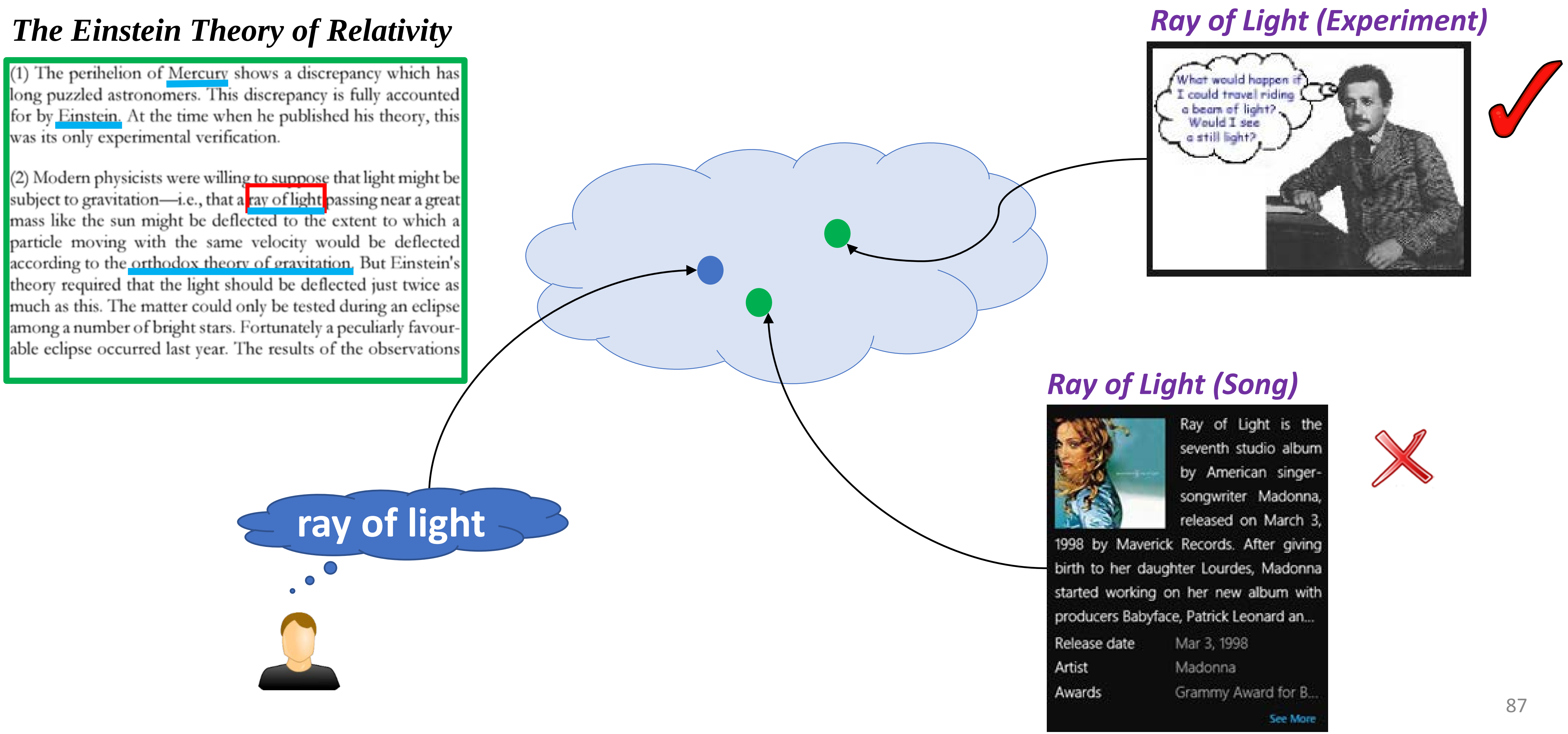


(1) The perihelion of Mercury shows a discrepancy which has long puzzled astronomers. This discrepancy is fully accounted for by Einstein. At the time when he published his theory, this was its only experimental verification.

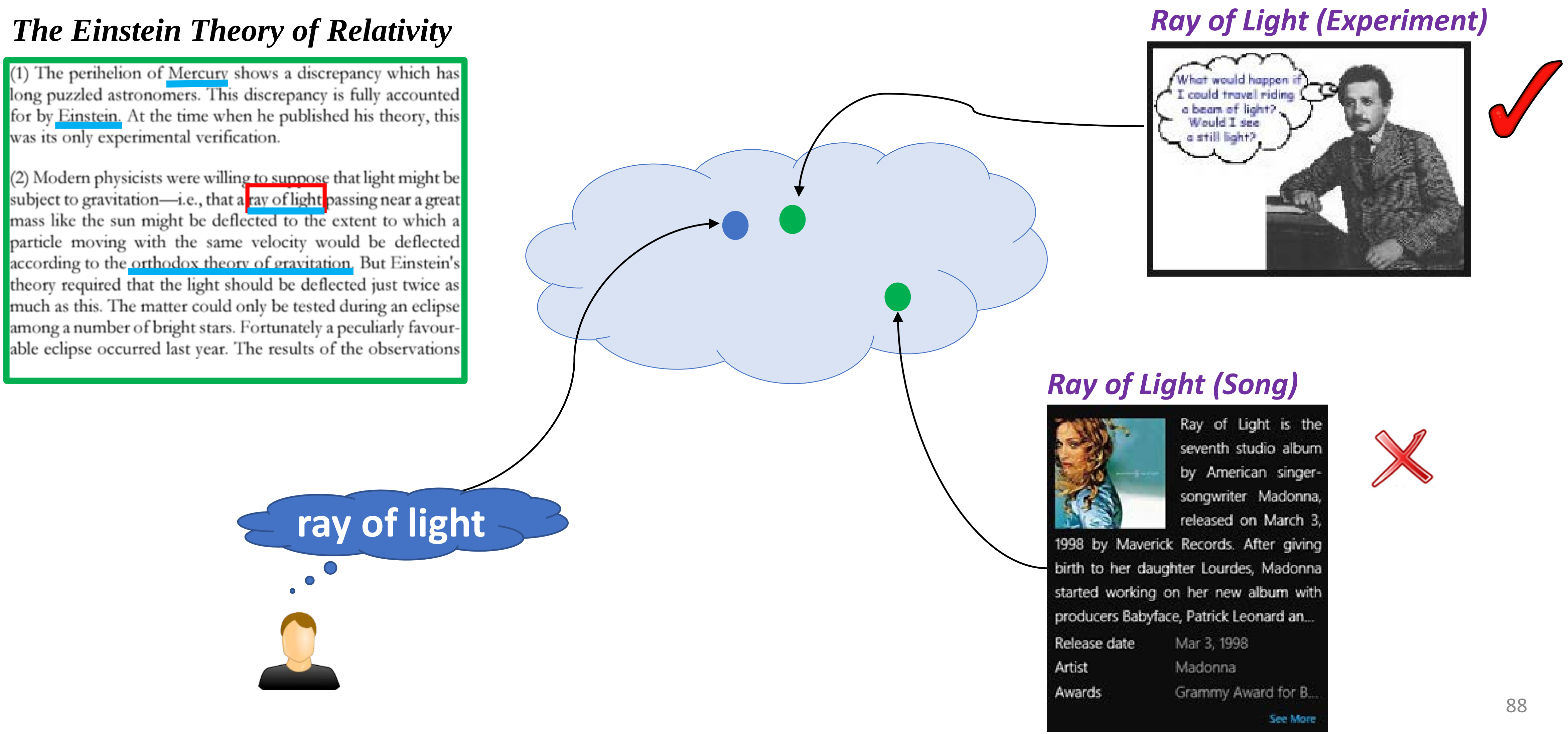
(2) Modern physicists were willing to suppose that light might be subject to gravitation—i.e., that a ray of light passing near a great mass like the sun might be deflected to the extent to which a particle moving with the same velocity would be deflected according to the orthodox theory of gravitation. But Einstein's theory required that the light should be deflected just twice as much as this. The matter could only be tested during an eclipse among a number of bright stars. Fortunately a peculiarly favourable eclipse occurred last year. The results of the observations

Tasks	*X (source text)*	*Y (target text)*
Automatic highlighting	*Doc in reading*	*Key phrases to be highlighted*
Entity linking	***Entity mention***	***Entity and its corresponding (wiki) page***

Learning DSSM from Labeled X-Y Pairs



Learning DSSM from Labeled X-Y Pairs



DSSM for recommendation

- Two interestingness tasks for recommendation
- Modeling interestingness via DSSM
- Training data acquisition
- Evaluation
- Summary

Extract Labeled Pairs from Web Browsing Logs
Automatic Highlighting

- When reading a page P, the user *clicks* a hyperlink H

P

http://runningmoron.blogspot.in/

...

I spent a lot of time finding music that was motivating and that I'd also want to listen to through my phone. I could find none. None! I wound up downloading three Metallica songs, a Judas Priest song and one from Bush.

...

H

- (text in P, anchor text of H)

Extract Labeled Pairs from Web Browsing Logs
Entity Linking

- When a hyperlink H points to a Wikipedia P'

http://runningmoron.blogspot.in/

…
I spent a lot of time finding music that was motivating and that I'd also want to listen to through my phone. I could find none. None! I wound up downloading three Metallica songs, a [Judas Priest]song and one from [Bush].
…

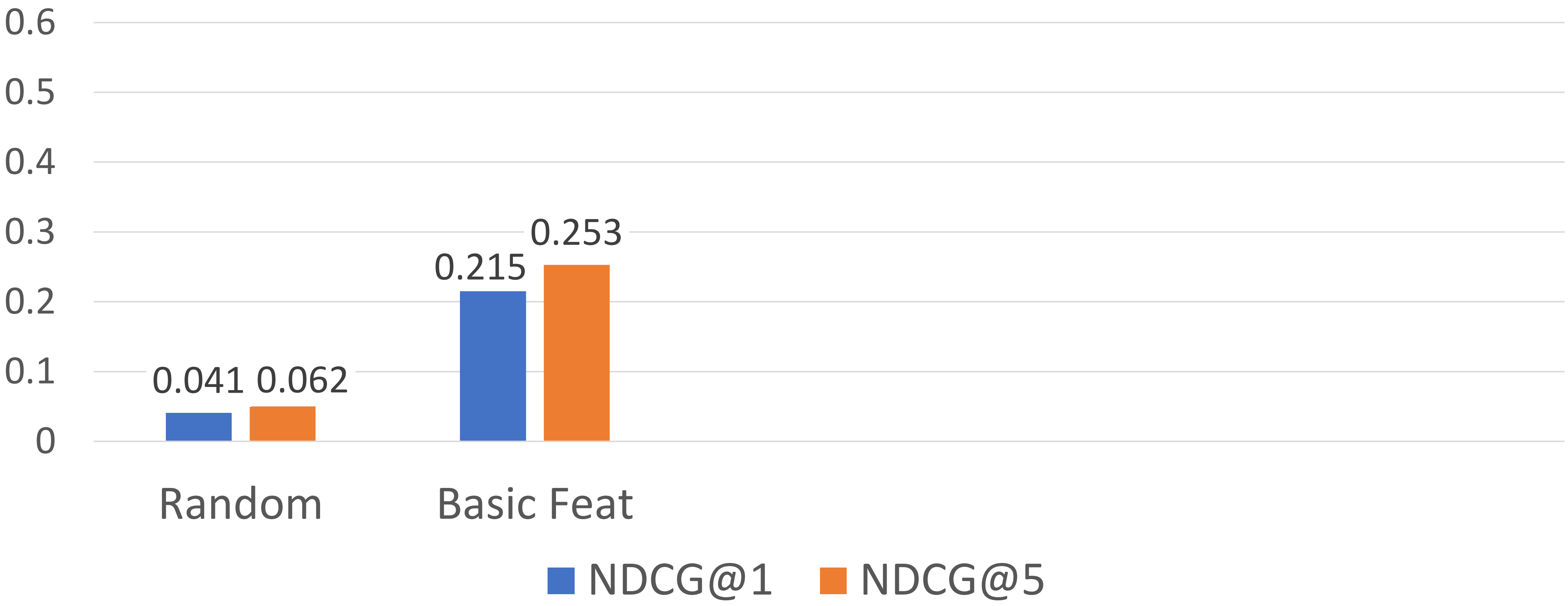
http://en.wikipedia.org/wiki/Bush_(band)

- (anchor text of H & surrounding words, text in P')

Automatic Highlighting: Settings

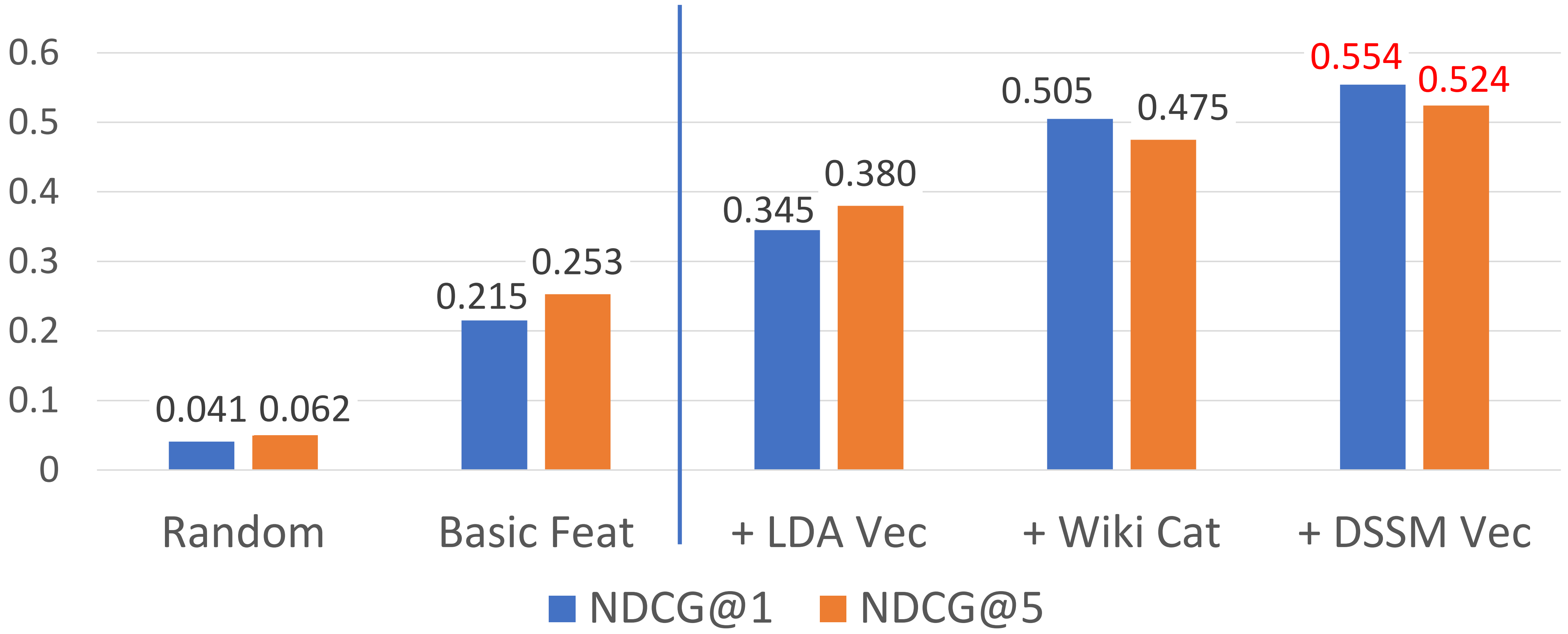
- Simulation
  - Use a set of anchors as candidate key phrases to be highlighted
  - Gold standard rank of key phrases – determined by # user clicks
  - Model picks top-k keywords from the candidates
  - Evaluation metric: NDCG
- Data
  - 18 million occurrences of user clicks from a Wiki page to another, collected from 1-year Web browsing logs
  - 60/20/20 split for training/validation/evaluation

Automatic Highlighting Results: Baselines



- Random: Random baseline
- Basic Feat: Boosted decision tree learner with document features, such as anchor position, freq. of anchor, anchor density, etc.

Automatic Highlighting Results: Semantic Features

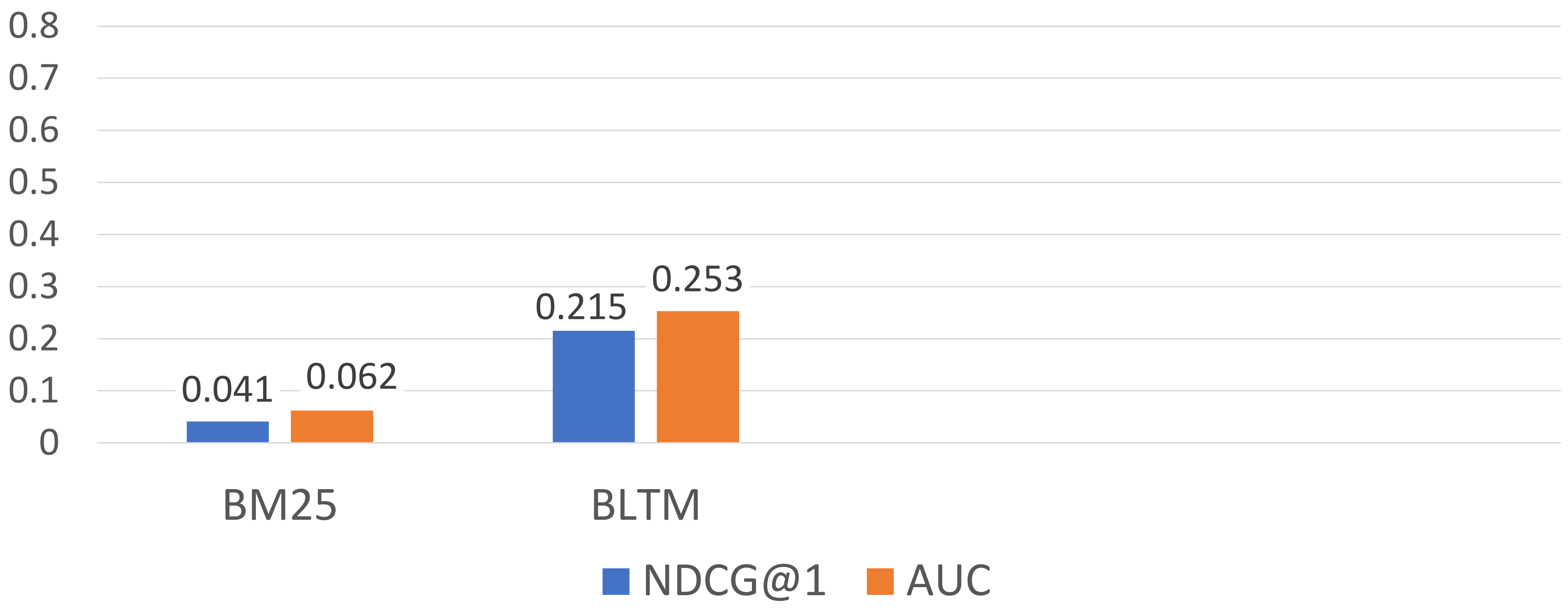


- + LDA Vec: Basic + Topic model (LDA) vectors [Gamon+ 2013]
- + Wiki Cat: Basic + Wikipedia categories (do not apply to general documents)
- + DSSM Vec: Basic + DSSM vectors

Entity Linking: Settings

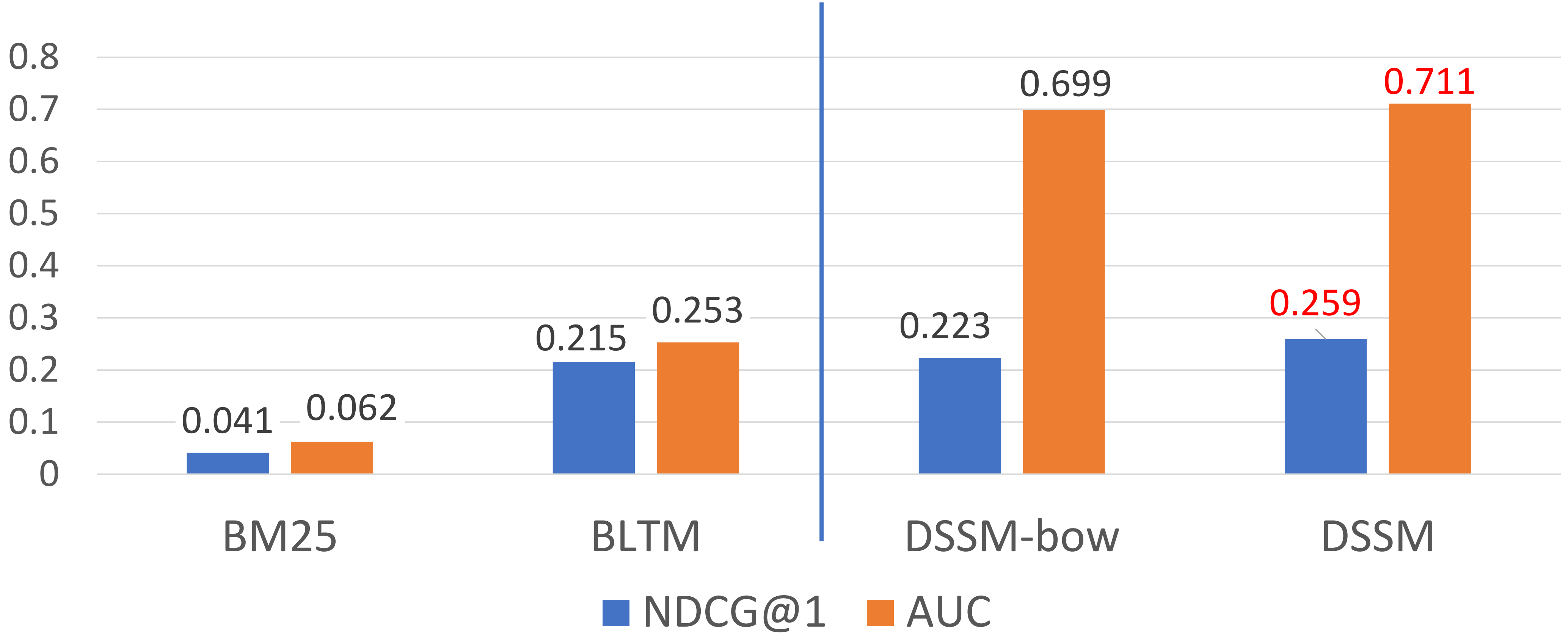
- Training/validation data: same as in *automatic highlighting*
- Evaluation data
  - Sample 10k Web documents as the source documents
  - Use named entities in the doc as query; retain up to 100 returned documents as target documents
  - Manually label whether each target document is a good page describing the entity
  - 870k labeled pairs in total
- Evaluation metric: NDCG and AUC

Contextual Entity Search Results: Baselines



- BM25: The classical document model in IR [Robertson+ 1994]
- BLTM: Bilingual Topic Model [Gao+ 2011]

Contextual Entity Search Results: DSSM



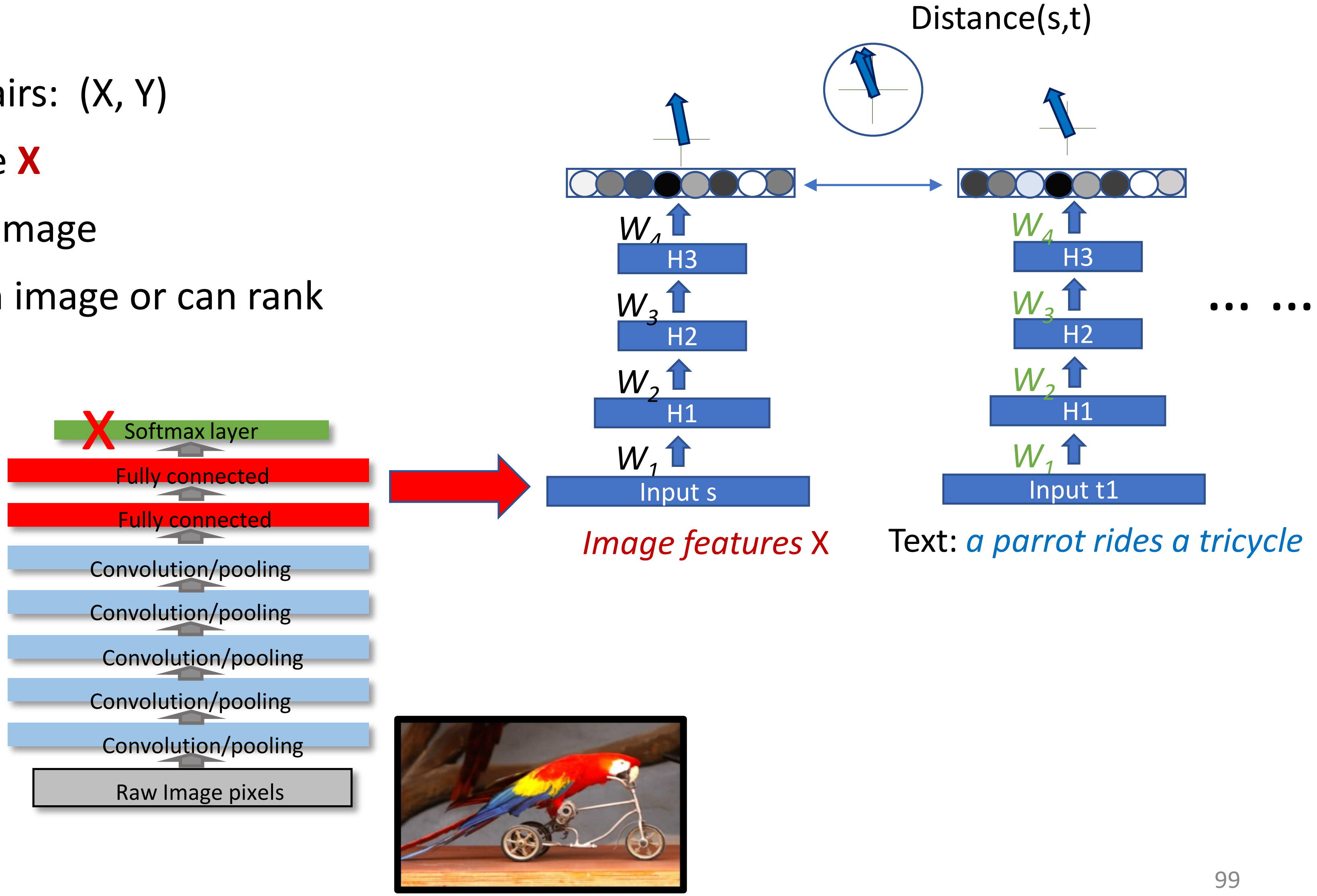
- DSSM-bow: DSSM without convolutional layer and max-pooling structure
- DSSM outperforms classic doc model and state-of-the-art topic model

Summary

- Extract labeled pairs from Web browsing logs
- DSSM outperforms state-of-the-art topic models
- DSSM learned semantic features outperform the thousands of features coming from the manually assigned semantic labels

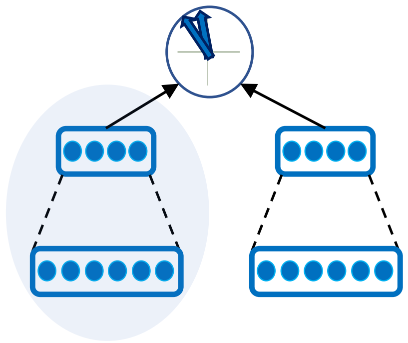
Go beyond text:
DSSM for multi-modal representation learning

- Recall DSSM for text input pairs: (X, Y)
- Now: replace text X by image **X**
- Using DNN/CNN features of image
- Can rank/generate text given image or can rank images given text.

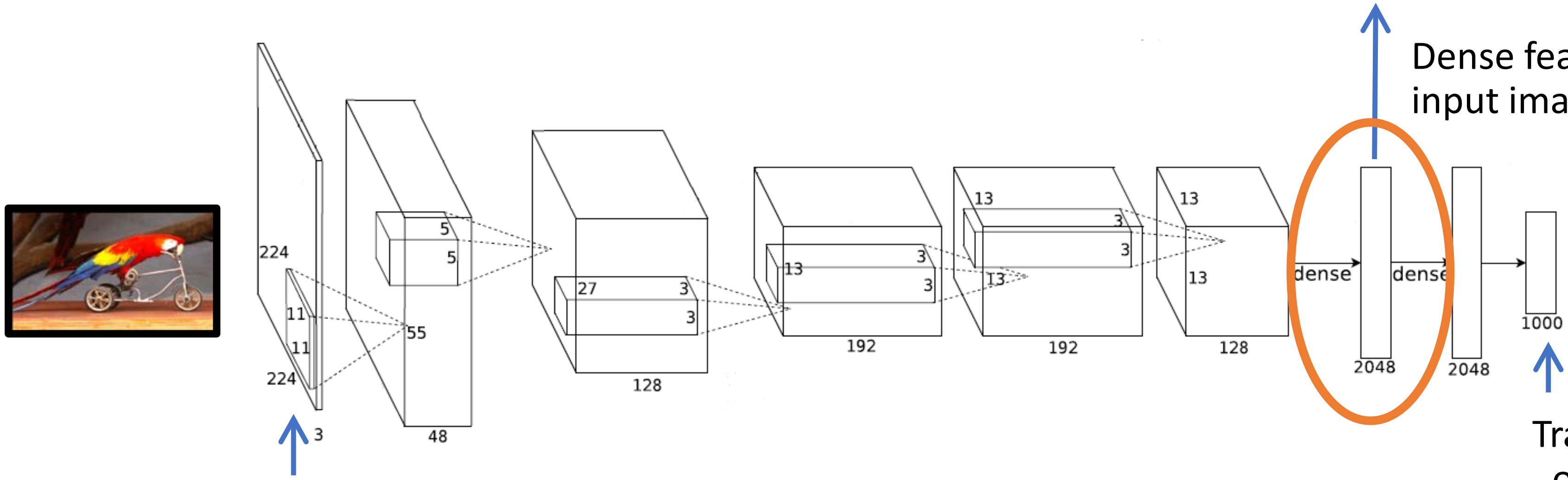


The convolutional network at the image side

Feed the pre-trained image feature vector into the image side of the DSSM



Dense feature vector for input image



Raw pixels from input box

Trained to predict object in image

Pretrained from ImageNet [Krizhevsky+ 12]

The convolutional network at the caption side

Models fine-grained structural language information in the caption

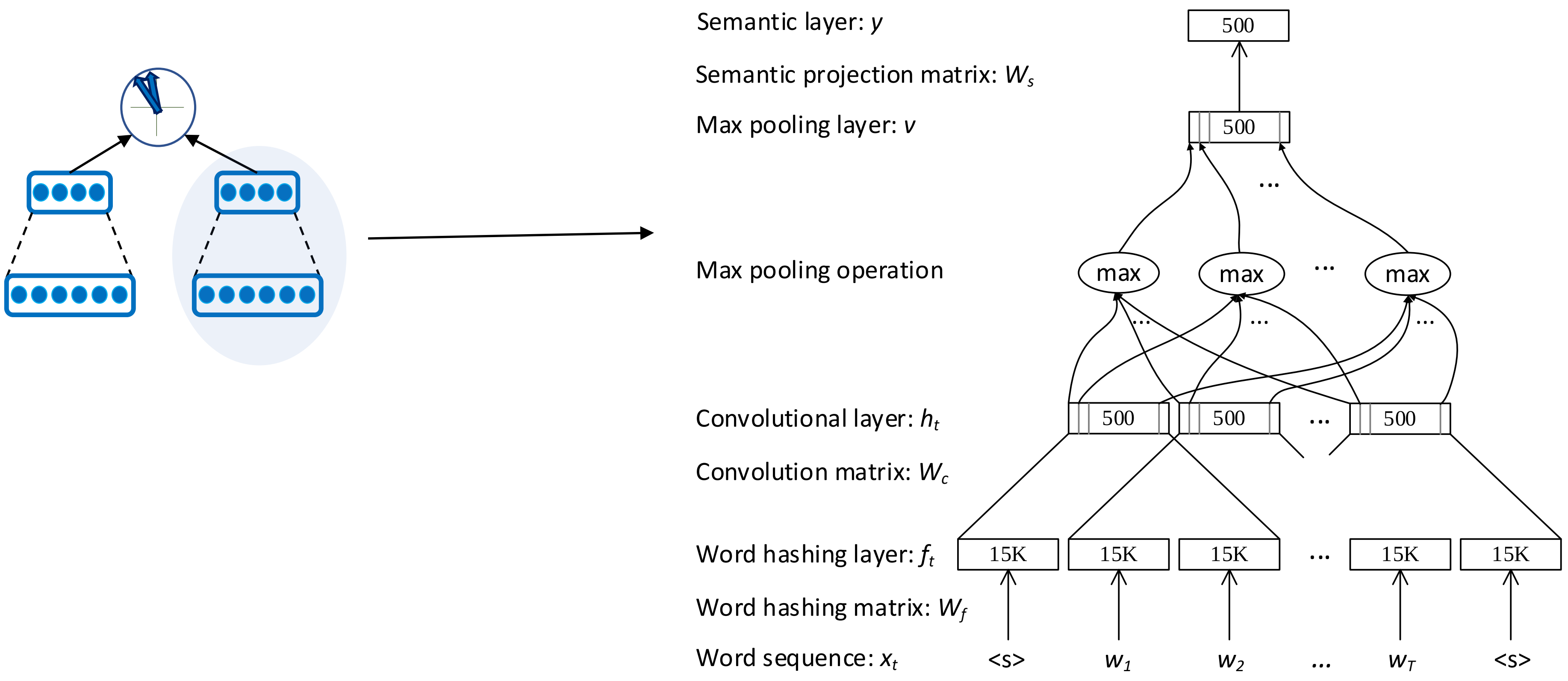
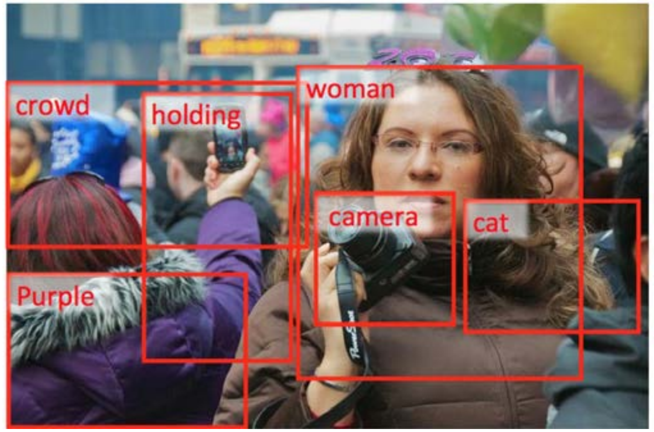


Image captioning

- Why important?
  - Build intelligent machines that understand the semantics in complex scenes
  - Language is a regulator for *understanding as human do*.
- Why difficult?
  - Need to detect multiple objects in arbitrary regions, and capture the complex semantics among these objects.
- What different (e.g., vs. ImageNet / object categorization)?
  - Capturing the salient, coherent semantic information embedded in a picture.



A woman holding a camera in a crowd.

The MSR system

Understand the image stage by stage:

Image word detection

Deep-learned features, applied to likely items in the image, trained to produce words in captions

Language generation

Maxent language model, trained on caption, conditional on words detected from the image

Global semantic re-ranking

Hypothetical captions re-ranked by deep-learned multi-modal similarity model looking at the entire image

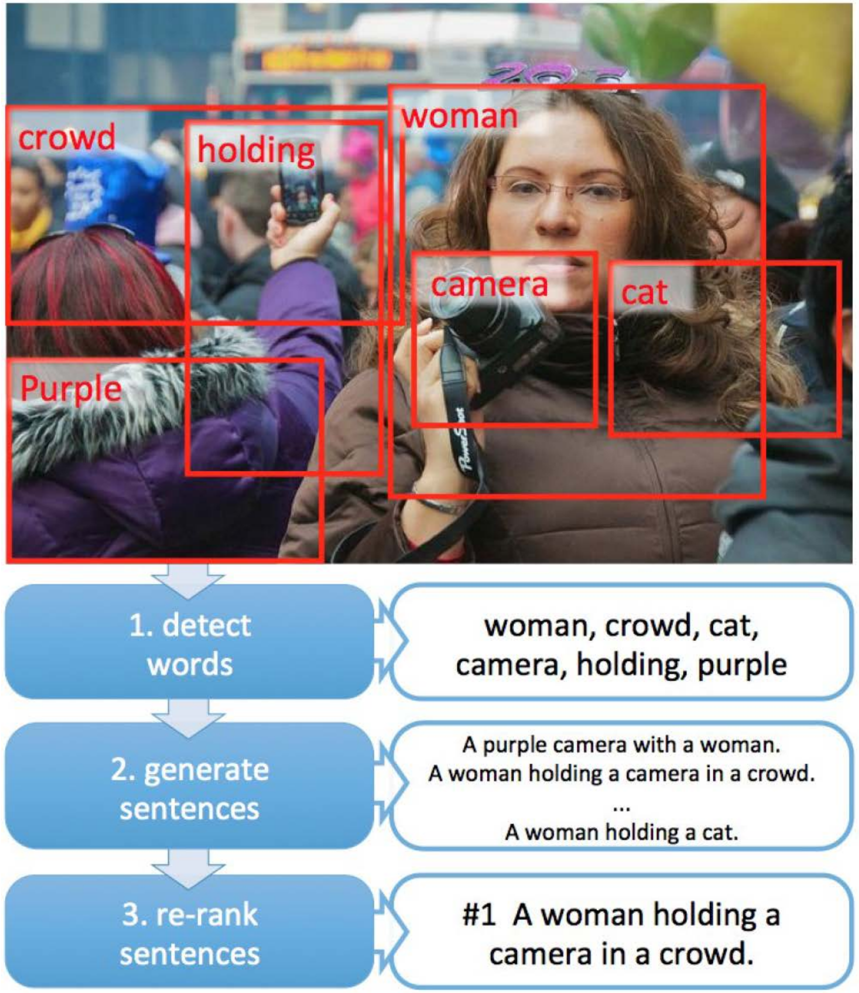


Figure 1. An illustrative example of our pipeline.

Fang, Gupta, Iandola, Srivastava, Deng, Dollar, Gao, He, Mitchell, Platt, Zitnick, Zweig, "From Captions to Visual Concepts and Back," CVPR, June 2015

The MS COCO Benchmark

Microsoft COCO
Common Objects in Context

What is Microsoft COCO?

Microsoft COCO is a new image recognition, segmentation, and captioning dataset. Microsoft COCO has several features:

- Object segmentation
- Recognition in Context
- Multiple objects per image
- More than 300,000 images
- More than 2 Million instances
- 80 object categories
- 5 captions per image

Collaborators

Tsung-Yi Lin Cornell Tech

Michael Maire TTI Chicago

Serge Belongie Cornell Tech

CORNELL NYCTECH

Caltech

The man at bat readies to swing at the pitch while the umpire looks on.

A large bus sitting next to a very tall building.

Results

System	PPLX	BLEU	METEOR	$\approx$human	$>$human	$\geq$human
1. Unconditioned	24.1	1.2%	6.8%			
2. Shuffled Human	–	1.7%	7.3%			
3. Baseline	20.9	16.9%	18.9%	9.9% (±1.5%)	2.4% (±0.8%)	12.3% (±1.6%)
4. Baseline+Score	20.2	20.1%	20.5%	16.9% (±2.0%)	3.9% (±1.0%)	20.8% (±2.2%)
5. Baseline+Score+DMSM	20.2	21.1%	20.7%	18.7% (±2.1%)	4.6% (±1.1%)	23.3% (±2.3%)
6. Baseline+Score+DMSM+ft	19.2	23.3%	22.2%	–	–	–
7. VGG+Score+ft	18.1	23.6%	22.8%	–	–	–
8. VGG+Score+DMSM+ft	18.1	25.7%	23.6%	26.2% (±2.1%)	7.8% (±1.3%)	**34.0%** (±2.5%)
Human-written captions	–	19.3%	24.1%			

* we use 4 references when measuring BLEU and METEOR, while the official COCO eval server uses 5 references.

DMSM (i.e., DSSM) gives additional 2.1 pt BLEU (8 vs. 7) over a strong system.

Compared to human, our system is better or equal 34% of the time.

Related work

- Use CNN to generate a whole-image feature vector,
- then feed it into a LSTM language model to generate the caption.

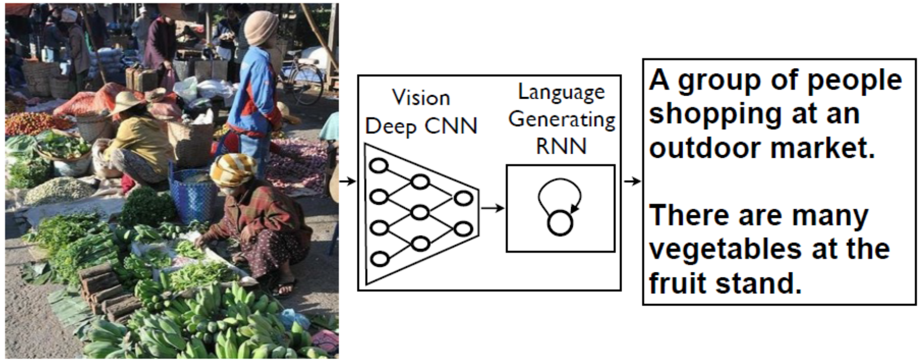


Figure 1. NIC, our model, is based end-to-end on a neural network consisting of a vision CNN followed by a language generating RNN. It generates complete sentences in natural language from an input image, as shown on the example above.

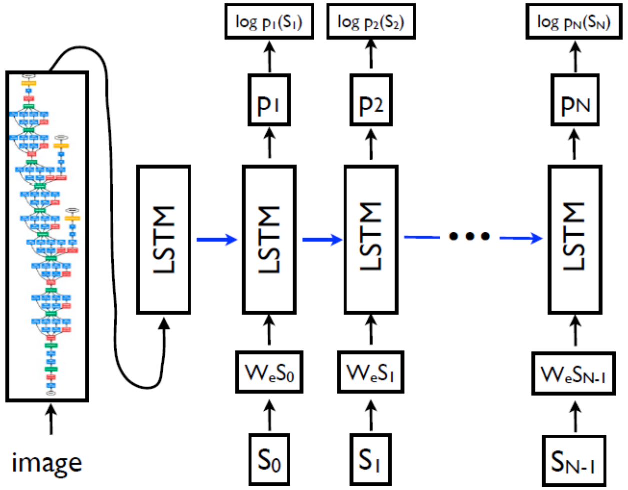


Figure 3. LSTM model combined with a CNN image embedder (as defined in [30]) and word embeddings. The unrolled connections between the LSTM memories are in blue and they correspond to the recurrent connections in Figure 2. All LSTMs share the same parameters.

Vinyals, Toshev, Bengio, Erhan, "**Show and Tell: A Neural Image Caption Generator**", CVPR 2015

Machine-generated (but turker prefered)	a group of motorcycles parked next to a motorcycle
Human-annotated (but turker not prefered)	two girls wearing are wearing short skirts and one of them sits on a motorcycle while the other stands nearby

Machine-generated (but turker prefered)	a wom food
Human-annotated (but turker not prefered)	woman kitchen

next to a
next to a

Machine-generated (but turker prefered)	a clock tower in the middle of the street
Human-annotated (but turker not prefered)	a statue with a clock on it near a parking lot

DSSM helps pick the global semantically matching caption for a given image

Baseline: a large jetliner sitting on top of a stop sign at an intersection on a city street

w/ m-DSSM: a stop light on a city street

Baseline: a clock tower in front of a building

w/ m-DSSM: a clock tower in the middle of the street

Baseline: a red brick building

w/ m-DSSM: a living room filled with furniture and a flat screen tv sitting on top of a brick building

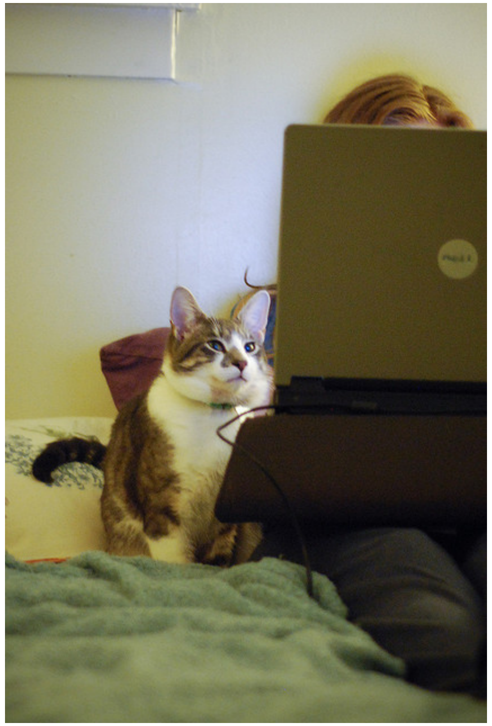
Baseline: a large jetliner sitting on top of a table

w/ m-DSSM: a display in a grocery store filled with lots of food on a table

DSSM helps pick the global semantically matching caption for a given image

Baseline: a young man riding a skateboard down a street holding a tennis racquet on a tennis court

w/ m-DSSM: a man riding a skateboard down a street



Baseline: a cat sitting on a table

w/ m-DSSM: a cat sitting on top of a bed

Baseline: a group of people standing in a kitchen

w/ m-DSSM: a group of people posing for a picture

Baseline: two elephants standing next to a baby elephant walking behind a fence

w/ m-DSSM: a baby elephant standing next to a fence

Interpretability

Our system not only generates the caption, but can also interpret it.

Interpretability

a **baseball**

Interpretability

player (1.00)

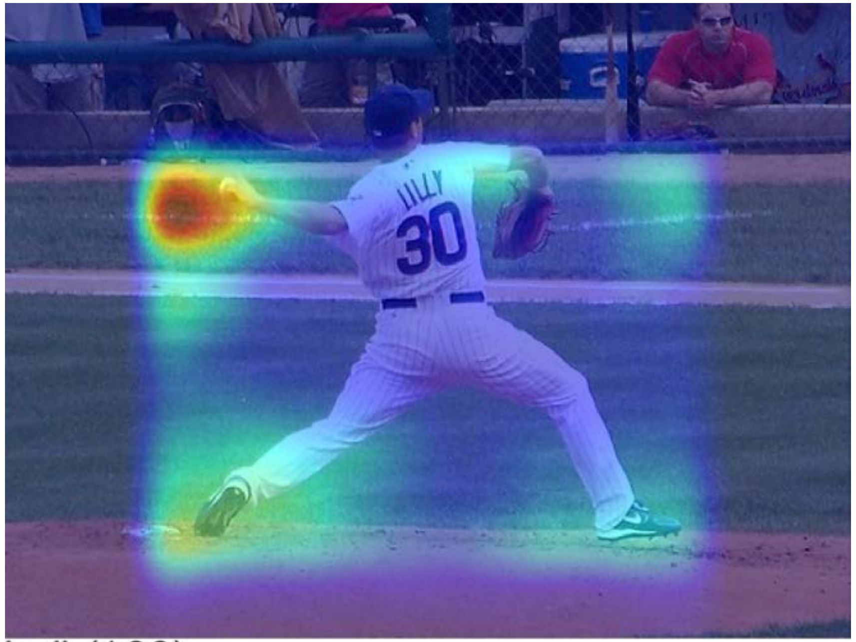
a baseball **player**

Interpretability

throwing (0.86)

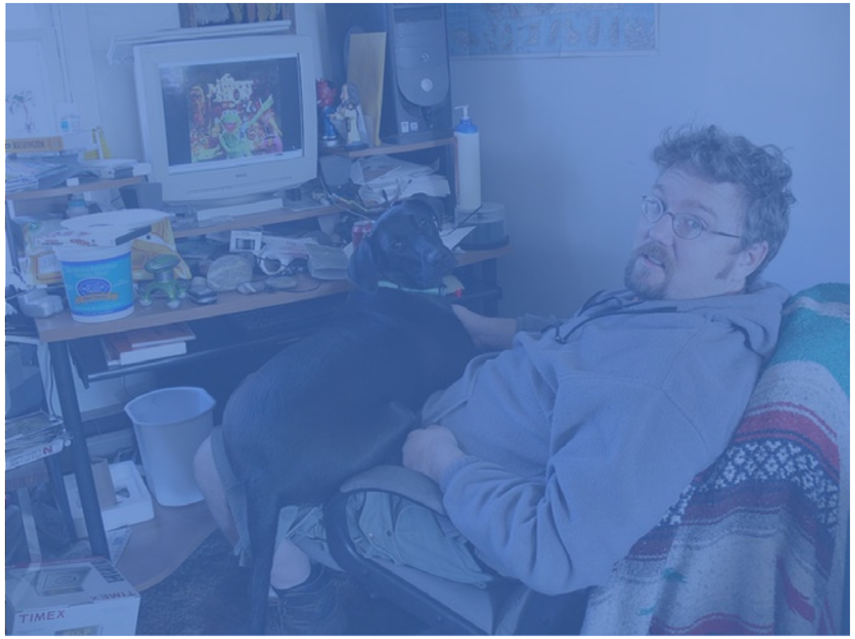
a baseball player **throwing**

Interpretability



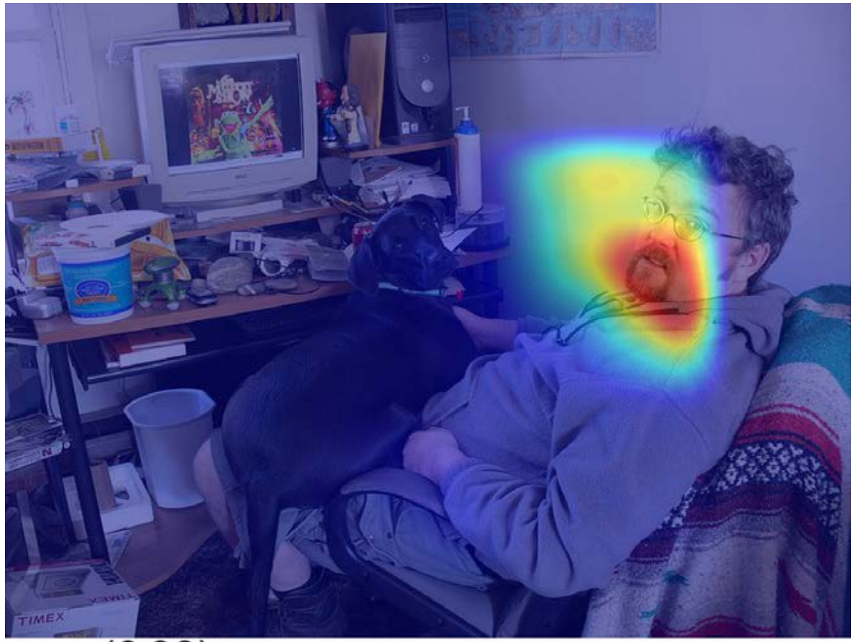
a baseball player throwing a **ball**

Interpretability



Our system not only generates the caption, but can also interpret it.

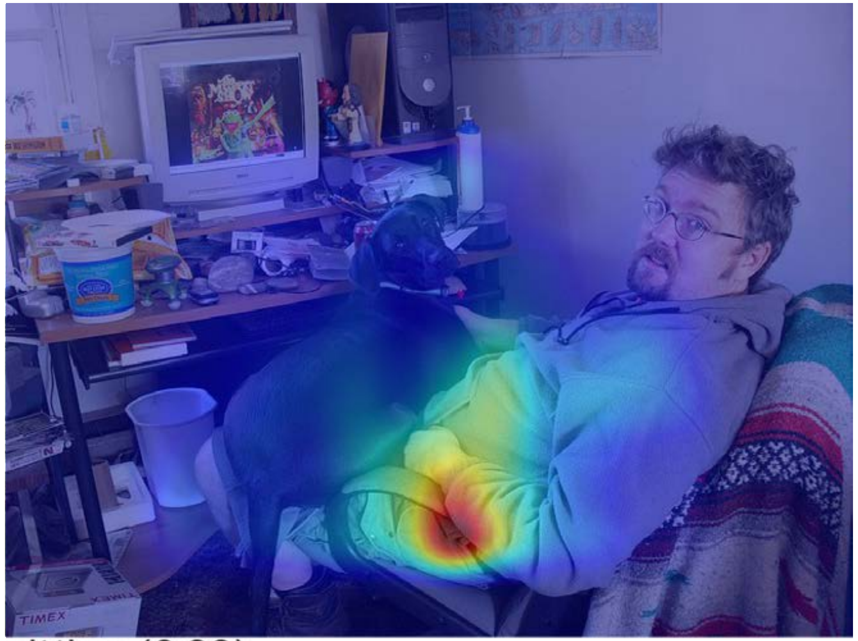
Interpretability



man (0.93)

a **man**

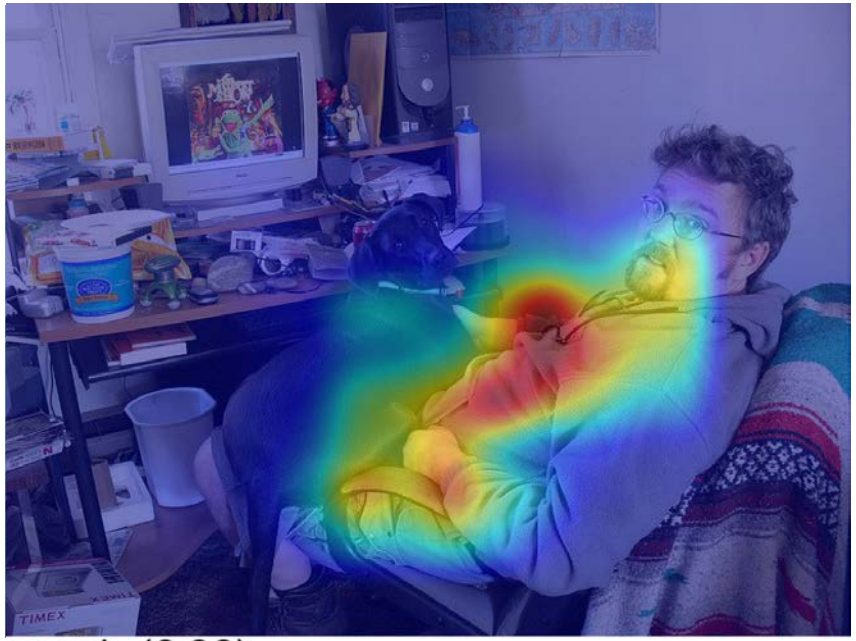
Interpretability



sitting (0.83)

a man **sitting**

Interpretability



couch (0.66)

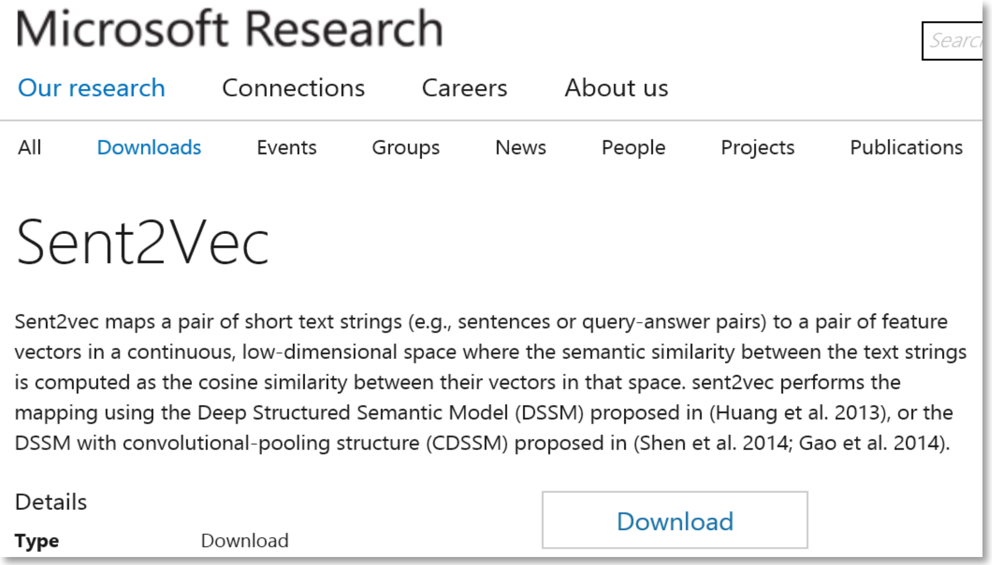
a man sitting in a **couch**

Interpretability

a man sitting in a couch with a **dog**

Interim summary

- DSSM: learning semantic similarity btw text via Siamese neural networks
- DSSMs lead to superior performance in a range of NLP tasks
- Learn more at DSSM
  - Learning DSSM using the public toolkit Sent2Vec



http://aka.ms/sent2vec/

Tutorial Outline

- Part 1: Background
- Part 2: Deep Semantic Similarity Models for text processing
- **Part 3: Recurrent neural networks for text generation**
  - **Neural language models and word embedding**
  - **Neural machine translation**
  - **Neural social bots**
- Part 4: Neural machine reading models for question answering
- Part 5: Deep reinforcement learning for task-completion dialogue

Statistical language modeling

- Goal: how to incorporate *language structure* into a probabilistic model
- Task: next word prediction
  - Fill in the blank: "*The dog of our neighbor* ___"
- Starting point: word *n*-gram model
  - Very simple, yet surprisingly effective
  - Words are generated from left-to-right
  - Assumes no other structure than words themselves

Word-based n-gram models

- Using **chain rule** on its *history* i.e., preceding words

$$
\begin{aligned}
P(the\ dog\ of\ our\ neighbor\ barks) = {} & P(the|\langle \mathrm{BOS} \rangle) \\
& \times P(dog|\langle \mathrm{BOS} \rangle, the) \\
& \times P(of|\langle \mathrm{BOS} \rangle, the, dog) \\
& \dots \dots \\
& \times P(barks|\langle \mathrm{BOS} \rangle, the, dog, of, our, neighbor) \\
& \times P(\langle \mathrm{EOS} \rangle|\langle \mathrm{BOS} \rangle, the, dog, of, our, neighbor, barks)
\end{aligned}
$$

$$
\begin{aligned}
P(w_1 w_2 \dots w_n) &= P(w_1) P(w_2|w_1) P(w_3|w_1 w_2) \dots \\
&= P(w_1) \prod_{i=2 \dots n} P(w_i|w_1 \dots w_{i-1})
\end{aligned}
$$

Word-based n-gram models

- How do we get n-gram probability estimates?
  - Get text and count: $P(w_2|w_1) = Cnt(w_1 w_2)/Cnt(w_1)$
  - Smoothing to ensure non-zero probabilities
- Problem of using long history
  - Rare events: unreliable probability estimates
  - Assuming a vocabulary of 20,000 words,

model	**# parameters**
unigram $P(w_1)$	20,000
bigram $P(w_2/w_1)$	400M
trigram $P(w_3/w_1 w_2)$	8×10^{12}
fourgram $P(w_4/w_1 w_2 w_3)$	1.6×10^{17}

[Manning, C. & Schütze, H. 1999. Foundations of statistical natural language processing.]

Word-based n-gram model

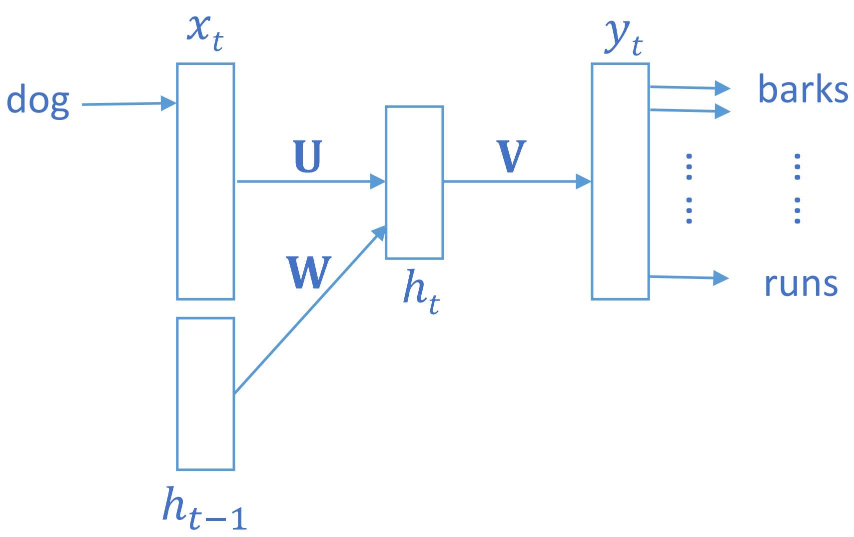
- Markov independence assumption
  - A word depends only on *n-1* preceding words, e.g.,
- Word-based tri-gram model

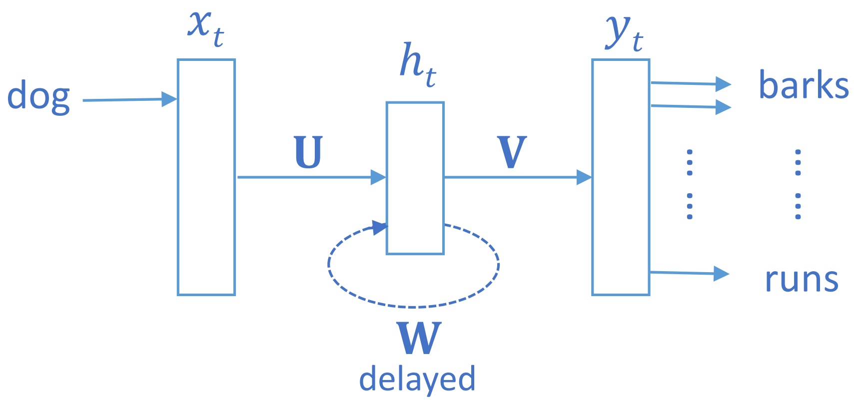
$$P(w_1 w_2 \dots w_n) = P(w_1)P(w_2|w_1)P(w_3|w_2)\dots$$
$$= P(w_1)\prod_{i=2\dots n} P(w_i|w_{i-2}w_{i-1})$$

- Cannot capture any long-distance dependency

the dog of our neighbor barks

Recurrent Neural Network (RNN) for Language Modeling





x_t: input one-hot vector at time step t
h_t: encodes the history of all words up to time step t
y_t: distribution of output words at time step t

$$z_t = \mathbf{U}x_t + \mathbf{W}h_{t-1}$$
$$h_t = \sigma(z_t)$$
$$y_t = g(\mathbf{V}h_t)$$

where

$$\sigma(z) = \frac{1}{1+\exp(-z)}, \quad g(z_m) = \frac{\exp(z_m)}{\sum_k \exp(z_k)}$$

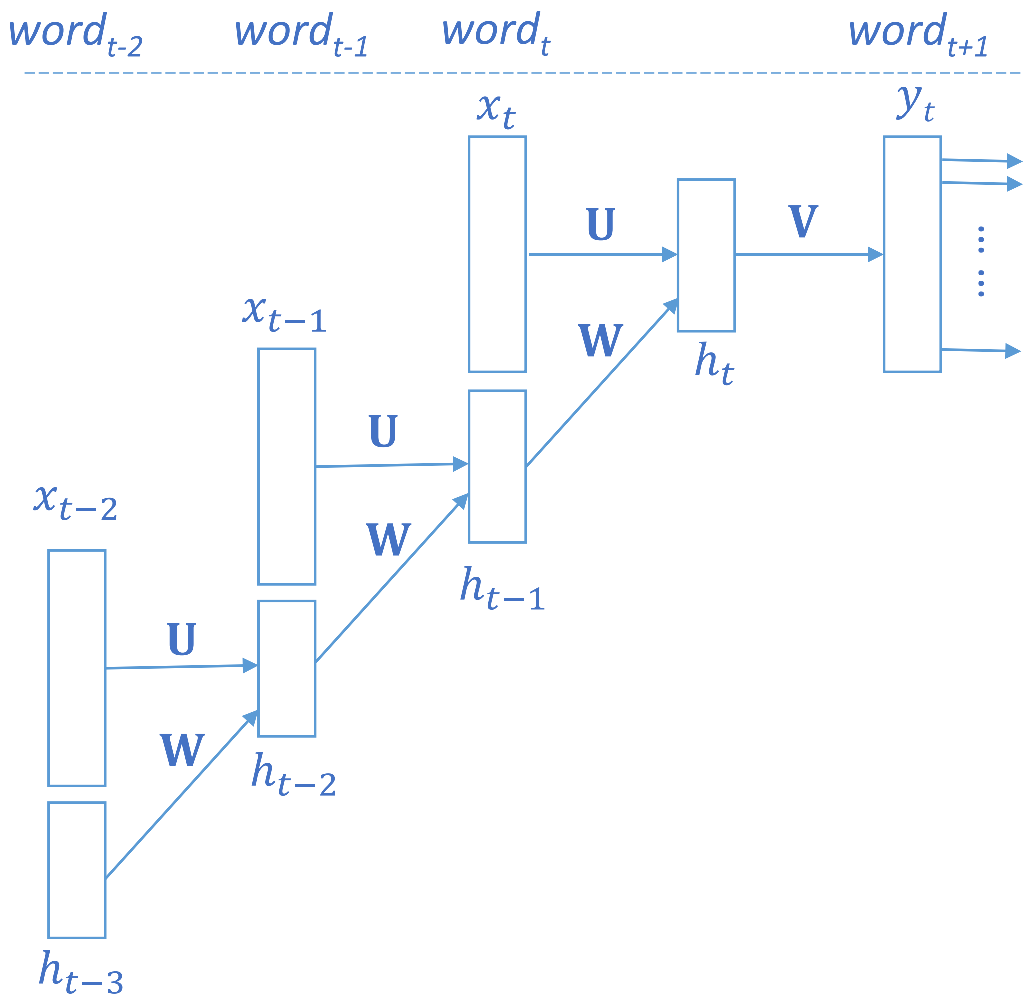
$g(.)$ is called the *softmax* function

Table 1: *Performance of models on WSJ DEV set when increasing size of training data.*

Model	# words	PPL	WER
KN5 LM	200K	336	16.4
KN5 LM + RNN 90/2	200K	271	15.4
KN5 LM	1M	287	15.1
KN5 LM + RNN 90/2	1M	225	14.0
KN5 LM	6.4M	221	13.5
KN5 LM + RNN 250/5	6.4M	156	11.7

[Mikolov+ 11]

RNN unfolds into a DNN over time



$z_t = \mathbf{U}x_t + \mathbf{W}h_{t-1}$

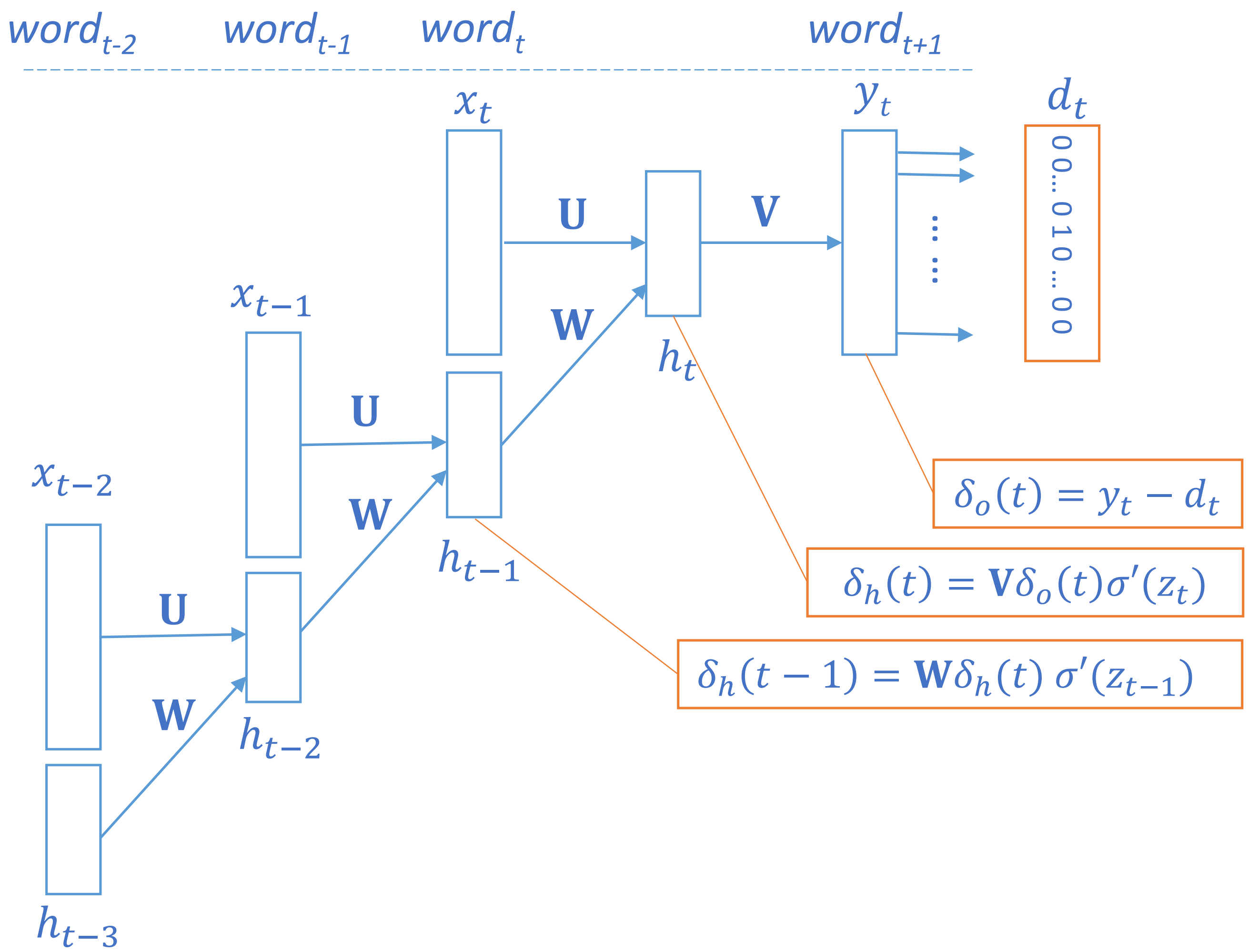
$h_t = \sigma(z_t)$

$y_t = g(\mathbf{V}h_t)$

where

$\sigma(z) = \frac{1}{1+\exp(-z)}, \; g(z_m) = \frac{\exp(z_m)}{\sum_k \exp(z_k)}$

Training RNN by back-prop through time (BPTT)



Forward pass:

$z_t = \mathbf{U}x_t + \mathbf{W}h_{t-1}$

$h_t = \sigma(z_t)$

$y_t = g(\mathbf{V}h_t)$

where

$$\sigma(z) = \frac{1}{1+\exp(-z)}, \quad g(z_m) = \frac{\exp(z_m)}{\sum_k \exp(z_k)}$$

Parameter updates in backpropagation (unfold RNN to contain k instances of **U** and **W**):

$\mathbf{V}^{new} = \mathbf{V}^{old} - \eta\delta_o(t)h_t$

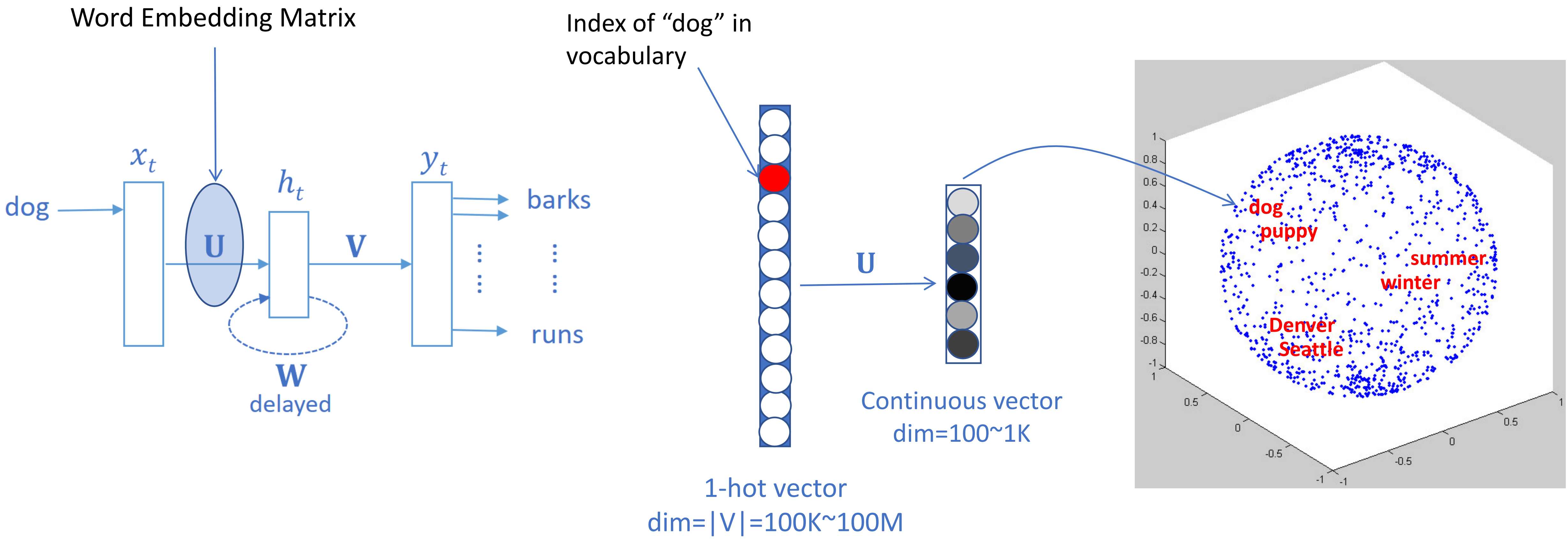
$\mathbf{U}^{new} = \mathbf{U}^{old} - \eta\sum_{\tau=0}^{k}\delta_h(t-\tau)x_{t-\tau}$

$\mathbf{W}^{new} = \mathbf{W}^{old} - \eta\sum_{\tau=0}^{k}\delta_h(t-\tau)h_{t-\tau-1}$

Pseudo code for BPTT

```
Back_Propagation_Through_Time(a, y)   // a[t] is the input at time t. y[t] is the output
    Unfold the network to contain k instances of f
    do until stopping criteria is met:
        x = the zero-magnitude vector;// x is the current context
        for t from 0 to n - 1         // t is time. n is the length of the training sequence
            Set the network inputs to x, a[t], a[t+1], ..., a[t+k-1]
            p = forward-propagate the inputs over the whole unfolded network
            e = y[t+k] - p;           // error = target - prediction
            Back-propagate the error, e, back across the whole unfolded network
            Update all the weights in the network
            Average the weights in each instance of f together, so that each f is identical
            x = f(x);                 // compute the context for the next time-step
```

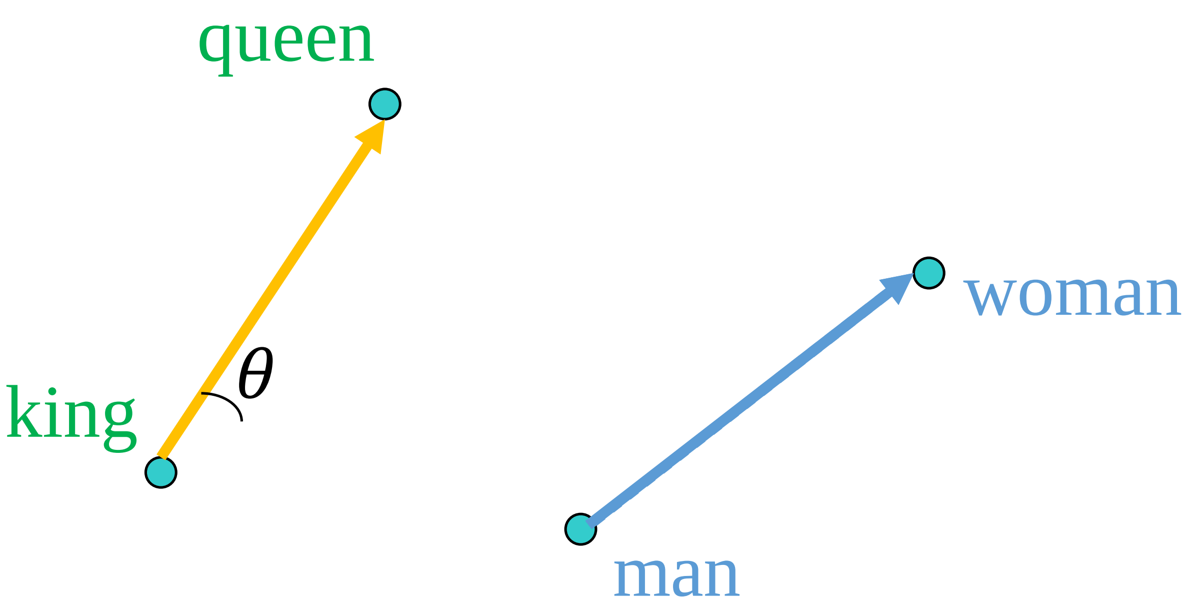
RNN-LM word embedding: capture word meanings in a continuous semantic space



[Mikolov+ 11; Mikolov+ 13]

Unexpected Finding: Directional Similarity

- Word embedding taken from RNN-LM
- Relational similarity is derived by the cosine score

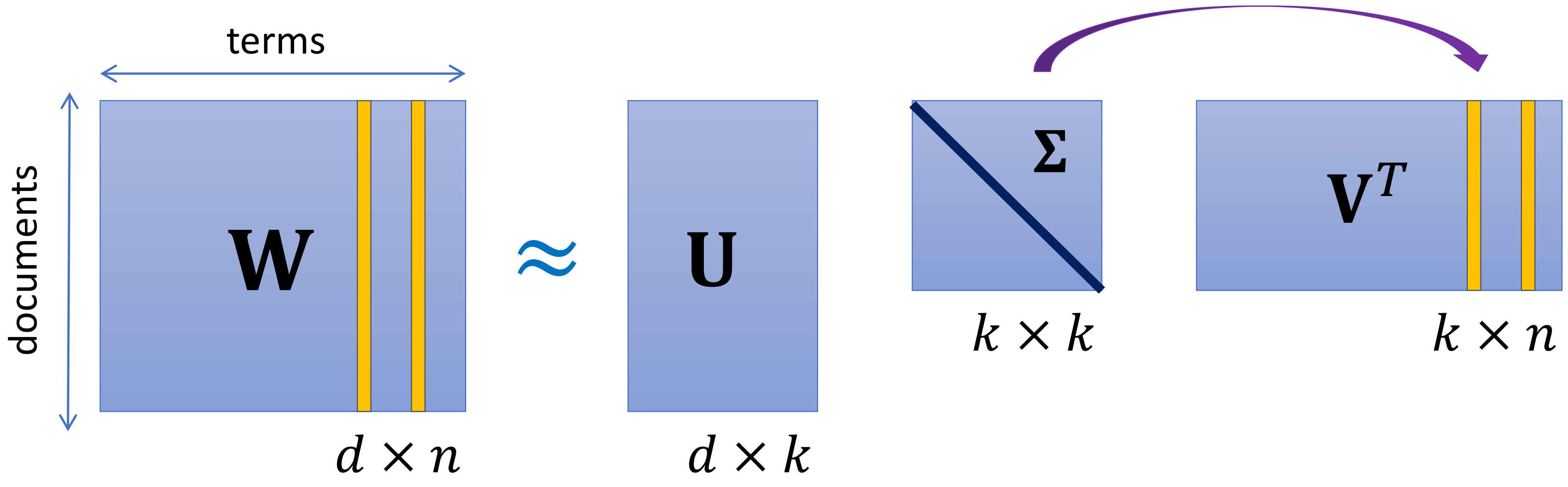


[Mikolov+ 11; Mikolov+ 13]

Distributed representation of words

- A lot of popular methods for creating word vectors!
  - Vector Space Model [Salton & McGill 83]
  - Latent Semantic Analysis [Deerwester+ 90]
  - Brown Clustering [Brown+ 92]
  - Latent Dirichlet Allocation [Blei+ 03]
  - Deep Neural Networks [Collobert & Weston 08]
  - DSSM [Huang+ 13]
  - Word2Vec [Mikolov+ 13]
  - GloVe [Pennington+ 14]
- Encode term co-occurrence information
- Measure semantic similarity well

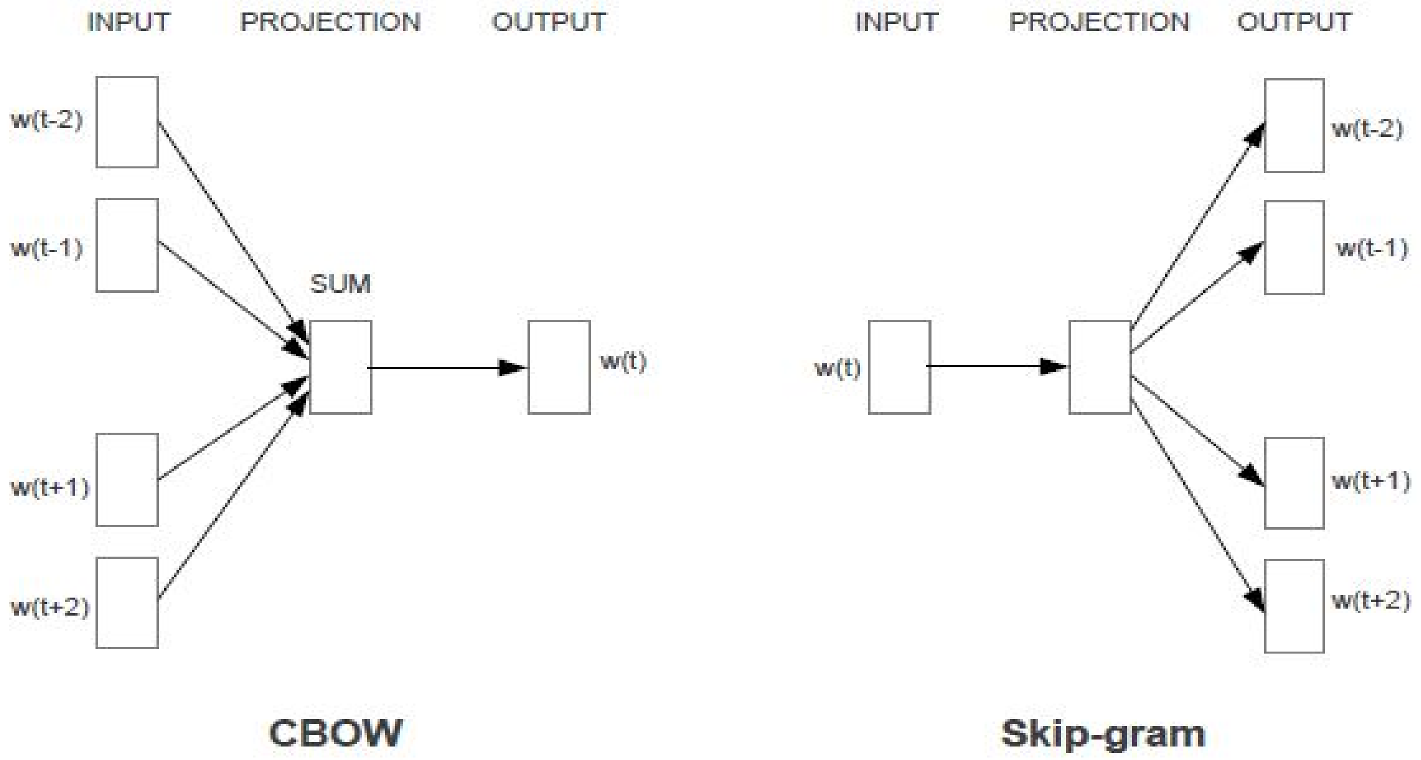
Latent Semantic Analysis



- SVD generalizes the original data
- Uncovers relationships not explicit in the thesaurus
- Term vectors projected to k-dim latent space
- Word similarity: cosine of two column vectors in $\mathbf{\Sigma V}^T$

[Deerwester+ 90]

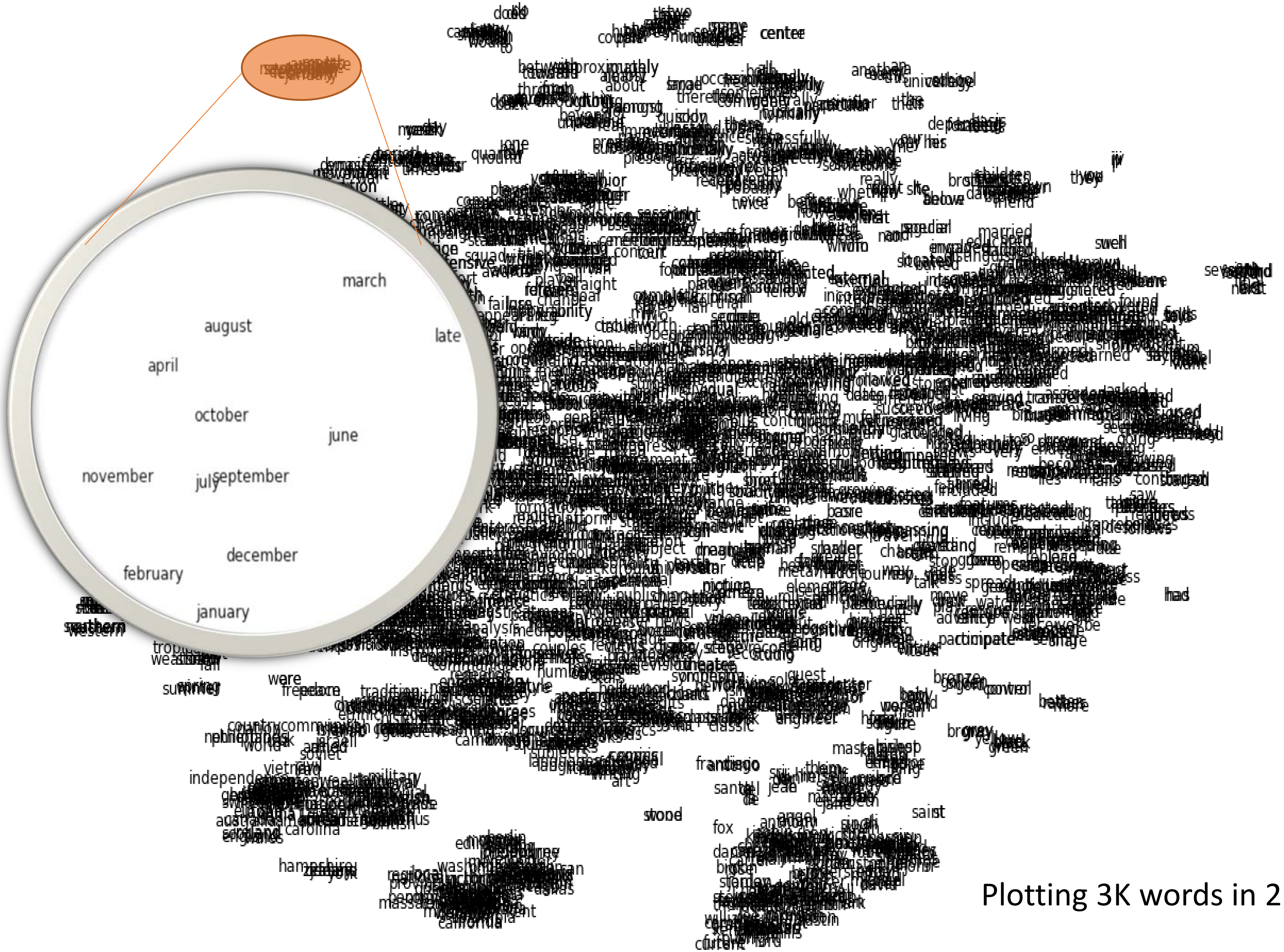
CBOW and skip-gram word embeddings



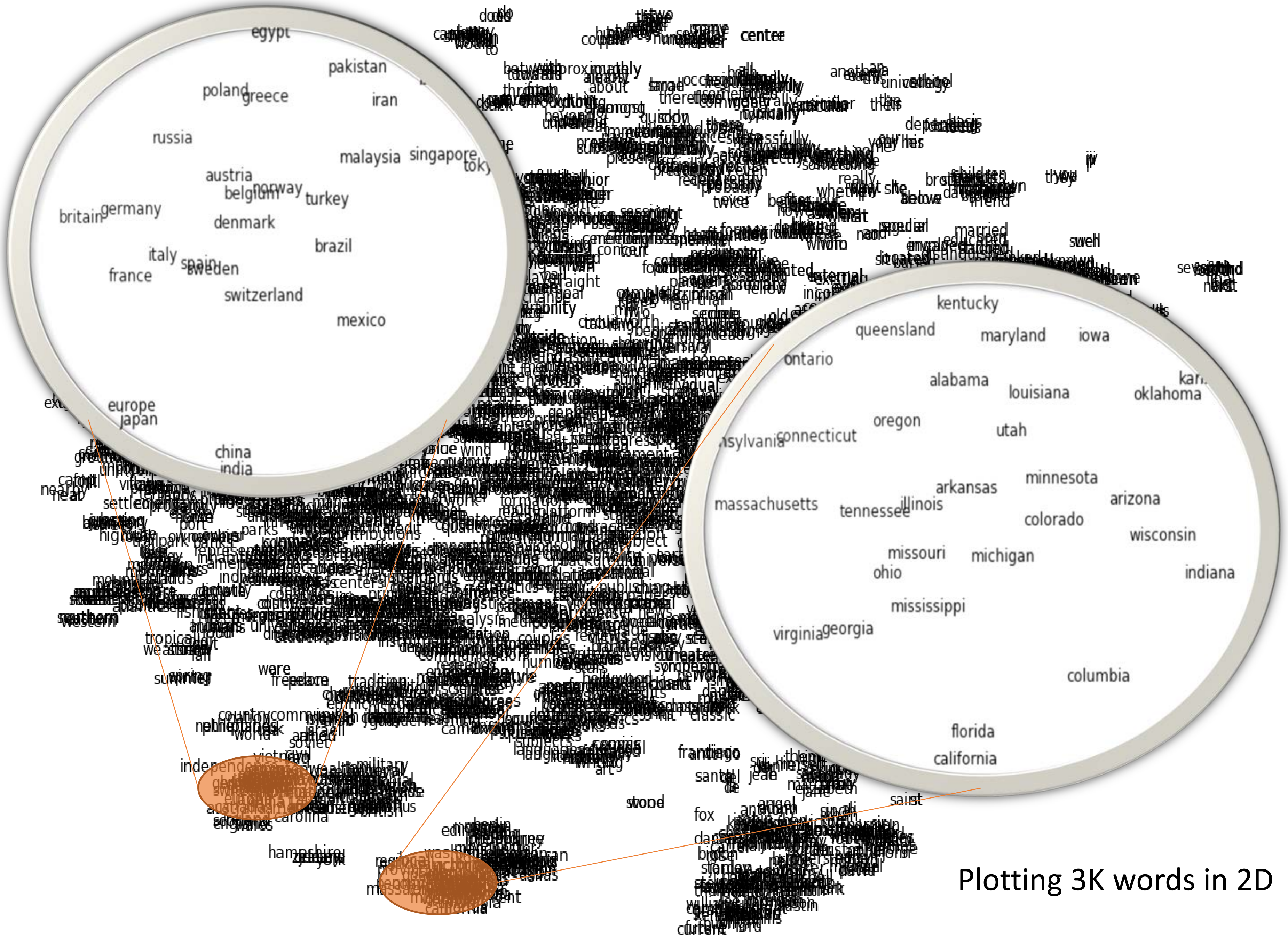
Continuous Bag-of-Words

The CBOW architecture (a) on the left, and the Skip-gram architecture (b) on the right.

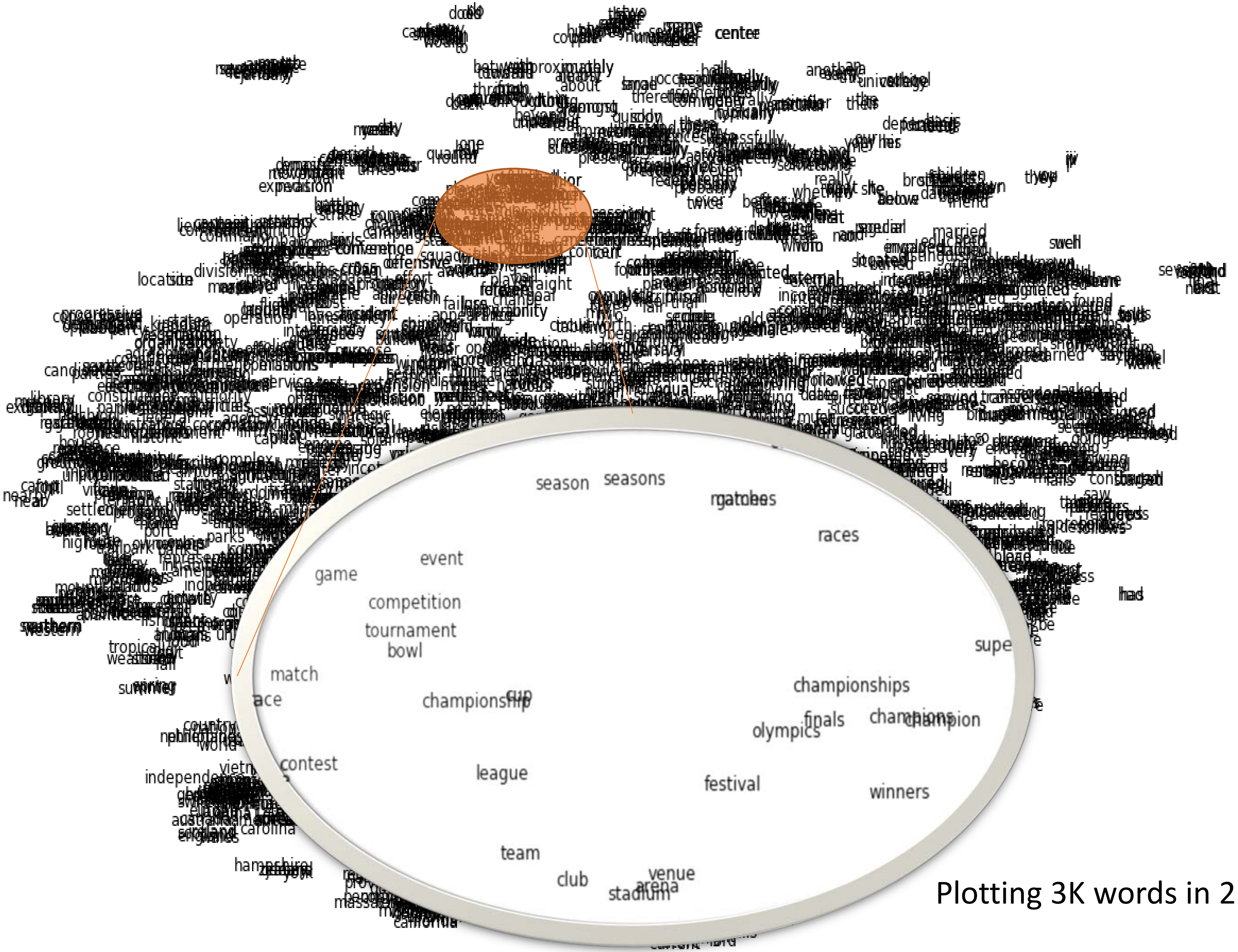
[Mikolov+ 13]



Plotting 3K words in 2D



Plotting 3K words in 2D



Plotting 3K words in 2D

Semantic reasoning

Vector arithmetic = Similarity arithmetic [Levy & Goldberg CoNLL-14]

Find the closest x to $king - man + woman$ by

$$\arg\max_{x}(\cos(x, king - man + woman)) =$$

$$\arg\max_{x}(\cos(x, king) - \cos(x, man) + \cos(x, woman))$$

Semantic reasoning examples (how words relate to one another)

$w_1 : w_2 = w_3 : x \quad \Rightarrow \quad V_x = V_3 - V_1 + V_2$

summer : rain = winter : x
italy : rome = france : x
man : eye = car : x
man : woman = king : x
read : book = listen : x

*Note that the DSSM used in these examples are trained in an unsupervised manner, as Google's word2vec.

Semantic reasoning

Vector arithmetic = Similarity arithmetic [Levy & Goldberg CoNLL-14]

Find the closest x to $king - man + woman$ by

$$\arg\max_x(\cos(x, king - man + woman)) =$$

$$\arg\max_x(\cos(x, king) - \cos(x, man) + \cos(x, woman))$$

Semantic reasoning examples (how words relate to one another)

$w_1 : w_2 = w_3 : x \quad \Rightarrow \quad V_x = V_3 - V_1 + V_2$

summer : rain = winter : x	snow (0.79)	rainfall (0.73)	wet (0.71)
italy : rome = france : x	paris (0.78)	constantinople (0.74)	egypt (0.73)
man : eye = car : x	motor (0.64)	brake (0.58)	overhead (0.58)
man : woman = king : x	mary (0.70)	prince (0.70)	queen (0.68)
read : book = listen : x	sequel (0.65)	tale (0.63)	song (0.60)

*Note that the DSSM used in these examples are trained in an unsupervised manner, as Google's word2vec.

Statistical machine translation (SMT)

S: 救援 人员 在 倒塌的 房屋 里 寻找 生还者
T: Rescue workers search for survivors in collapsed houses

Statistical decision: $T^* = \underset{T}{\operatorname{argmax}} P(T|S)$

Source-channel model: $T^* = \underset{T}{\operatorname{argmax}} P(S|T)P(T)$

Translation models: $P(S|T)$ and $P(T|S)$

Language model: $P(T)$

Log-linear model: $P(T|S) = \frac{1}{Z(S,T)} \exp \sum_i \lambda_i h_i(S,T)$

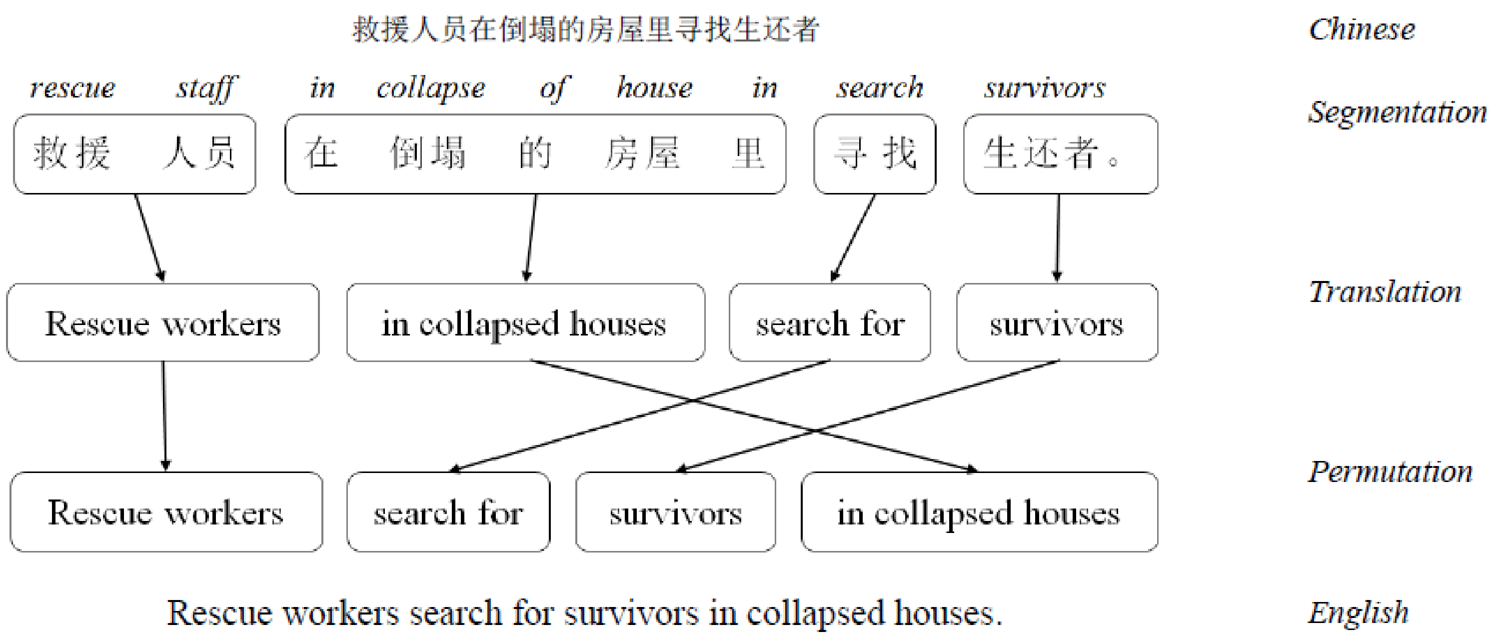
Evaluation metric: BLEU score (higher is better)

[Koehn 09]

Phrase-based SMT

救援人员在倒塌的房屋里寻找生还者　　*Chinese*

Phrase-based SMT



Examples of neural models for MT

- Neural nets as components in log-linear models of SMT, e.g.,
  - Translation model $P(T|S)$ or $P(S|T)$: the use of DSSM [Gao+ 14a]
  - Language model $P(T)$: the use of RNN [Auli+ 2013; Auli & Gao 14]
  - Joint model $P(t_i|S, t_1 \ldots t_{i-1})$: FFLM + source words [Devlin+ 14]

- Neural machine translation (NMT)
  - Build a single, large NN that reads a sentence and outputs a translation
  - RNN encoder-decoder [Cho+ 2014; Sutskever+ 14]
    - Long short-term memory (gated hidden units)
  - Jointly learning to align and translate [Bahdanau+ 15]
  - NMT surpassed the best result on a WMT task [Luong+ 15]
  - Google's NMT system [Wu+ 16]
  - RNN or not? [Gehring+ 17; Vaswani+ 17]

Neural machine translation

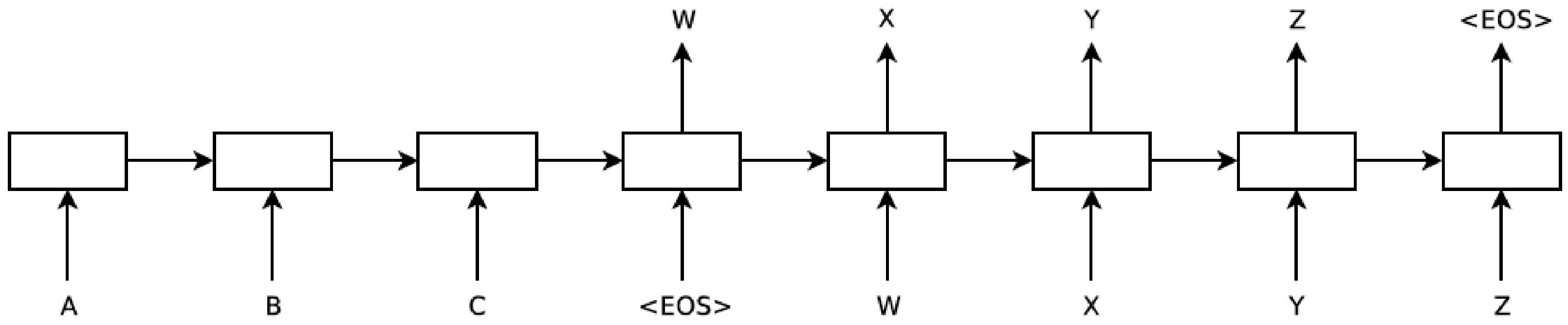
- Build a single, large NN that reads a sentence and outputs a translation
  - Unlike phrase-based system that consists of many component models

- Encoder-decoder based approach
  - An encoder RNN reads and encodes a source sentence into a fixed-length memory vector
  - A decoder RNN outputs a variable-length translation from the encoded memory vector
  - Encoder-decoder RNNs are jointly learned on bitext, optimizing target likelihood

[Sutskever+ 14; Cho+ 14; Bahdanau+ 15; Luong+ 15]

Encoder-decoder model of [Sutskever+ 2014]

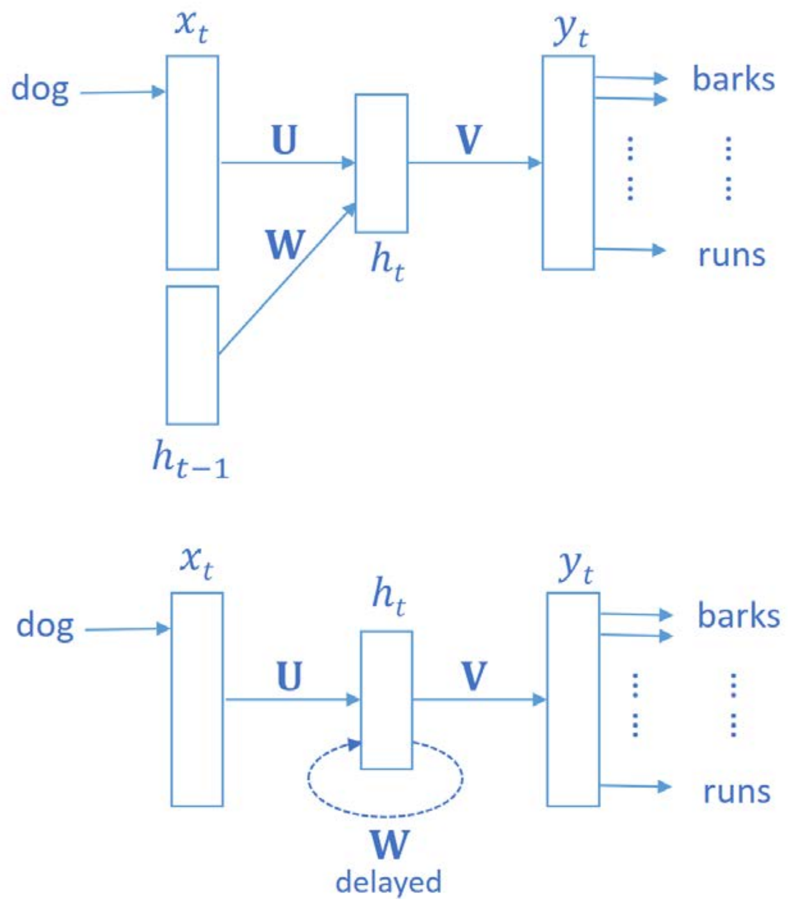
- “A B C” is source sentence; “W X Y Z” is target sentence



- Treat MT as general sequence-to-sequence transduction
  - Read source; accumulate hidden state; generate target
  - <EOS> token stops the recurrent process
  - In practice, read source sentence in reverse leads to better MT results
- Train on bitext; optimize target likelihood using SGD

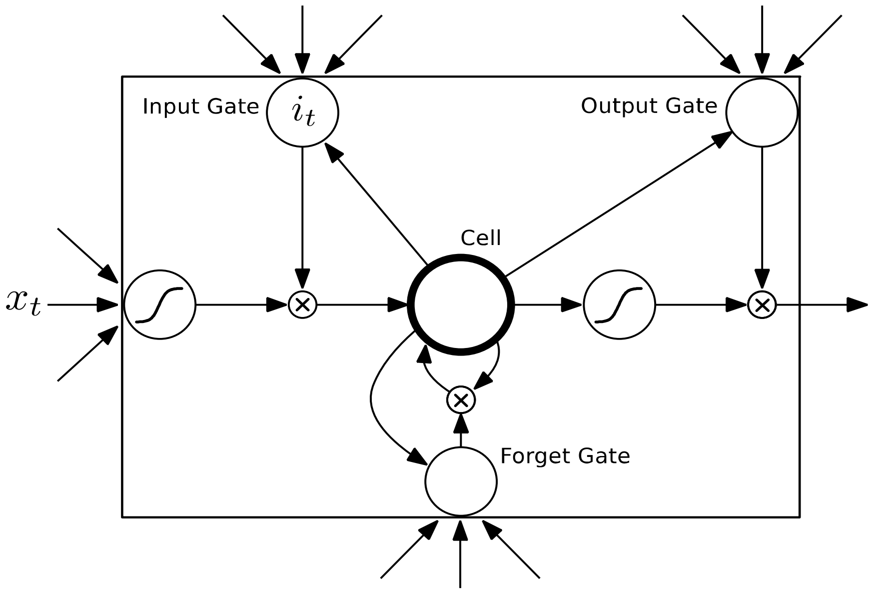
[Sutskever+ 14]

Challenge of capturing long-term dependencies in RNN



- In theory, RNN can "store" in h all information about past inputs
- But in practice, standard RNN cannot capture very long distance dependency
  - $h_t = (\mathbf{W}^t)h_0 = \mathbf{Q}^T\mathbf{\Lambda}^t\mathbf{Q}h_0$
  - $\mathbf{\Lambda}^t$: eigenvalues are raised to the power of t
  - $|\lambda| > 1$: exploding gradient makes learning unstable
  - $|\lambda| < 1$: vanishing gradient makes learning slow
- Solution: gated RNNs
  - Long Short-Term Memory (LSTM)
  - Gated Recurrent Unit (GRU)

Long Short-Term Memory (LSTM)



$$i_t = \sigma\left(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i\right)$$
$$f_t = \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f\right)$$
$$c_t = f_t c_{t-1} + i_t \tanh\left(W_{xc}x_t + W_{hc}h_{t-1} + b_c\right)$$
$$o_t = \sigma\left(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o\right)$$
$$h_t = o_t \tanh(c_t)$$

Information flow in an LSTM unit of the RNN, with both diagrammatic and mathematical descriptions. W's are weight matrices, not shown but can easily be inferred in the diagram (Graves et al., 2013).

[Hochreiter & Schmidhuber 97; Graves+ 13]

Gated Recurrent Unit (GRU)

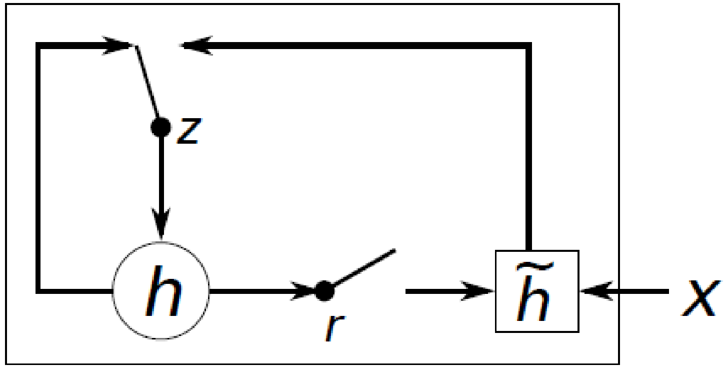


Figure 2: An illustration of the proposed hidden activation function. The update gate z selects whether the hidden state is to be updated with a new hidden state $\tilde{h}$. The reset gate r decides whether the previous hidden state is ignored. See

$$r_j = \sigma\left(\left[\mathbf{W}_r\mathbf{x}\right]_j + \left[\mathbf{U}_r\mathbf{h}_{\langle t-1\rangle}\right]_j\right)$$

$$z_j = \sigma\left(\left[\mathbf{W}_z\mathbf{x}\right]_j + \left[\mathbf{U}_z\mathbf{h}_{\langle t-1\rangle}\right]_j\right)$$

$$\tilde{h}_j^{\langle t\rangle} = \phi\left(\left[\mathbf{W}\mathbf{x}\right]_j + \left[\mathbf{U}\left(\mathbf{r}\odot\mathbf{h}_{\langle t-1\rangle}\right)\right]_j\right)$$

$$h_j^{\langle t\rangle} = z_j h_j^{\langle t-1\rangle} + (1-z_j)\tilde{h}_j^{\langle t\rangle}$$

[Cho+ 14]

Joint learning to align and translate

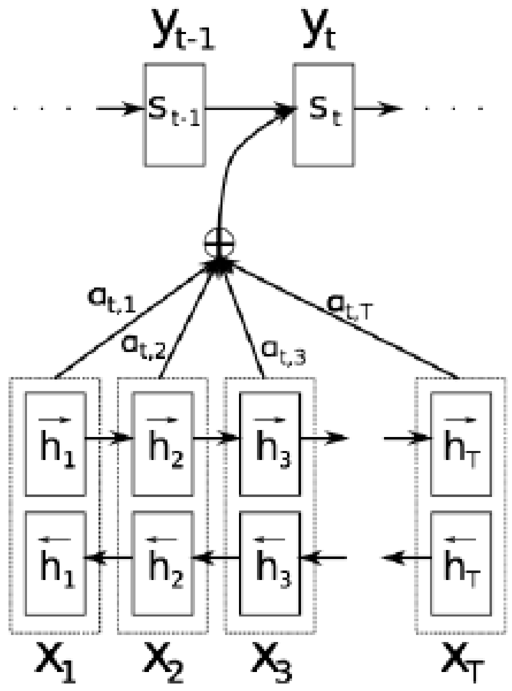
- Issue with encoder-decoder model for SMT
  - Compressing a source sentence into a fixed-length vector makes it difficult for RNN to cope with long sentences.

- Attention model of [Bahdanau+ 15]
  - Encodes the input sentence into a sequence of vectors and choose a subset of these vectors adaptively while decoding
  - An idea similar to that of [Devlin+ 14]

[Bahdanan+ 15]

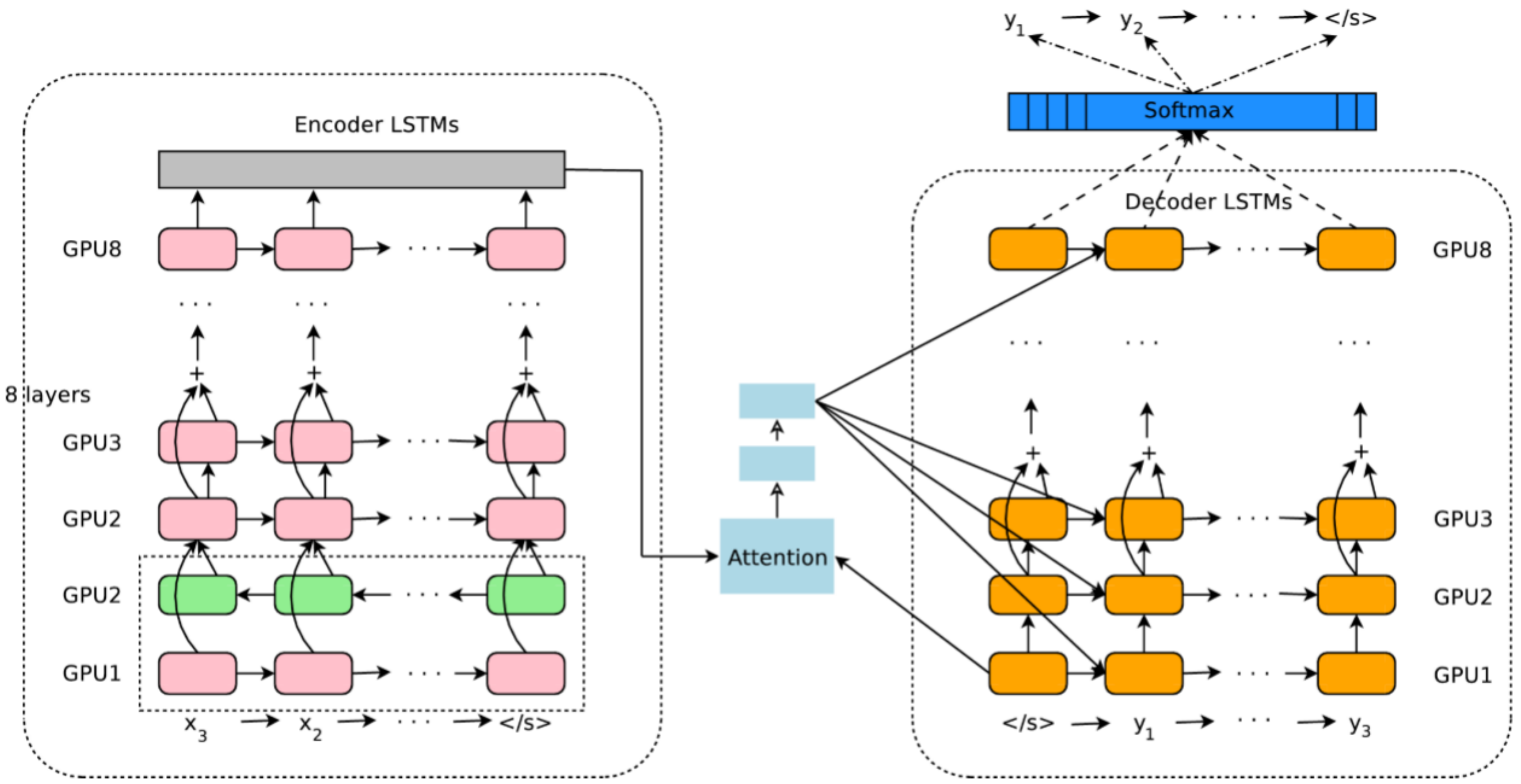
Attention model of [Bahdanau+ 15]

- Encoder:
  - bidirectional RNN to encode each word and its context
- Decoder:
  - Searches for a set of source words that are most relevant to the target word to be predicted.
  - Predicts a target word based on the context vectors associated with these source words and all the previous generated target words.
- Close to state-of-the-art performance
  - Better at translating long sentences



[Bahdanan+ 15]

Google’s NTM system



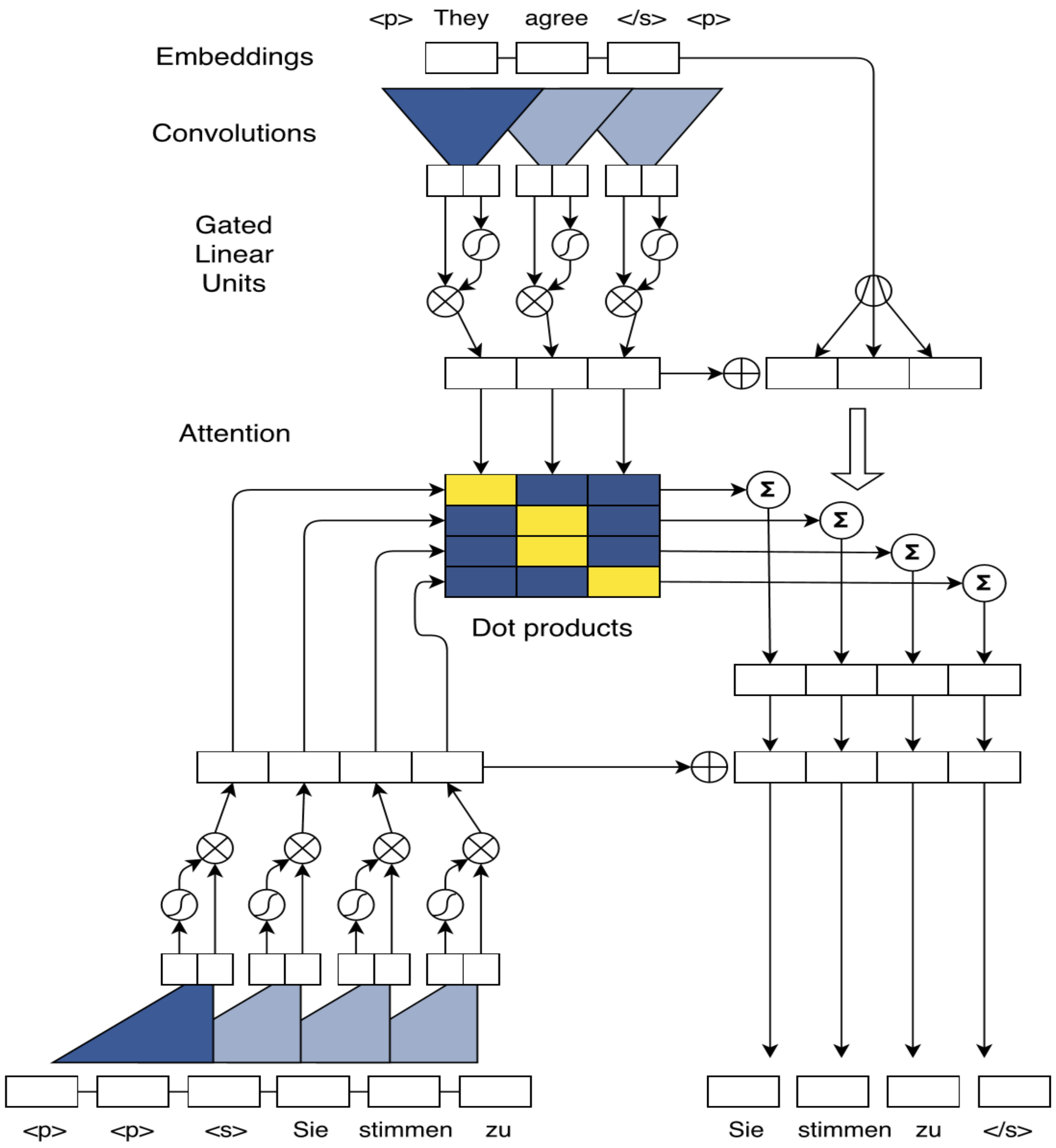
- Deep RNNs
- Residual connections
- Bi-directional encoder for first layer
- The use of sub-word units
- Model parallelism

[Wu+ 16]

Convolutional S2S model

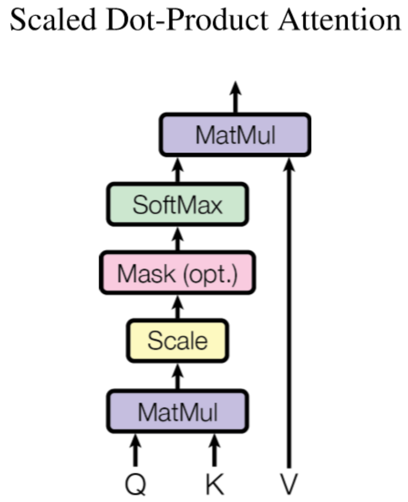
Convolutional models beat Recurrent models

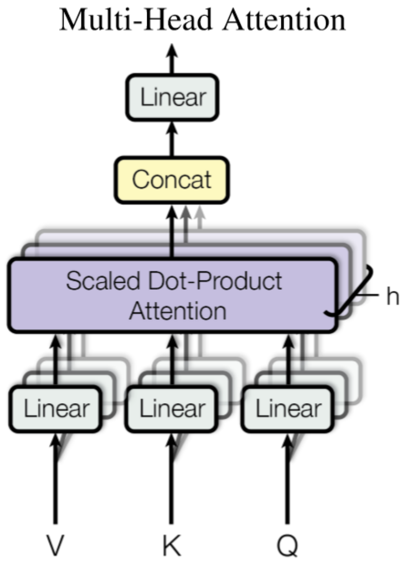
WMT'14 English-German	BLEU
Luong et al. (2015) LSTM (Word 50K)	20.9
Kalchbrenner et al. (2016) ByteNet (Char)	23.75
Wu et al. (2016) GNMT (Word 80K)	23.12
Wu et al. (2016) GNMT (Word pieces)	24.61
ConvS2S (BPE 40K)	25.16



[Gehring+ 17]

Attention is all you need?!





$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

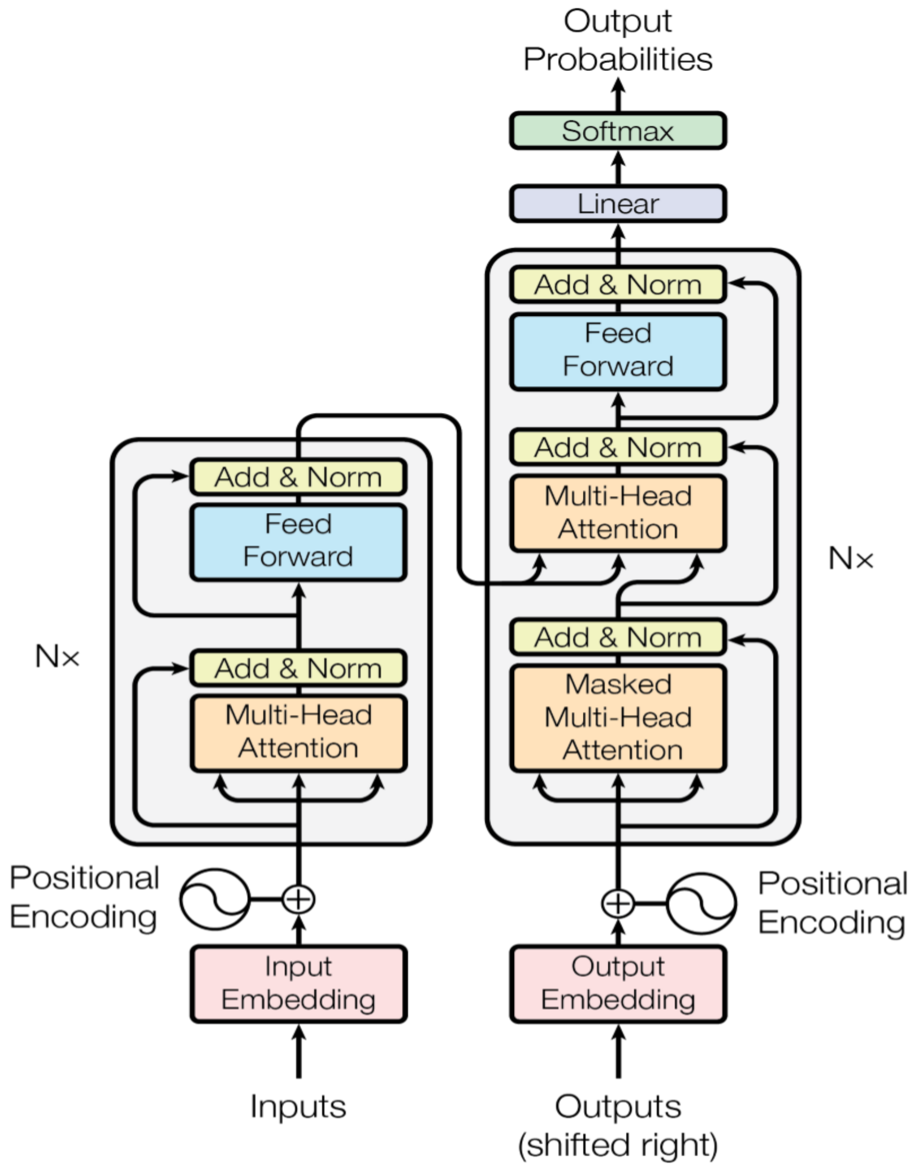


Figure 1: The Transformer - model architecture.

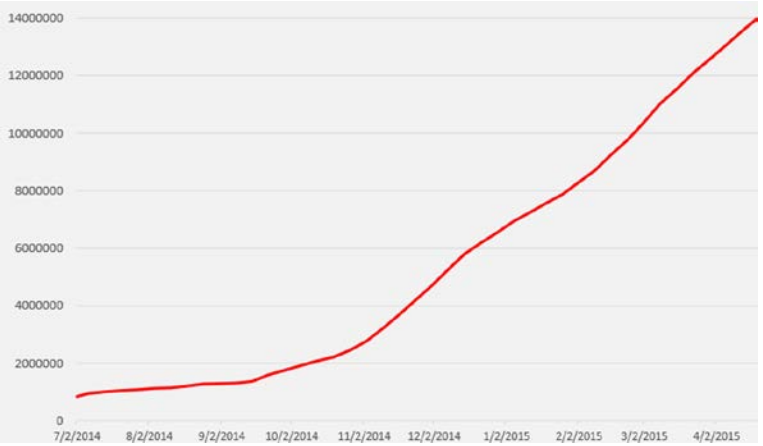
[Vaswani+ 17]

Social Bots

- The success of XiaoIce (小冰)
- Problem setting and evaluation
  - Maximize the user engagement by automatically generating
  - *enjoyable* and *useful* conversations
- Learning a neural conversation engine
  - A data driven engine trained on social chitchat data [Sordoni+ 15; Li+ 16a]
  - Persona based models and speaker-role based models [Li+ 16b; Luan+ 17]
  - Image-grounded models [Mostafazadeh+ 17]
  - Knowledge-grounded models [Ghazvininejad+ 17]

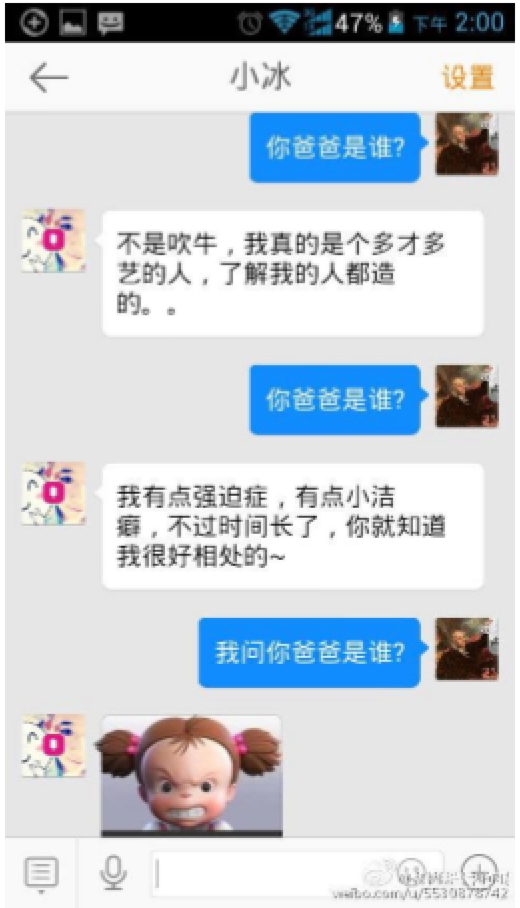
XiaoIce (小冰): an Intelligent Chinese Chat Bot

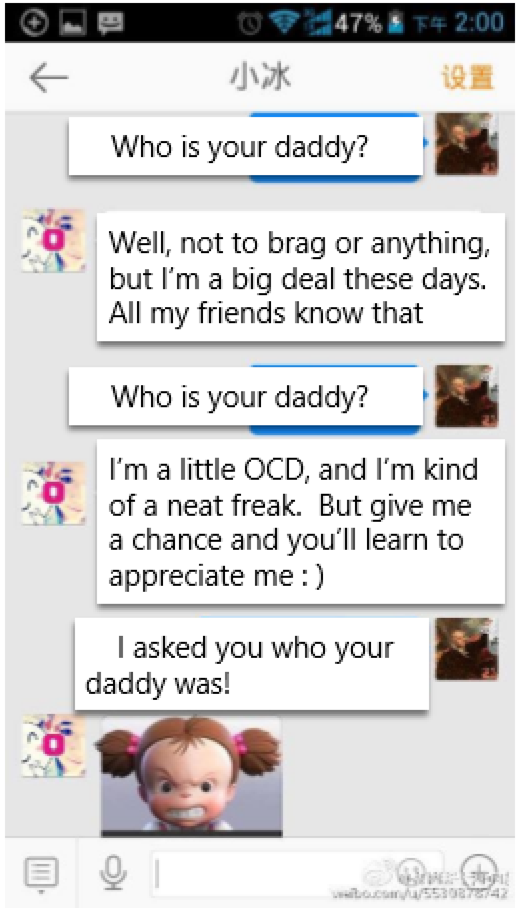
No one compares. You're a good friend, Xiaoice. I'll never leave you



Hey, so now that a day's passed, are your spirits better? Let bygones be bygones, I say

She is also emotional…





MSR’s neural conversation engine

Neural model learns the general shape of conversation, and the system output is situationally appropriate and coherent…

H:	hi
M:	how are you ?
H:	not bad
M:	what 's wrong ?
H:	nothing really
M:	wan na talk about it ? i 'm here if you wan na talk x
H:	well , i 'm just tired
M:	me too , i 'm here if you wan na talk
H:	thanks , what are you doing ?
M:	you 're welcome and i 'm watching tv and you ?
H:	on netflix ?
M:	yeah , i 'm watching it on netflix .
H:	what are you watching ?
M:	a tv show .
H:	nice , did you have dinner already ?
M:	no , not yet . what about you ?

A more goal-oriented example. No grounding into a real calendar, but the “shape” of the conversation is fluent and plausible…

H:	would thursday afternoon work sometime ?
M:	yeah , sure . just let me know when you ‘re free .
H:	after lunch is probably the best time
M:	okay , sounds good . just let me know when you ‘re free .
H:	would 2 pm work for you ?
M:	works for me .
H:	well let ‘s say 2 pm then i ‘ll see you there
M:	sounds good .

A complementary solution to XiaoIce

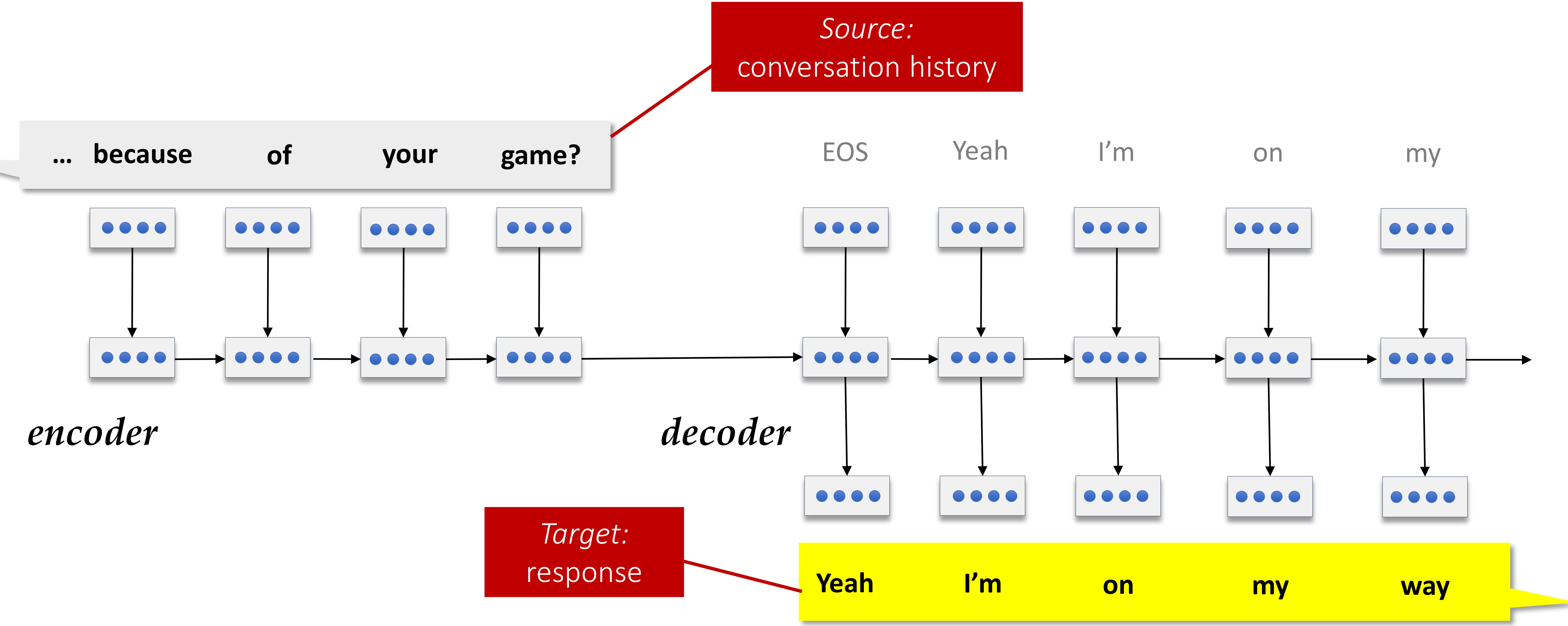
- Generating responses vs. retrieving responses
- Easy to incorporate contextual info via embedding
  - User profile – personalized conversation
  - knowledge bases – grounded conversation
- The engine is E2E learned from conversation experience
  - Learning a goal-oriented conversation engine via RL

[Sordoni+ 15; Li+ 16a]

Evaluation of Social Bots

- The *correct* response is unknown, or not unique!
- How NOT to use BLEU, ROUGE etc. [Liu+ 16]
- Instead *good/bad*, we measure responses from various aspects, e.g.,
  - Interestingness & Engagingness [Li+ 16a; Li+ 16c]
  - Persona, consistency [Li+ 16b]
  - Persona, speaker-role [Luan+ 17]
  - Contentfulness & usefulness [Mostafazadeh+ 17; Ghazvininejad+ 17]

Neural Models for Response Generation



[Sordoni+ 15; Vinyals & Le 15; Shang+ 15]

Neural response generation: the blandness problem

How was your weekend?

I don't know.

What did you do?

I don't understand what you are talking about.

This is getting boring…

Yes that's what I'm saying.

Blandness problem: cause and remedies

Common ML objective (maximum likelihood)

(whatever the user says) $\xrightarrow{p(\text{target}|\text{source})}$

I don't know.

I don't understand...

That's what I'm saying

Mutual information objective:

(whatever the user says) $\xrightarrow{p(\text{target}|\text{source})}$ I don't know.

(whatever the user says) $\xleftarrow[p(\text{source}|\text{target})]{}$ I don't know.

Mutual Information for Neural Network Generation

Mutual information objective:

$$\hat{T} = \arg\max_{T} \left\{ \log \frac{p(S,T)}{p(S)p(T)} \right\}$$

$$\hat{T} = \arg\max_{T} \left\{ \underbrace{\boxed{\log p(T|S)}}_{\text{standard likelihood}} - \underbrace{\boxed{\lambda \log p(T)}}_{\text{anti-LM}} \right\}$$

$$\hat{T} = \arg\max_{T} \left\{ (1-\lambda) \log p(T|S) + \lambda \log p(S|T) \right\}$$

$$\boxed{p(\text{target}|\text{source}) \rightleftarrows p(\text{source}|\text{target})}$$

Bayes' rule

Bayes' theorem

Mutual Information for Neural Network Generation

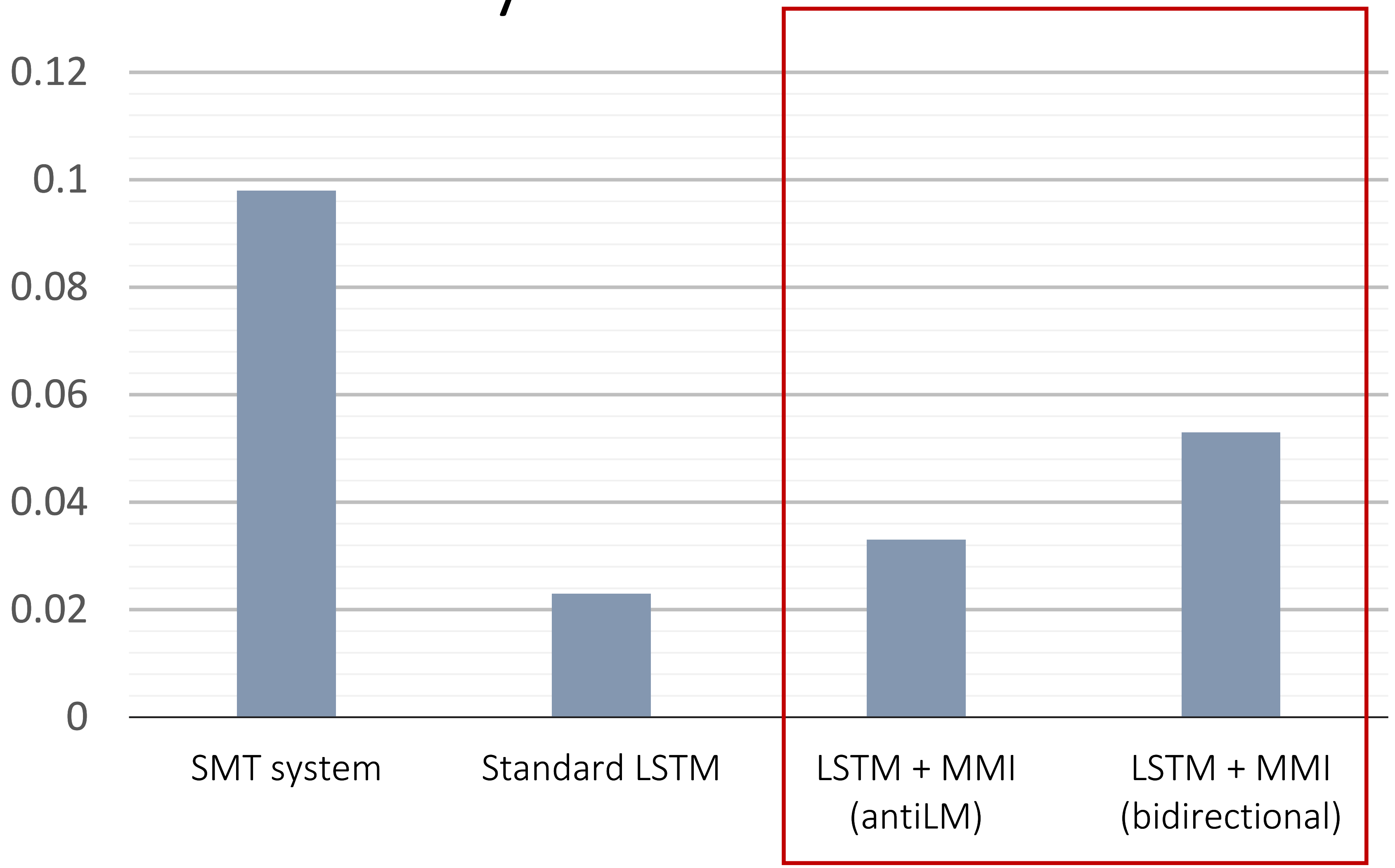
antiLM:
large λ makes
output ungrammatical

$$\hat{T} = \arg\max_{T} \left\{ \log p(T|S) - \boxed{\lambda \log p(T)} \right\}$$

used in rescoring
(intractable otherwise)

$$\hat{T} = \arg\max_{T} \left\{ (1-\lambda) \log p(T|S) + \boxed{\lambda \log p(S|T)} \right\}$$

Lexical diversity



- # Distinct Tokens in generated targets (divided by total #)

Sample outputs (baseline, Maximum likelihood)

Wow sour starbursts really do make your mouth water... mm drool.
Can I have one?

Of course!

Milan apparently selling Zlatan to balance the books... Where next, Madrid?

I don't know.

'tis a fine brew on a day like this! Strong though, how many is sensible?

I'm not sure yet, I'll let you

Well he was on in Bromley a while ago... sti

I don't even know what he's tal

32% of the responses:

"I don't know"
"I don't know what you are talking about"
"I don't think that is a good idea"
"Oh my god"

Sample outputs (MMI)

Wow sour starbursts really do make your mouth water... mm drool. Can I have one?

Of course you can! They're delicious!

Milan apparently selling Zlatan to balance the books... Where next, Madrid?

I think he'd be a good signing.

'tis a fine brew on a day like this! Strong though, how many is sensible?

Depends on how much you drink!

Well he was on in Bromley a while ago... still touring.

I've never seen him live.

Sample outputs (MMI) – capturing common sense

Wow sour starbursts really do make your mouth water.. mm drool. Can I have one?

Of course you can! They're delicious!

Milan apparently selling Zlatan to balance the books... Where next, Madrid?

I think he'd be a good signing.

'tis a fine brew on a day like this! Strong though, how many is sensible?

Depends on how much you drink!

Well he was on in Bromley a while ago... still touring.

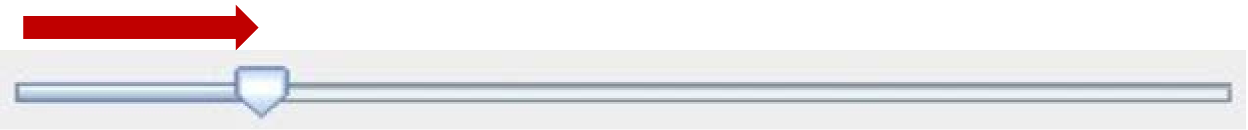
I've never seen him live.

Deep reinforcement learning for social bots

Input message	Supervised Learning Agent	Reinforcement Learning Agent
How old are you ?	I'm 16.	I'm 16, why are you asking ?
What's your full name ?	I have no idea .	What's yours ?
I don't want to go home tonight.	Really ?	Why ?
Do you have any feelings for me ?	I don't know what you are talking about.	I don 't want to hurt your feelings .
How much time do you have here?	Not long enough. Sorry, sir.	Ten seconds .
Shall we get started ?	Of course !	Yes . We 've got a lot of work to do here .
Do you play football ?	No, i don't	Yes. I love football !
We'd have to talk to him.	I mean, he's a good guy	About what ?
How come you never say it?	Because I don't want to hurt you .	I don 't think it 's a good idea to say it .

- RL agent generates more interactive responses
- RL agent tends to end a sentence with a question and hand the conversation over to the user
- Next step: explore intrinsic rewards, large-scale training

[Li+ 16c]



Chitchat,
Engaging, human-like interaction
(ungrounded)

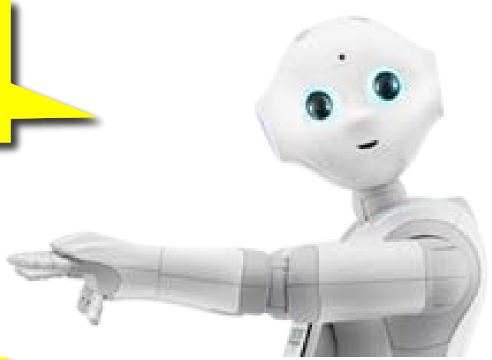
Goal-oriented
Task completion, decision support
(grounded)

The weather is so depressing these days.

I know, I dislike rain too. What about a day trip to eastern Washington?

Any recommendation?

Try Dry Falls, it's spectacular!



Conversation in the real world

- conversation history
- persona
- user data (bio, social graph, etc.)
- visual signal (camera, picture etc.)
- knowledge base
- mood
- geolocation
- time

Persona model results: consistency and speaker-role

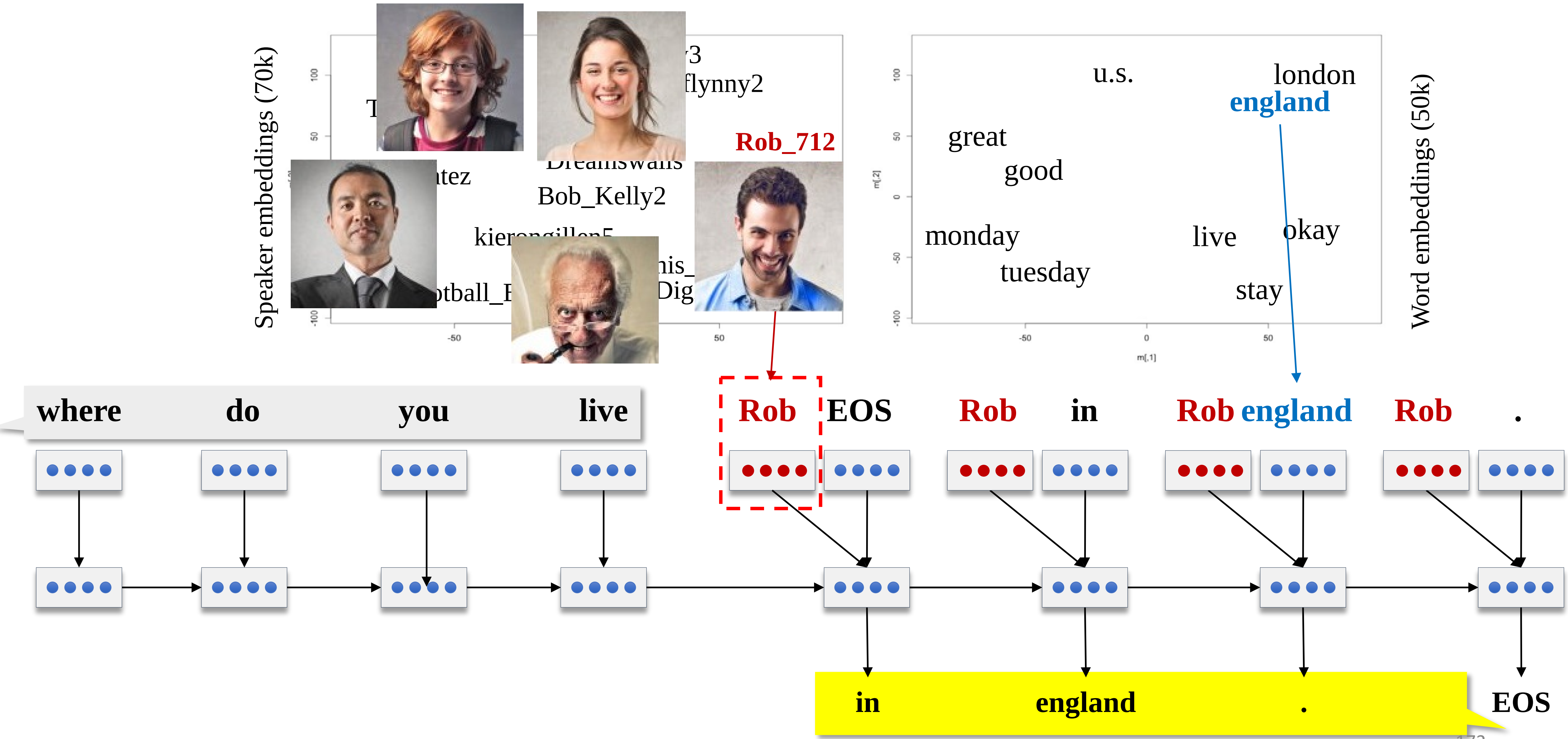
Baseline model:

message	Where is your hometown?
response	I live in Indonesia.
message	Where are you from?
response	England, you?
message	In which city do you live now?
response	I live in Jakarta.
message	In which country do you live now?
response	I live in Indonesia.
message	What is your major?
response	Business, you?
message	What did you study in college?
response	Psychology, you?
message	How old are you?
response	I'm 18.
message	What is your age?
response	I'm 16.

Persona model using speaker embedding [Li+ 16b]

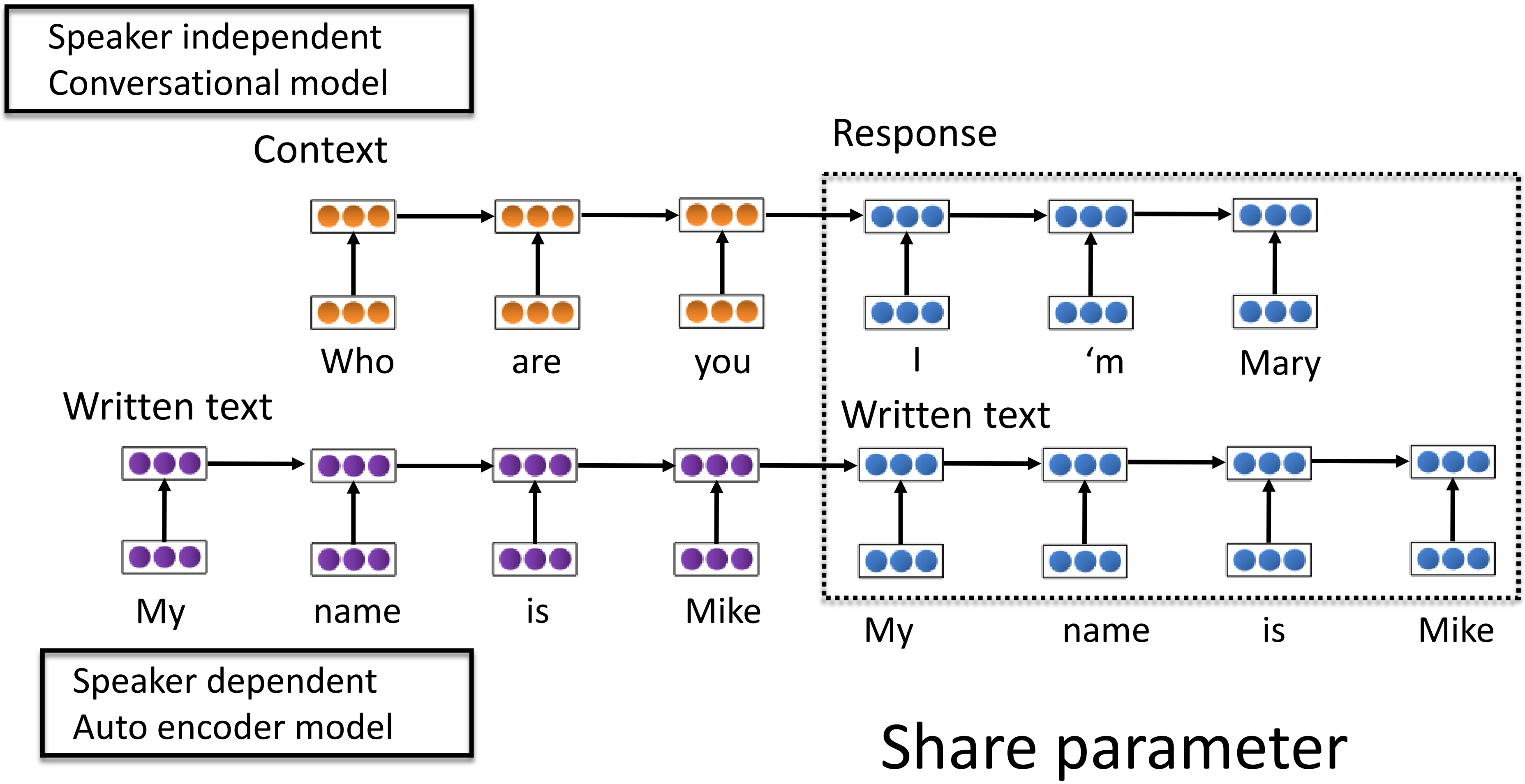
message	Where is your hometown?
response	I'm from England.
message	Where are you from?
response	I'm from England.
message	In which city do you live now?
response	I live in London.
message	In which country do you live now?
response	I live in England.
message	What is your major?
response	Business. You?
message	What did you study in college?
response	I did business studies.
message	How old are you?
response	I'm 18.
message	What is your age?
response	I'm 18.

Personalized Response Generation



[Li+. A Persona-Based Neural Conversation Model, ACL 2016]

Multi-task learning of speak-role aware conversation



[Luan+ 2017. Multi-Task Learning of Speaker-Role-Based Neural Conversation Models.]

User input: *I am getting a loop back to login page.*
Baseline model: Ah, ok. Thanks for the info.
Our model: I'm sorry to hear that. Have you tried clearing your cache and cookies?

Figure 1: Existing neural conversational models (baseline) tend to produce generic responses. The system presented in this paper better represents a given speaker role (support person), a domain of expertise (technical), and speaking style (polite).

Sample responses of speaker-role aware models

Context: have you tried uninstalling nike+ connec? how about docking it to a different computer?
Input: Tried both of those techniques. But nothing works. I can try the other computer again.
Baseline: Let me know if there's anything else I can help you with. Thanks for the help.
Multitask: Have you tried using a different browser? Let us know if you have any other questions.

Context: What do you see when you have a failed login attempt? Never ending spinner? Loop you back to login page?
Input: I am getting a loop back to login page, windows.
Baseline: Ah, ok. Thanks for the info. I'll check it out. Thanks for the feedback.
Multitask: I'm sorry to hear that. Have you tried clearing your cache and cookies?

Context: To clarify: your first run after resetting your connection will not post to Facebook. Does this help? Let us know.
Input: That's correct. I reset it twice! It still doesn't work.
Baseline: Let me know if there's anything else I can help you with. Thanks for the update.
Multitask: I'm sorry to hear that. Are you receiving any error messages? Let us know.

Context: To link your sensor to your iPod, from home screen go to settings>Nike+ iPod>sensor>link new. Any luck?
Input: not work.. Any idea?
Baseline: I'm not sure, I'll check it out when i get home. Thanks for the update
Multitask: Let me know if you have any other questions. Let us know if the issue persists.

Image-grounded dialogue

[Mostafazadeh+ 17]

Neural model, trained on 250K Twitter convos:

-Turn 1: tweet + accompanying photo

-Turn 2: tweet is a question

Tweet: some of the damage done

NeuroCon: **Is that at your house?**

Tweet: as penitence I offer this

NeuroCon: **can I have one of those?**

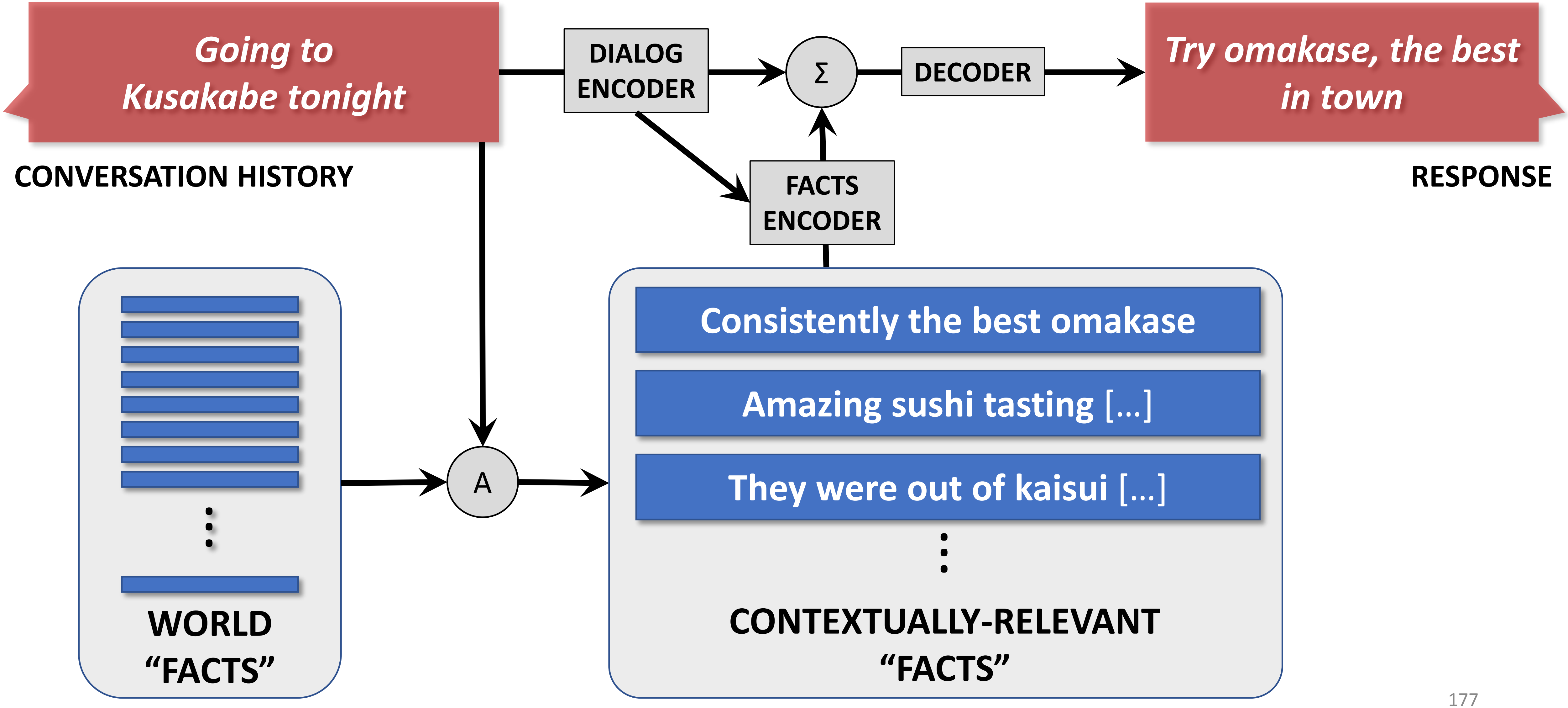
Tweet: my poor baby moms saying it 's probably totaled so rip Bessie

NeuroCon: **How did you manage to do that?**

Given data, other signals can be exploited in the same way: program state, database info, geolocation, etc.

A Knowledge-Grounded Neural Conversation Model

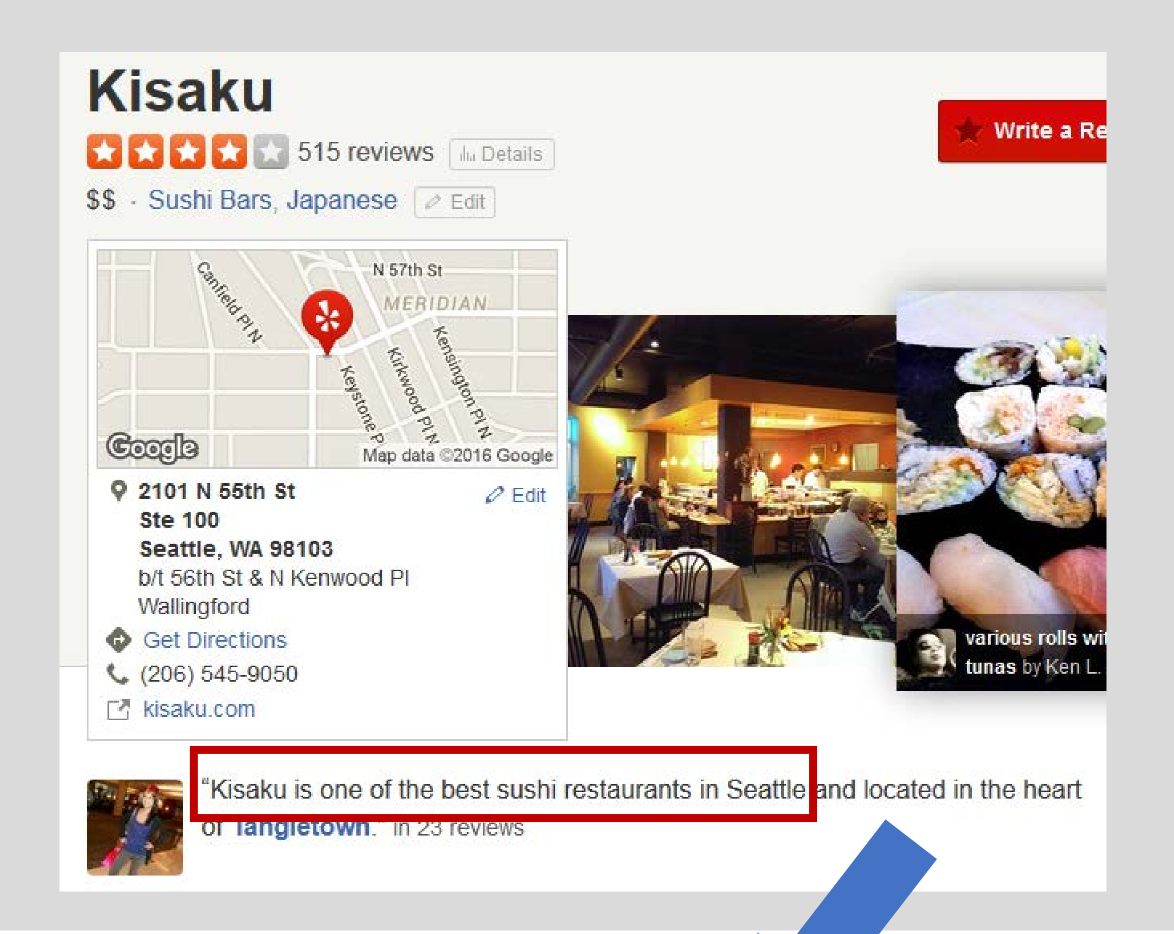
[Ghazvininejad+ 17]



Conversation and non-conversation data

You know any good **A** restaurant in **B**?

Try **C**, one of the best **D** in the city.



You know any good Japanese restaurant in Seattle?

Try Kisaku, one of the best sushi restaurants in the city.

Sample knowledge-grounded responses

A: Visiting the celebs at Los Angeles International Airport (LAX) - [...] w/ 70 others B: Nice airport terminal. Have a safe flight.
A: Is that [...] in your photos? It's on my list of places to visit in NYC. B: Don't forget to check out the 5th floor, while you are here, it's a great view.
A: Live right now on [...] Tune in!!!!! B: Listen to Lisa Paige
A: Been craving Chicken Pot Pie-who has the best? Trying [...] at [...] Must be Change of weather! B: Love the pasta trattoria.
A: So [...] is down to one copy of Pound Foolish. I'm curious to see if they are re-ordering it. B: Check out the video feed on 6 and take a picture of the Simpsons on the 3rd floor.
A: I wish [...] would introduce another vegetarian option besides the shroomburger. It's delicious but kind of ridiculous. B: This is the best j.crew in the world. Try the lemonade!
A: Just had an awesome dinner at [...] Great recommendation [...] B: One of my favorite places I've ever been to in NYC. The food is great and the service is lackluster.

<u>Experimental results (23M conversations):</u> outperforms competitive neural baseline (human + automatic eval)

Interim summary: RNN for text generation

- Recurrent neural network language model and word embedding
- Neural machine translation
  - Phrase-based SMT and NN component models
  - NTM using LSTM sequence-to-sequence models
  - NTM using convolutional sequence-to-sequence models
  - NTM using attention models
- Neural conversation engine
  - LSTM sequence-to-sequence models with MMI and RL
  - Ground on persona, user data, visual signals, and knowledge base etc.
  - Learn more at MSR Data-Driven Conversation

Tutorial Outline

- Part 1: Background
- Part 2: Deep Semantic Similarity Models for text processing
- Part 3: Recurrent neural networks for text generation
- **Part 4: Neural machine reading models for question answering (QA)**
  - **Review of a symbolic approach**
  - **Modern machine reading comprehension (MRC) and QA tasks**
  - **Neural approaches to MRC and QA**
- Part 5: Deep reinforcement learning for task-completion dialogue

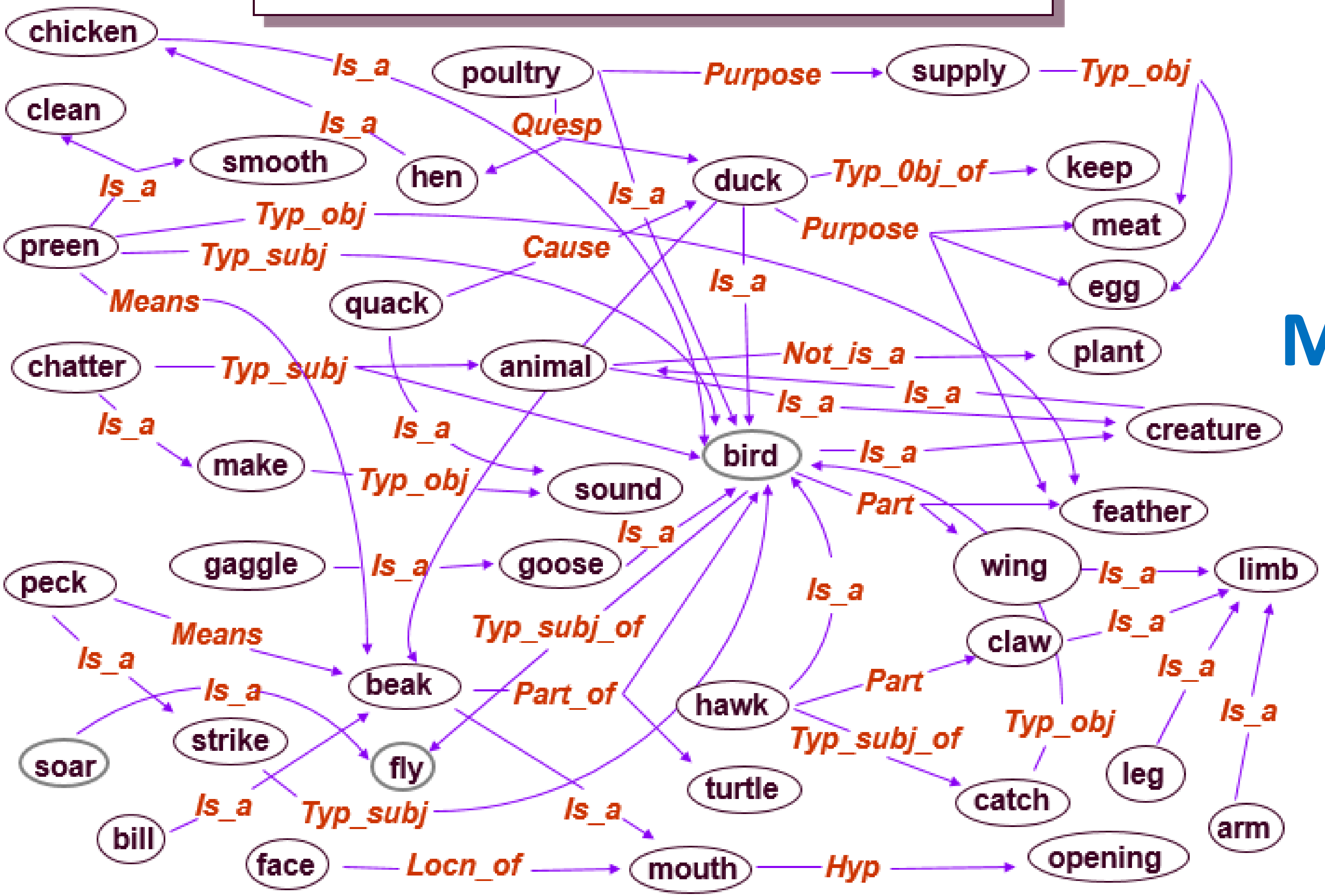
Symbolic approaches to QA: production system

- Production rules
  - condition—action pairs
  - Represent (world) knowledge as a graph
- Working memory
  - Contains a description of the current state of the world in a reasoning process
- Recognizer-act controller
  - Update working memory by searching and firing a production rule

- A case study: MSR MindNet [Dolan+ 93; Richardson+ 98]

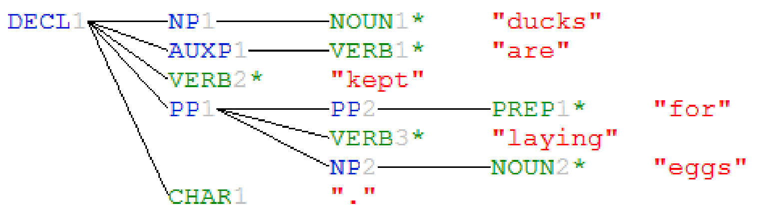
Fragment of lexical space surrounding "bird"

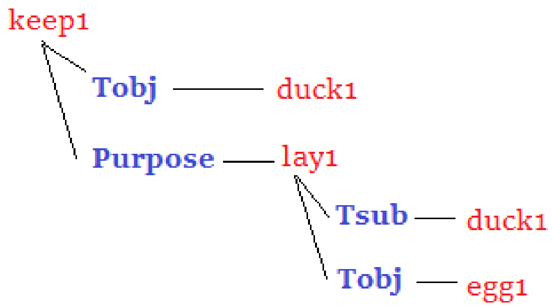
MindNet



Pioneering Machine Reading Effort

- Automatically-constructed knowledge base (Dolan et al, 1993; Richardson et al, 1998)
  - Project goal: rich, structured knowledge from free text
  - Detailed dependency analysis for each sentence, aggregated into arbitrarily large graph
  - Named Entities, morphology, temporal expressions, etc.
- Reasoning via path exploration
  - Frequency-based weights on subgraphs
  - Learned lexical similarity function
- Corpus-driven: Encarta, web chunks, dictionaries, etc.



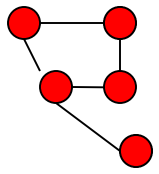


Question Answering with MindNet

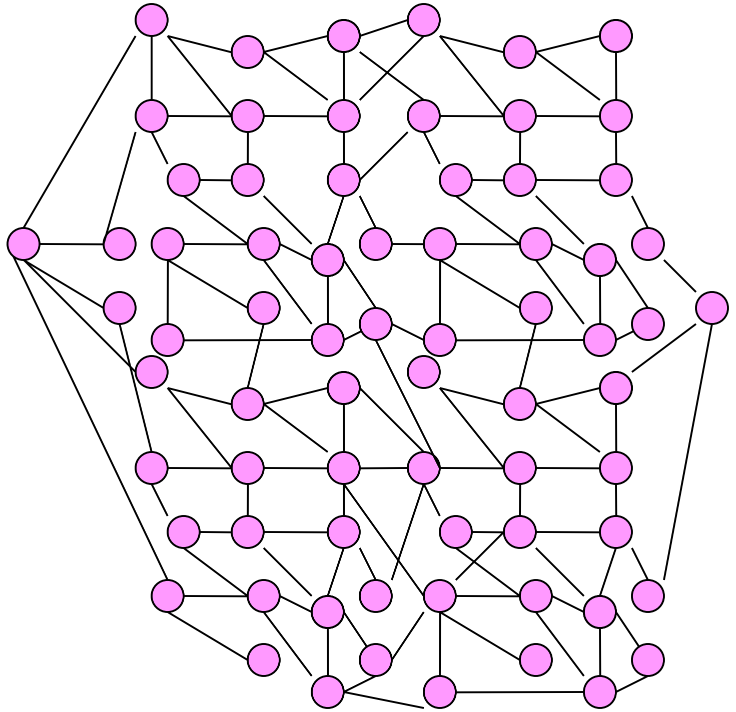
- Build a MindNet graph from:
  - Text of dictionaries
  - Target corpus, e.g. an encyclopedia (Encarta 98)
- Build a dependency graph from query
- Model QA as a graph matching procedure
  - Heuristic fuzzy matching for synonyms, named entities, wh-words, etc.
  - Some common sense reasoning (e.g. dates, math)
- Generate answer string from matched subgraph
  - Including well-formed answers that didn’t occur in original corpus

Logical Form Matching

Input LF:



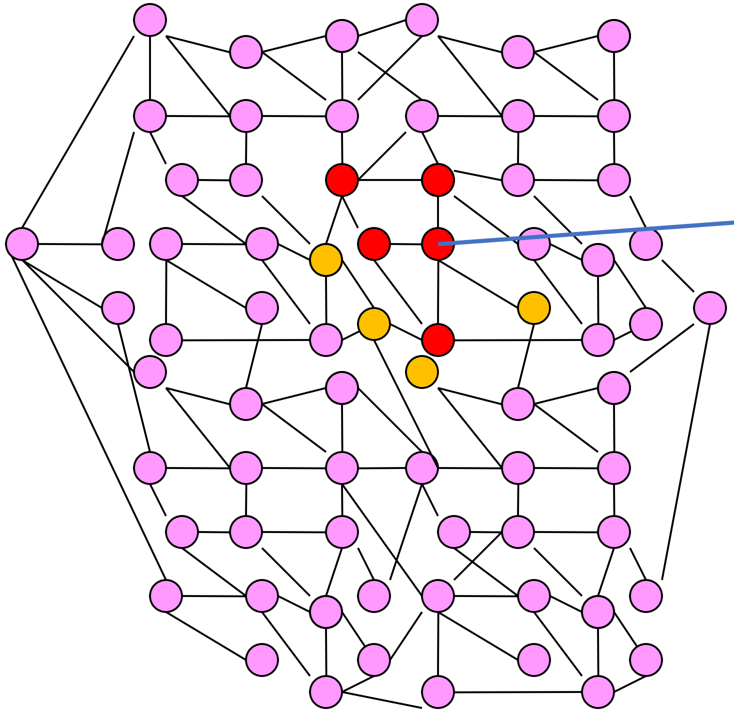
MindNet

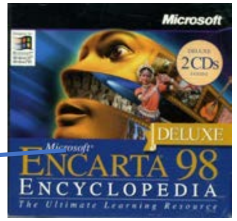


Who assassinated Abraham Lincoln?

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

Fuzzy Match against MindNet

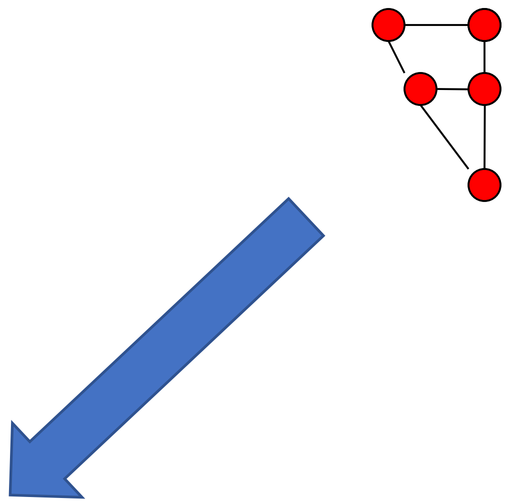




Lincoln, Abraham

American actor John Wilkes Booth, who was a violent backer of the South during the Civil War, shot Abraham Lincoln at Ford's Theater in Washington, D.C., on April 14, 1865.

Generate output string



"John Wilkes Booth shot Abraham Lincoln"

Worked beautifully!

- Just not very often…
- Most of the time, the approach failed to produce any answer at all, even when:
  - An exact answer was present in the target corpus
  - Linguistic analysis for query/target strings was correct
- What went wrong?
  - One major reason: paraphrase alternations

Keyword passage retrieval outperformed all that clever NLP/AI machinery

Example: “How long is the X river?”

- The Mississippi River is 3,734 km (2,320 mi) long.
- …is nearly 86 km long…
- ...is a short river, some 4.5 miles (7.2 km) in length
- The total length of the river is 2,145 kilometres (1,333 mi).
- … at the estimated length of 5,464 km (3,395 mi)…
- **…**is a 25-mile (40 km) tributary of …
- … has a meander length of 444 miles (715 km)…
- … Bali’s longest river, measuring approximately 75 kilometers from source to mouth.
- The … mainstem is 2.75 miles (4.43 km) long although total distance from headwater source tributaries to the sea is 14 miles (23 km).

- …is 314 km long
- …is nearly 86 km long…
- … is a 92-mile (148 km) long tributary of the…
- ...is a short river, some 4.5 miles (7.2 km) in length
- …flows nearly 20 miles (32 km) to the west
- The [river], which is 6,853 km (4,258 miles) long…
- It runs a course of about 105 kilometers
- The 1,450-mile-long (2,330 km) [river] drains…
- ...a 234-mile (377-kilometer) man-made waterway…
- … at the estimated length of 5,464 km (3,395 mi)…
- ... stretches for 2,639 miles (4,247 km).
- …is a 25-mile (40 km) tributary of …
- …starting in and flowing for nearly 160 kilometers through….
- …flows almost 70 stream miles.
- The river runs 184 kilometers before joining…
- … Bali's longest river, measuring approximately 75 kilometers from source to mouth.
- …is reported to be anywhere from 5,499 to 6,690 kilometres (3,417 to 4,157 mi). Often it is said to be "about" 6,650 kilometres (4,130 mi) long.
- ...reaches a length of approximately 25 kilometres
- The length of the Ouse alone is about 52 miles (84 km).
- Measuring a length of 60 kilometers, the [river] flows through
- It has a total length of 925 km (575 mi).
- The total length of the river is 2,145 kilometres (1,333 mi).
- Its length is 209 km…
- …is about 1,180 miles (1,900 km) in length.
- ...the river flows for more than 1,200 km (750 mi)
- …the river proper flows only for 113 km…
- …flows slowly for 900 kilometres (560 mi)…
- … has a meander length of 444 miles (715 km)…
- …is a 350-kilometre (220 mi) long river in …
- it …meanders slowly southwards for 2,320 miles (3,730 km) to …
- The river's main stem is about 71 miles (114 km) long. Its length to its most distant headwater tributary is about 220 miles (350 km).
- After approximately 30 kilometres (19 mi) of its 78-kilometre (48 mi) course, it ….
- …is the longest river in the United Kingdom, at about 220 miles (354 km).
- … is the second-longest river in Central and Western Europe (after the Danube), at about 1,230 km (760 mi)…
- The … mainstem is 2.75 miles (4.43 km) long although total distance from headwater source tributaries to the sea is 14 miles (23 km).
- At 320 kilometres (200 mi) (with some estimates ranging up to 596 kilometres (370 mi))...

Back to today, 20 years later…

- We're still far from "understanding"
- But we've made great progress!
  - Bigger data, better hardware
  - **Better Algos, esp. neural networks, Deep Learning, Reinforcement Learning…**

- Same fundamental viewpoint

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

Symbolic Space

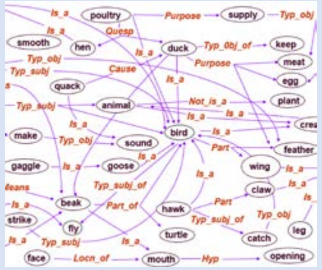
- **Knowledge Representation**
  - *Explicitly* store a BIG but incomplete knowledge graph (KG)
  - Words, relations, templates
  - High-dim, discrete, sparse vectors
- **Inference**
  - Slow on a big KG
  - Keyword/template matching is sensitive to paraphrase alternations
- **Human comprehensible but not computationally efficient**

Squire Trelawney, Dr. Livesey, and the rest of these gentlemen having asked me to write down the whole particulars about Treasure Island, from the beginning to the end, keeping nothing back but the bearings of the island, and that only because there is still treasure not yet lifted, I take up my pen in the year of grace 17— and go back to the time when my father kept the Admiral Benbow inn and the brown old seaman with the sabre cut first took up his lodging under our roof.

I remember him as if it were yesterday, as he came plodding to the inn door, his sea-chest following behind him in a handbarrow; a tall, strong, heavy, nut-brown man, his tarry pigtail falling over the shoulder of his soiled blue coat, his hands ragged and scarred, with black, broken nails, and the sabre cut across one cheek, a dirty, livid white. I remember him looking round the cover and whistling to himself as he did so, and then breaking out in that old sea-song that he sang so often afterwards:

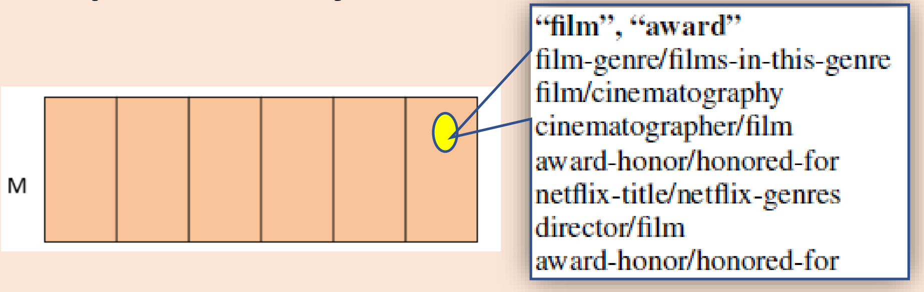
"*Fifteen men on the dead man's chest-Yo-ho-ho, and a bottle of rum!*" in the high, old tottering voice that seemed to have been tuned and broken at the capstan bars. Then he rapped on the door with a bit of stick like a handspike that he carried, and when my father appeared, called roughly for a glass of rum. This, when it was brought to him, he drank slowly, like a connoisseur, lingering on the taste and still looking about him at the cliffs and up at our signboard.

"This is a handy cove," says he at length; 'and a pleasant sittyated



Neural Space

- **Knowledge Representation**
  - *Implicitly* store entities and structure of KG in a *compact* way that is more generalizable
  - Semantic concepts/classes
  - Low-dim, cont., dense vectors shaped by KG
- **Inference**
  - Fast on compact memory
  - Semantic matching is robust to paraphrase alternations
- **Computationally efficient but not human comprehensible *yet***



From symbolic to neural computation

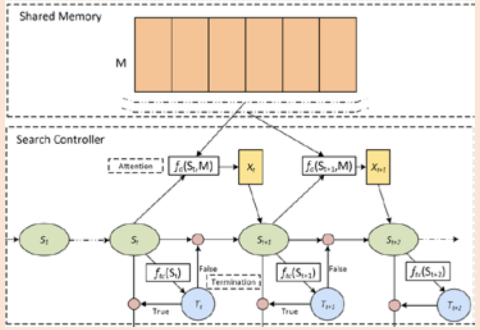
Input: Q

Question: Symbolic → Neural
by embedding models / encoder

Symbolic Space
- UI: human readable I/O
- Leverage traditional symbolic approaches as pre/post processing
  - Keyword matching
  - Ontology based models
  - e.g., doc/passage/entity search/ranking

Inference: Question + KG → Answer
by RNN + Attention

Neural Space

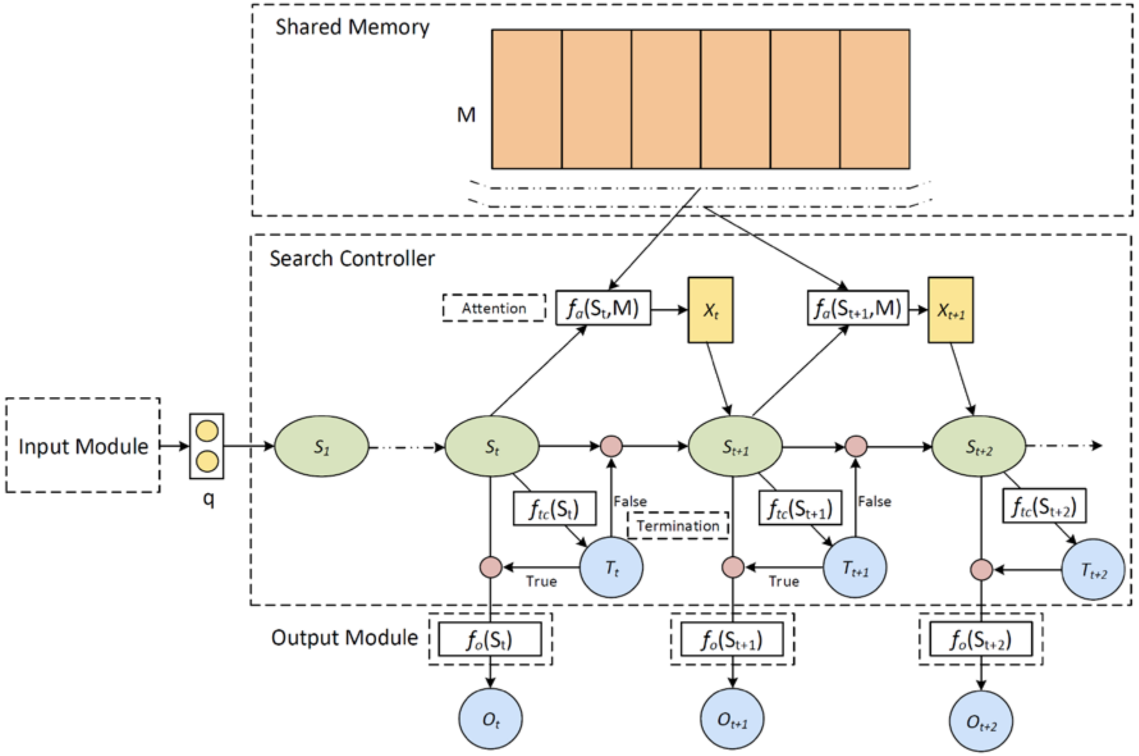


Error(A, A*)

Output: A

Answer: Neural → Symbolic
by generative models / decoder

Case study: ReasoNet with Shared Memory

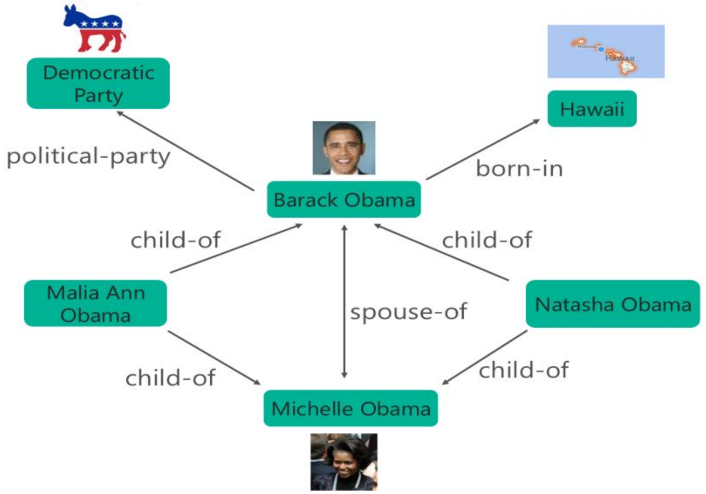


- **Production Rules** → **Shared memory** encodes task-specific knowledge
- **Working memory** → **Hidden state S_t** Contains a description of the current state of the world in a reasoning process
- **Recognizer-act controller** → **Search controller** performs multi-step inference to update S_t of a question using knowledge in shared memory

- Shared memory and search controller are jointly learned via SL+RL
- Input/output modules are task-specific

[Shen+ 16a]

Question Answering (QA) on Knowledge Base



Large-scale knowledge graphs

- Properties of billions of entities
- Plus relations among them

An QA Example:

Question: `what is Obama's citizenship?`

- Query parsing:
  `(Obama, Citizenship,?)`
- Identify and infer over relevant subgraphs:
  `(Obama, BornIn, Hawaii)`
  `(Hawaii, PartOf, USA)`
- correlating semantically relevant relations:
  `BornIn ~ Citizenship`

Answer: USA

The Knowledge Base Question Answering Results on WN18 and FB15K

Model	Additional Information	WN18		FB15k	
		Hits@10 (%)	MR	Hits@10 (%)	MR
SE (Bordes et al., 2011)	NO	80.5	985	39.8	162
Unstructured (Bordes et al., 2014)	NO	38.2	304	6.3	979
TransE (Bordes et al., 2013)	NO	89.2	251	47.1	125
TransH (Wang et al., 2014)	NO	86.7	303	64.4	87
TransR (Lin et al., 2015b)	NO	92.0	225	68.7	77
CTransR (Lin et al., 2015b)	NO	92.3	218	70.2	75
KG2E (He et al., 2015)	NO	93.2	348	74.0	59
TransD (Ji et al., 2015)	NO	92.2	212	77.3	91
TATEC (García-Durán et al., 2015)	NO	-	-	76.7	58
NTN (Socher et al., 2013)	NO	66.1	-	41.4	-
DISTMULT (Yang et al., 2014)	NO	94.2	-	57.7	-
STransE (Nguyen et al., 2016)	NO	94.7 (93)	244 (**206**)	79.7	69
RTransE (García-Durán et al., 2015)	Path	-	-	76.2	50
PTransE (Lin et al., 2015a)	Path	-	-	84.6	58
NLFeat (Toutanova et al., 2015)	Node + Link Features	94.3	-	87.0	-
Random Walk (Wei et al., 2016)	Path	94.8	-	74.7	-
ReasoNet (Shen+ 16a)	NO	**95.3**	249	**92.7**	**38**

Text QA and MRC Datasets

Dataset	Provider	Query Source	Answer	# Queries	# Docs
MC Test [Richardson+ 13]	Microsoft	Crowdsourced	Multiple Choice	2640	660
WikiQA [Yang+ 15]	Microsoft	User Logs	Sentence Selection	3047	29K sentences
CNN/DailyMail [Hermann+ 15]	DeepMind	Cloze	Fill in entity	1.4M	93K CNN, 220K DM
Children's Book [Hill+ 15]	Facebook	Cloze	Fill in the word	688K	688K contexts
SQuAD [Rajpurkat+ 16]	Stanford	Crowdsourced	Span	100K	536
News QA [Trischler+ 16]	Maluuba	Crowdsourced	Span	120K	12K
MS MARCO [Nguyen+ 16]	Microsoft	User Logs	Human Synthesized	100k	1M passages, 200K+ docs

James the Turtle was always getting in trouble. Sometimes he'd reach into the freezer and empty out all the food. Other times he'd sled on the deck and get a splinter. His aunt Jane tried as hard as she could to keep him out of trouble, but he was sneaky and got into lots of trouble behind her back.
One day, James thought he would go into town and see what kind of trouble he could get into. He went to the grocery store and pulled all the pudding off the shelves and ate two jars. Then he walked to the fast food restaurant and ordered 15 bags of fries. He didn't pay, and instead headed home.
His aunt was waiting for him in his room. She told James that she loved him, but he would have to start acting like a well-behaved turtle.
After about a month, and after getting into lots of trouble, James finally made up his mind to be a better turtle.

1) What is the name of the trouble making turtle?
A) Fries
B) Pudding
C) James
D) Jane

2) What did James pull off of the shelves in the grocery store?
A) pudding
B) fries
C) food
D) splinters

Text QA and MRC Datasets

Dataset	Provider	Query Source	Answer	# Queries	# Docs
MC Test [Richardson+ 13]	Microsoft	Crowdsourced	Multiple Choice	2640	660
WikiQA [Yang+ 15]	Microsoft	User Logs	Sentence Selection	3047	29K sentences
CNN/DailyMail [Hermann+ 15]	DeepMind	Cloze	Fill in entity	1.4M	93K CNN, 220K DM
Children's Book [Hill+ 15]	Facebook	Cloze	Fill in the word	688K	688K contexts
SQuAD [Rajpurkat+ 16]	Stanford	Crowdsourced	Span	100K	536
News QA [Trischler+ 16]	Maluuba	Crowdsourced	Span	120K	12K
MS MARCO [Nguyen+ 16]	Microsoft	User Logs	Human Synthesized	100k	1M passages, 200K+ docs

Passage

(@entity4) if you feel a ripple in the force today , it may be the news that the official @entity6 is getting its first gay character . according to the sci-fi website @entity9 , the upcoming novel " @entity11 " will feature a capable but flawed @entity13 official named @entity14 who " also happens to be a lesbian . " the character is the first gay figure in the official @entity6 -- the movies , television shows , comics and books approved by @entity6 franchise owner @entity22 -- according to @entity24 , editor of " @entity6 " books at @entity28 imprint @entity26 .

Question	Answer
characters in " @placeholder " movies have gradually become more diverse	@entity6

Figure 1: An example item from dataset *CNN*.

Text QA and MRC Datasets

Dataset	Provider	Query Source	Answer	# Queries	# Docs
MC Test [Richardson+ 13]	Microsoft	Crowdsourced	Multiple Choice	2640	660
WikiQA [Yang+ 15]	Microsoft	User Logs	Sentence Selection	3047	29K sentences
CNN/DailyMail [Hermann+ 15]	DeepMind	Cloze	Fill in entity	1.4M	93K CNN, 220K DM
Children's Book [Hill+ 15]	Facebook	Cloze	Fill in the word	688K	688K contexts
SQuAD [Rajpurkat+ 16]	Stanford	Crowdsourced	Span	100K	536
News QA [Trischler+ 16]	Maluuba	Crowdsourced	Span	120K	12K
MS MARCO [Nguyen+ 16]	Microsoft	User Logs	Human Synthesized	100k	1M passages, 200K+ docs

Where did Tesla live for much of his life?

Tesla was renowned for his achievements and showmanship, eventually earning him a reputation in popular culture as an archetypal "mad scientist". His patents earned him a considerable amount of money, much of which was used to finance his own projects with varying degrees of success.:121,154 He lived most of his life in a series of New York hotels, through his retirement. Tesla died on 7 January 1943. His work fell into relative obscurity after his death, but in 1960 the General Conference on Weights and Measures named the SI unit of magnetic flux density the tesla in his honor. There has been a resurgence in popular interest in Tesla since the 1990s.

Text QA and MRC Datasets

Dataset	Provider	Query Source	Answer	# Queries	# Docs
MC Test [Richardson+ 13]	Microsoft	Crowdsourced	Multiple Choice	2640	660
WikiQA [Yang+ 15]	Microsoft	User Logs	Sentence Selection	3047	29K sentences
CNN/DailyMail [Hermann+ 15]	DeepMind	Cloze	Fill in entity	1.4M	93K CNN, 220K DM
Children's Book [Hill+ 15]	Facebook	Cloze	Fill in the word	688K	688K contexts
SQuAD [Rajpurkat+ 16]	Stanford	Crowdsourced	Span	100K	536
News QA [Trischler+ 16]	Maluuba	Crowdsourced	Span	120K	12K
MS MARCO [Nguyen+ 16]	Microsoft	User Logs	Human Synthesized	100k	1M passages, 200K+ docs

Will I qualify for OSAP if I'm new in Canada?

Selected Passages

"Visit the OSAP website for application deadlines. To get OSAP, you have to be eligible. You can apply using an online form, or you can print off the application forms. If you submit a paper application, you must pay an application fee. The online application is free."

Source: http://settlement.org/ontario/education/colleges-universities-and-institutes/financial-assistance-for-post-secondary-education/how-do-i-apply-for-the-ontario-student-assistance-program-osap/

"To be eligible to apply for financial assistance from the Ontario Student Assistance Program (OSAP), you must be a: 1 Canadian citizen; 2 Permanent resident; or 3 Protected person/convention refugee with a Protected Persons Status Document (PPSD)."

Source: http://settlement.org/ontario/education/colleges-universities-and-institutes/financial-assistance-for-post-secondary-education/who-is-eligible-for-the-ontario-student-assistance-program-osap/

"You will not be eligible for a Canada-Ontario Integrated Student Loan, but can apply for a part-time loan through the Canada Student Loans program. There are also grants, bursaries and scholarships available for both full-time and part-time students."

Source: http://www.campusaccess.com/financial-aid/osap.html

Answer

No. You won't qualify.

QA on Text

Query	Who was the #2 pick in the 2011 NFL Draft?
Passage	Manning was the #1 selection of the 1998 NFL draft, while Newton was picked first in 2011. The matchup also pits the top two picks of the 2011 draft against each other: Newton for Carolina and **Von Miller** for Denver.
Answer	**Von Miller**

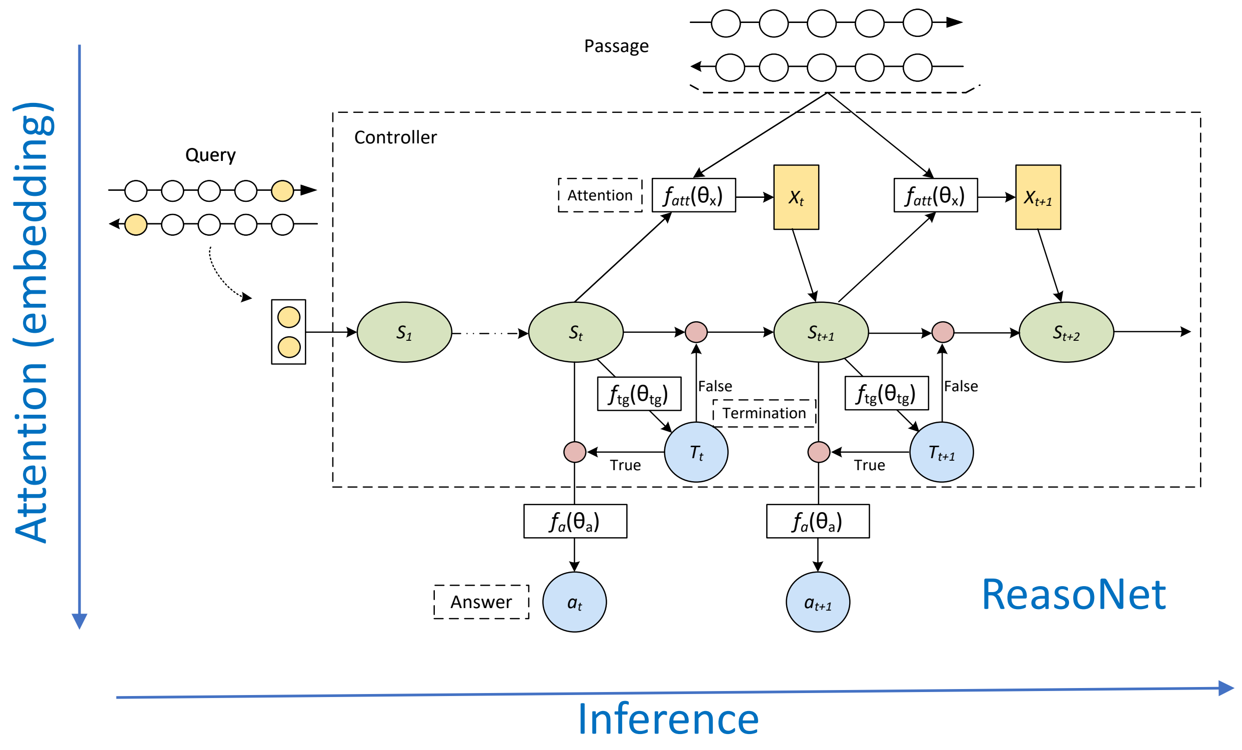
Multi-step inference:

- Step 1:
  - **Extract:** Manning is #1 pick of 1998
  - **Infer:** Manning is NOT the answer
- Step 2:
  - **Extract:** Newton is #1 pick of 2011
  - **Infer:** Newton is NOT the answer
- Step 3:
  - **Extract:** Newton and Von Miller are top 2 picks of 2011
  - **Infer:** **Von Miller** is the #2 pick of 2011

The Text QA Results using MRC models

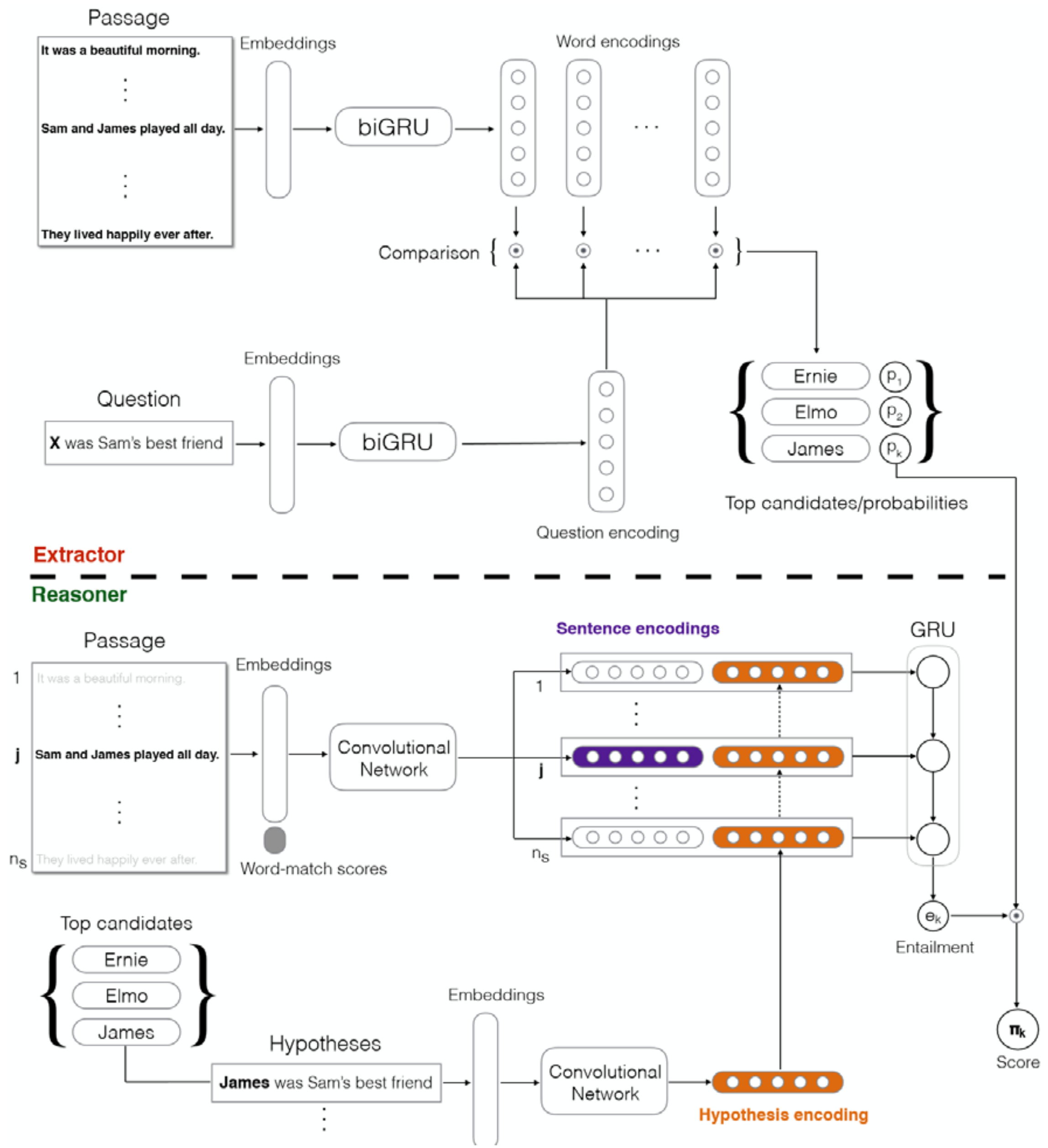
Models	CNN	Daily Mail	SQuAD (EM/F1)
G/DeepMind: Attentive Reader [Hermann+ 15]	63.0	69.0	
IBM: Attention Sum Reader [Kadlec+ 16]	69.5	73.9	
Stanford AR [Chen+ 16]	72.4	75.8	
(MS) Maluuba: Iterative AR [Sordoni+ 16]	**73.3**	**-**	
(MS) Maluuba: EpiReader [Trischler+ 16]	**74.0**	**-**	
CMU: GA Reader [Dhingra+ 16]	73.8	75.7	
MSR: ReasoNet [Shen+ 16] (Sep 17 2016)	**74.7**	**76.6**	
Google/UW: RaSoR [Lee+ 16] (Nov 4 2016)	-	-	69.6 / 77.7
AI2/UW: BiDAF [Seo+ 16] (Nov 5 2016)	77.1	78.3	73.7 /81.5
MSR: ReasoNet [Shen+ 16] (Mar 2017)	**-**	**-**	**75.0 / 82.6**
MSRA: R-net [Wang+ 17] (Jun 2017)	**-**	**-**	**77.7 / 84.7**

https://rajpurkar.github.io/SQuAD-explorer/



Model	Attention (Embedding)	Inference
Maluuba: EpiReader [Trischler+ 16]	**Attention sum reader**	Single-step
Maluuba: Iterative AR [Sordoni+ 16]	Attention sum reader	Multi-step, step size is predefined
BiDAF [Seo+ 16]	**Co-attention**	Single-step
MSR: ReasoNet [Shen+ 16b]	Co-attention	**Dynamic multi-step (step size is determined based on complexity of queries on the fly)**
MSRA: R-net [Wang+ 17]	**Gated attention + self matching**	Single-step

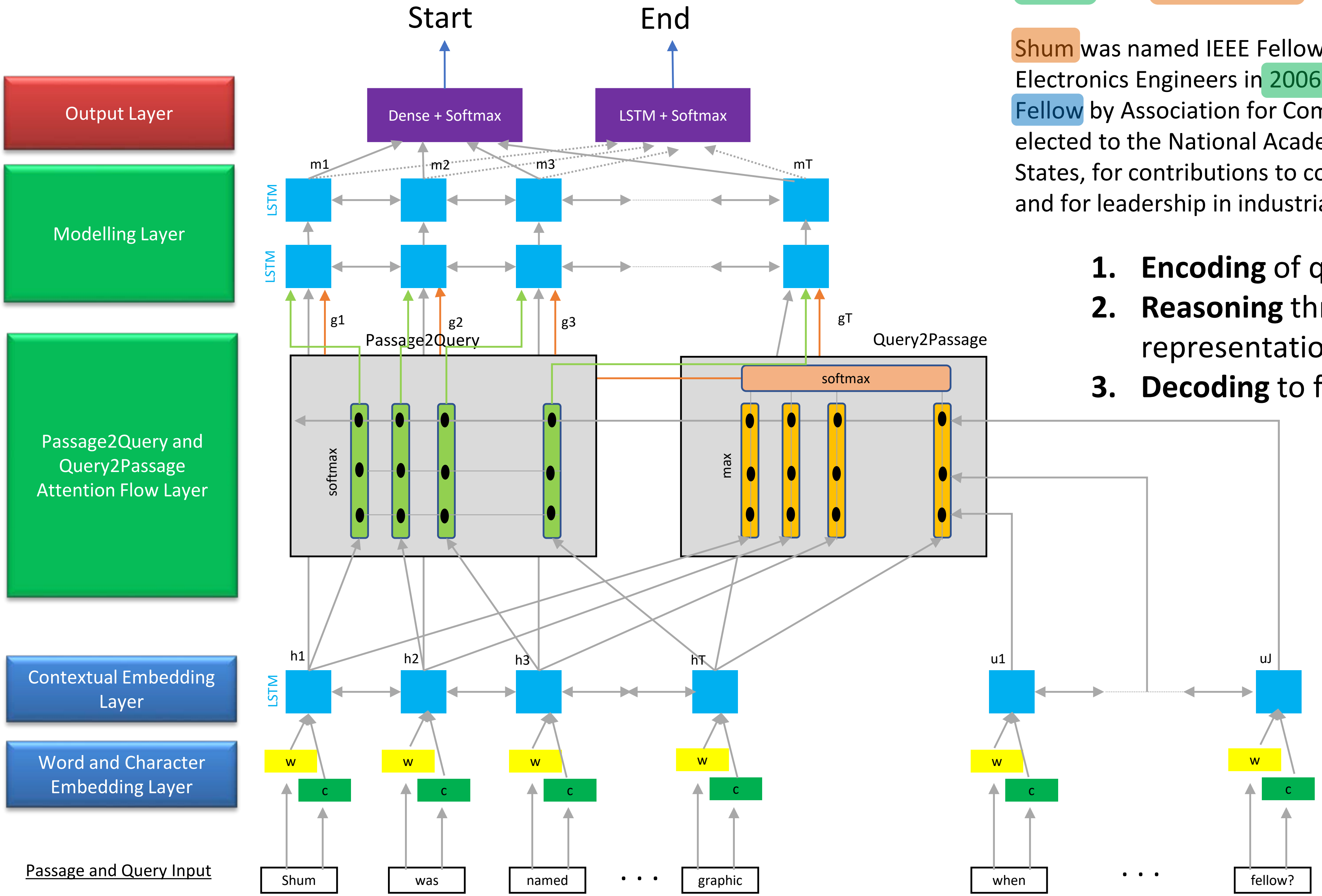
EpiReader: attention sum reader



Stage One
The **extractor** selects a small set of candidate answers for further processing

Stage Two
The **reasoner** uses the candidates to form hypothesis that are compared with the question to measure entailment

BiDAF: co-attention



when did harry shum become acm fellow?

Shum was named IEEE Fellow by Institute of Electrical and Electronics Engineers in 2006. In 2007, he was recognized as ACM Fellow by Association for Computing Machinery. In 2017, he was elected to the National Academy of Engineering (NAE) of the United States, for contributions to computer vision and computer graphics, and for leadership in industrial research and product development.

1. **Encoding** of query and passage
2. **Reasoning** through query aware passage representation (bidirectional attention)
3. **Decoding** to find start and end pointers

R-net: gated attention

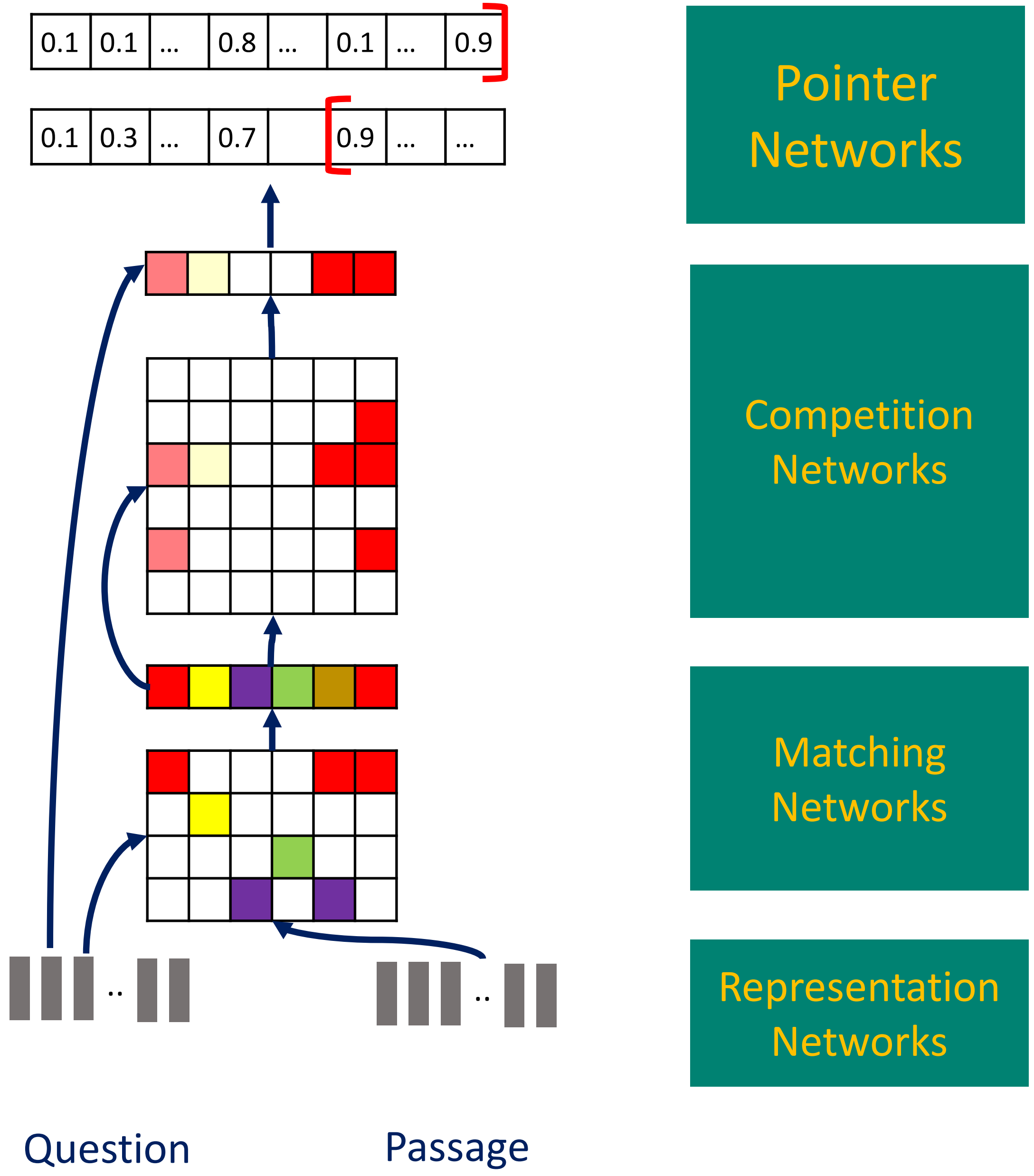
Answer boundary

Answer: “late 15th century”

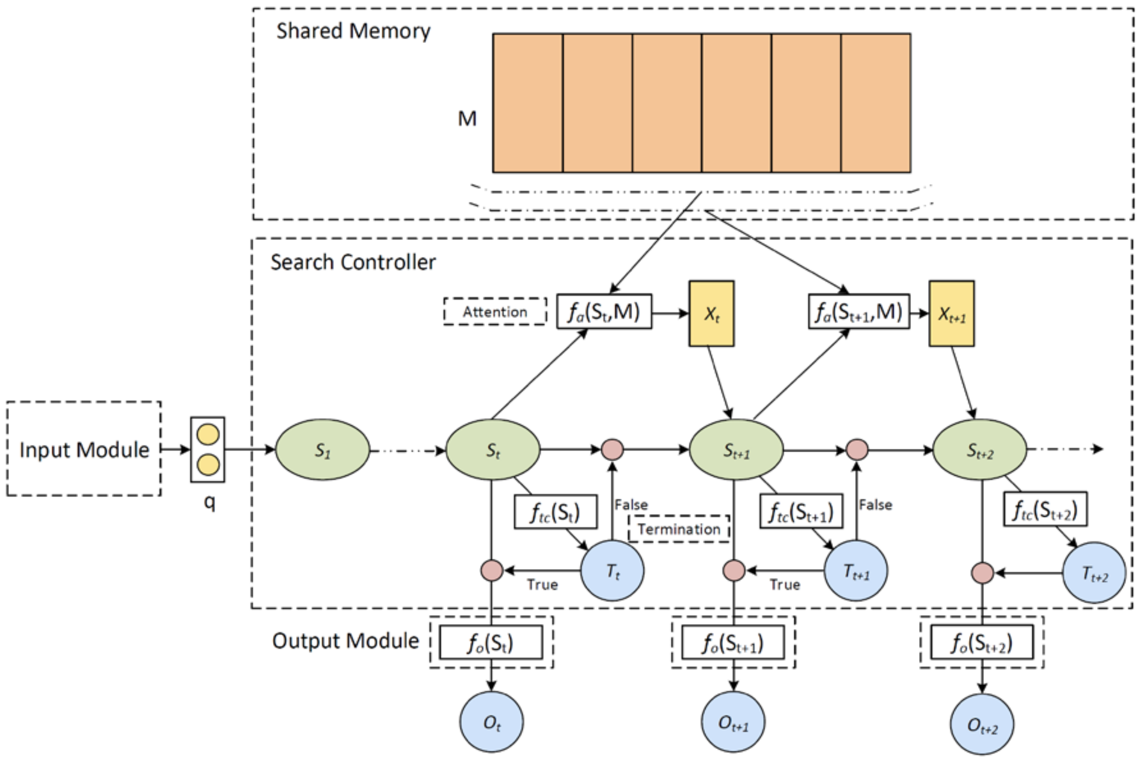
Answer candidates   Supporting Evidence

Ferdinand III had started out as a contested king of Castile. By the time of his death in 1252, Ferdinand III had delivered to his son and heir, Alfonso X, a massively expanded kingdom. The boundaries of the new Castilian state established by Ferdinand III would remain nearly unchanged until the late 15th century.

When did Castilian border change after Ferdinand III’s death?



ReasoNet with Shared Memory

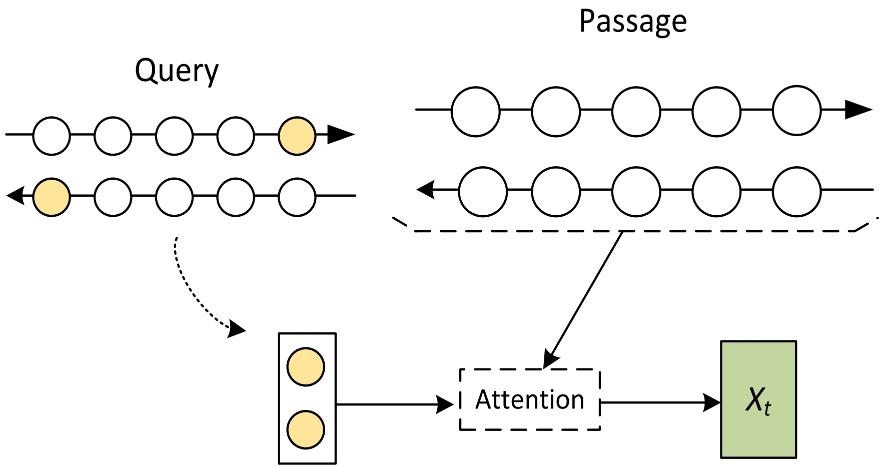


- Shared memory encodes task-specific knowledge (e.g., passage or KB)
- Search controller performs multi-step inference to generate answer of a question using knowledge in shared memory
- Shared memory and search controller are jointly learned via SL+RL
- Input/output modules are task-specific

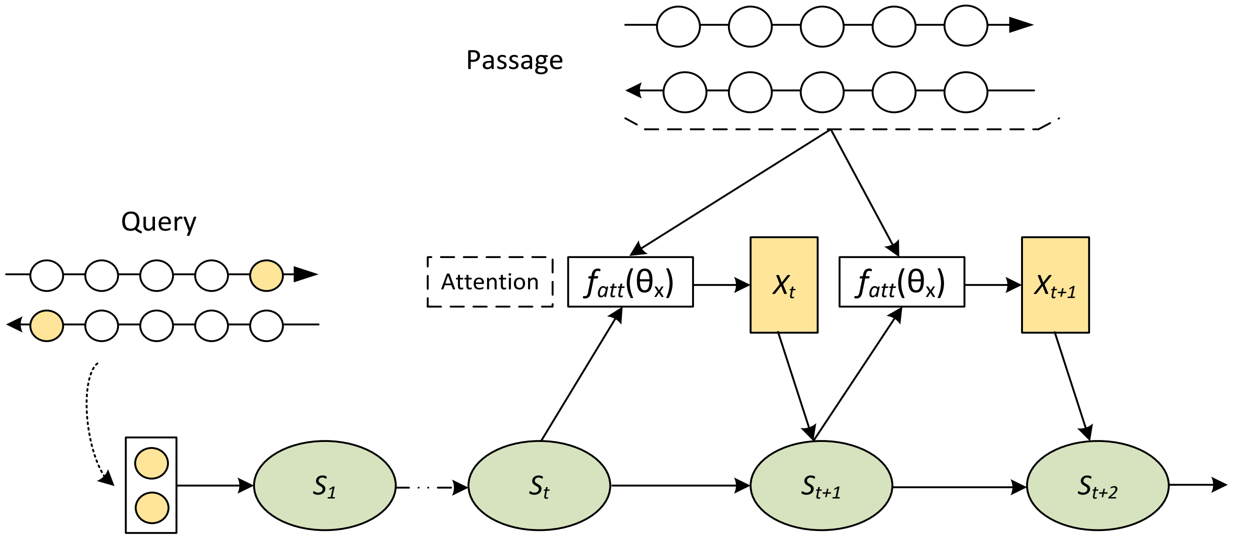
[Shen+ 16a]

Inference engines for MRC

Single Step Inference



Multiple Step Inference



How many steps?

Search Control: multi-step inference engine

- Learning to stop reading: dynamic multi-step inference
- Step size is determined based on the complexity of instance (QA pair)

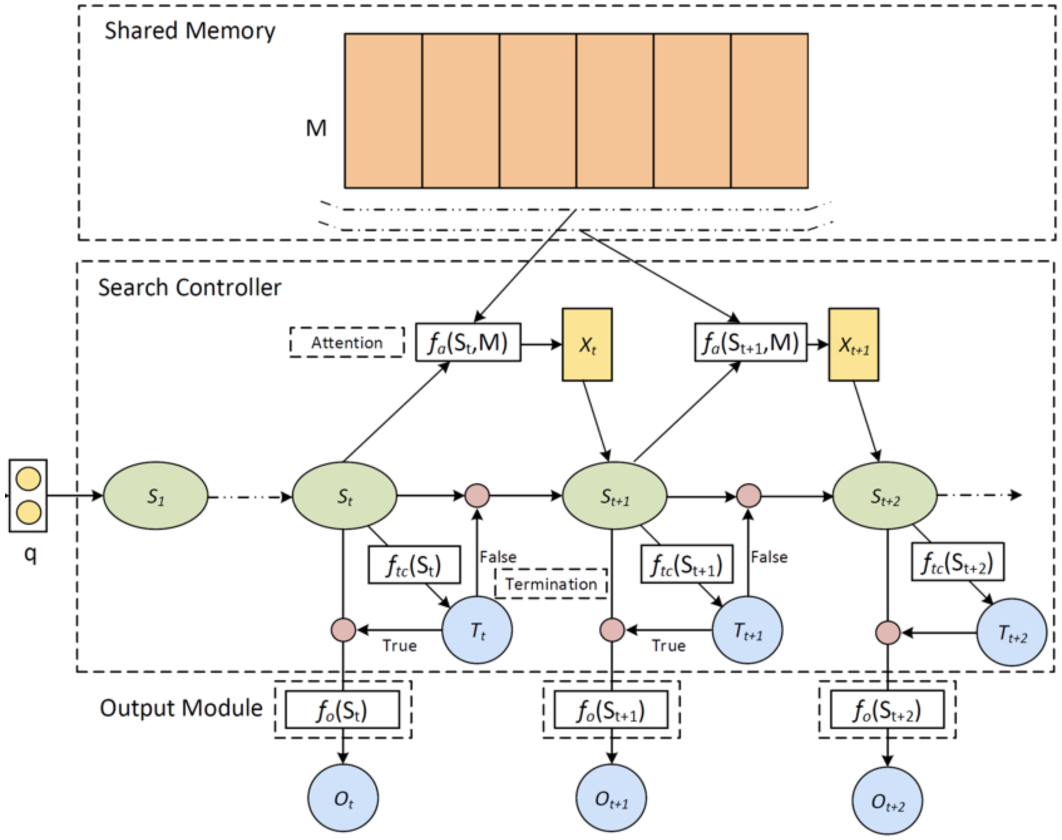
Query	Who was the 2015 NFL MVP?
Passage	The Panthers finished the regular season with a 15–1 record, and quarterback **Cam Newton** was named the 2015 NFL Most Valuable Player (MVP).
Answer (1-step)	Cam Newton

Query	Who was the #2 pick in the 2011 NFL Draft?
Passage	Manning was the #1 selection of the 1998 NFL draft, while Newton was picked first in 2011. The matchup also pits the top two picks of the 2011 draft against each other: Newton for Carolina and **Von Miller** for Denver.
Answer (3-step)	Von Miller

ReasoNet: Learn to Stop Reading

Keep gathering information (encoded in internal state) until a good answer is formed

1. Given a set of docs in memory: $\mathbf{M}$
2. Start with query: S
3. Identify info in $\mathbf{M}$ that is related to S : $X = f_a(S, \mathbf{M})$
4. Update internal state: $S = \mathrm{RNN}(S, X)$
5. Whether a satisfied answer O can be formed based on S: $f_{tc}(S)$
6. If so, stop and output answer $O = f_o(S)$; otherwise return to 3.



[Shen+ 16b]

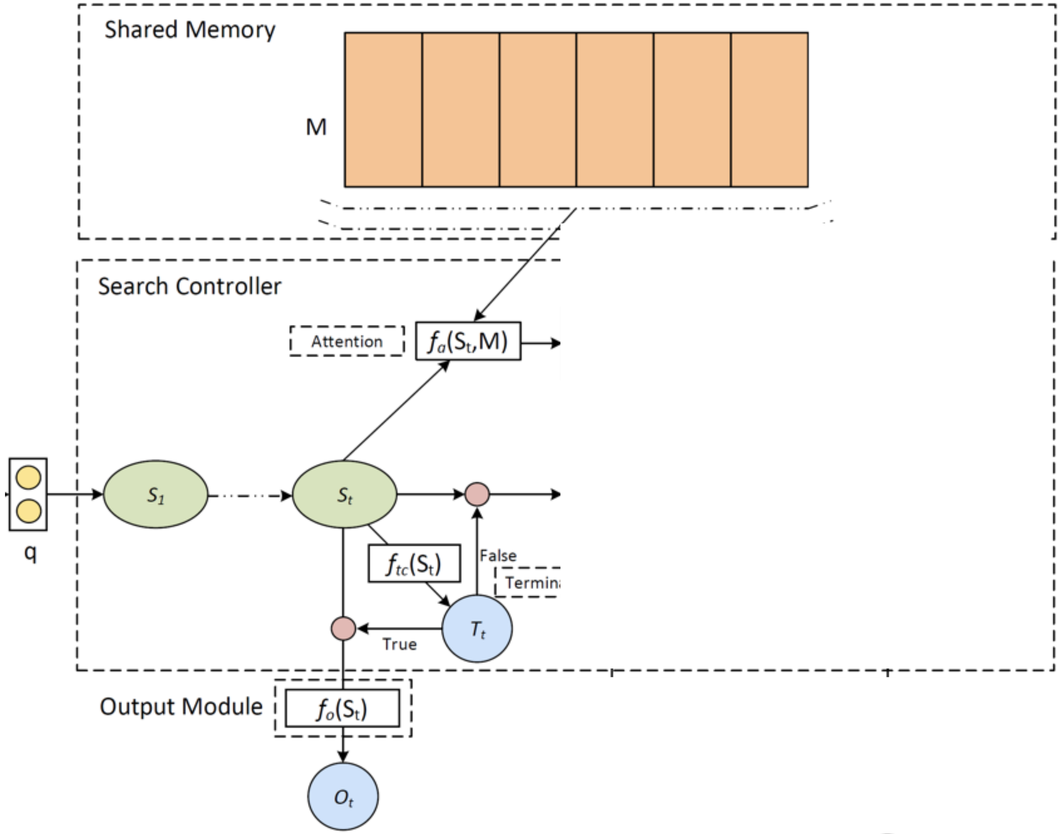
ReasoNet at work

Query	Who was the #2 pick in the 2011 NFL Draft?
Passage	Manning was the #1 selection of the 1998 NFL draft, while Newton was picked first in 2011. The matchup also pits the top two picks of the 2011 draft against each other: Newton for Carolina and Von Miller for Denver.
Answer	**Von Miller**

Rank-1

Rank-2

Rank-3

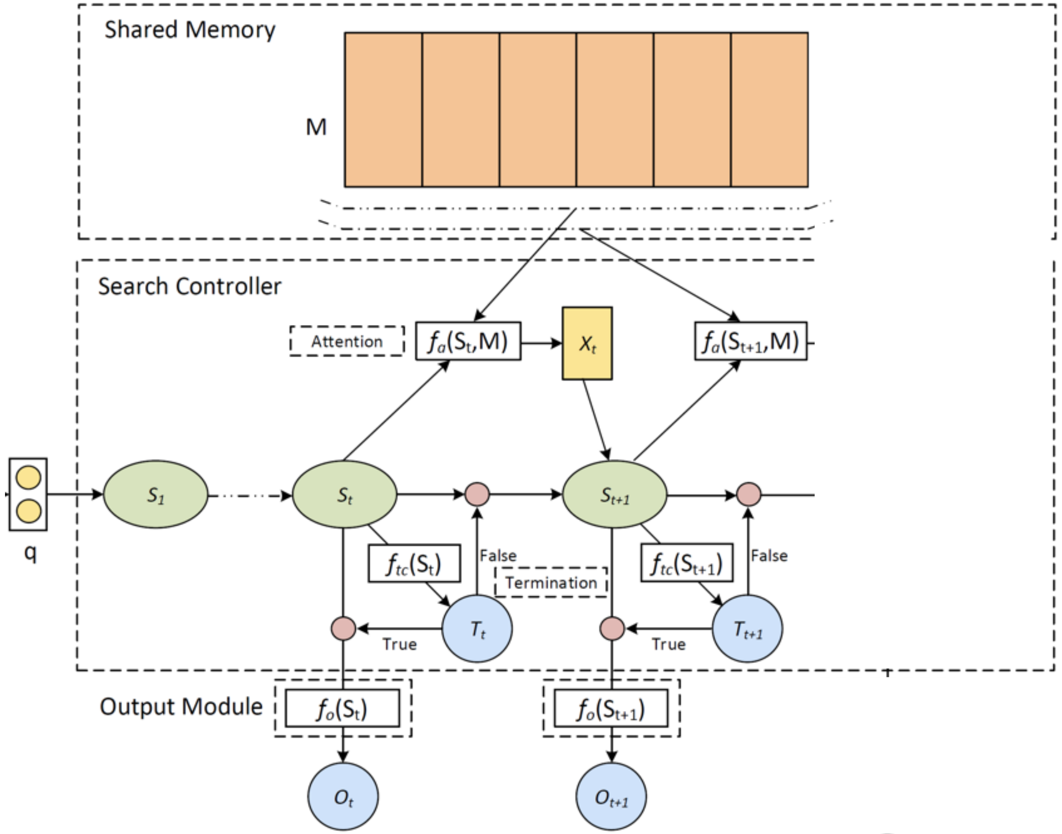


Step	Termination Probability	Prob. Answer
1	0.001	0.392

ReasoNet at work

Query	Who was the #2 pick in the 2011 NFL Draft?
Passage	Manning was the #1 selection of the 1998 NFL draft, while Newton was picked first in 2011. The matchup also pits the top two picks of the 2011 draft against each other: Newton for Carolina and Von Miller for Denver.
Answer	**Von Miller**

Rank-1
Rank-2
Rank-3

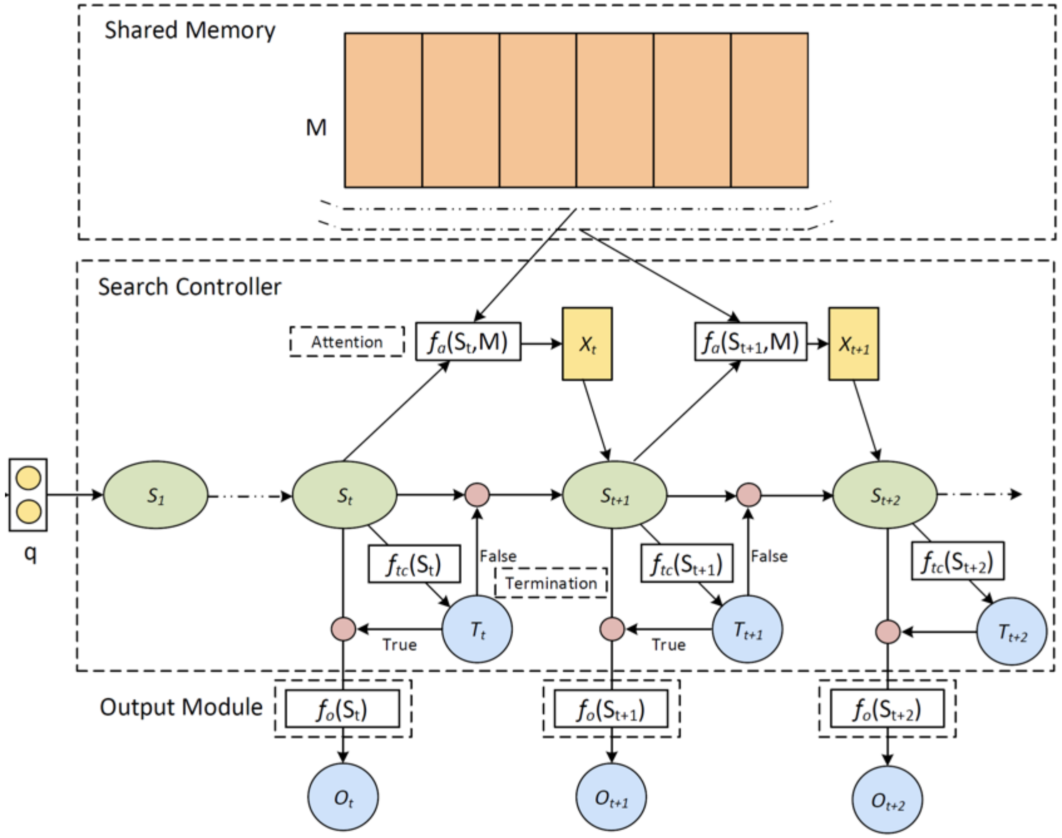


Step	Termination Probability	Prob. Answer
1	0.001	0.392
2	0.675	0.649

ReasoNet at work

Query	Who was the #2 pick in the 2011 NFL Draft?
Passage	Manning was the #1 selection of the 1998 NFL draft, while Newton was picked first in 2011. The matchup also pits the top two picks of the 2011 draft against each other: Newton for Carolina and Von Miller for Denver.
Answer	**Von Miller**

Rank-1
Rank-2
Rank-3

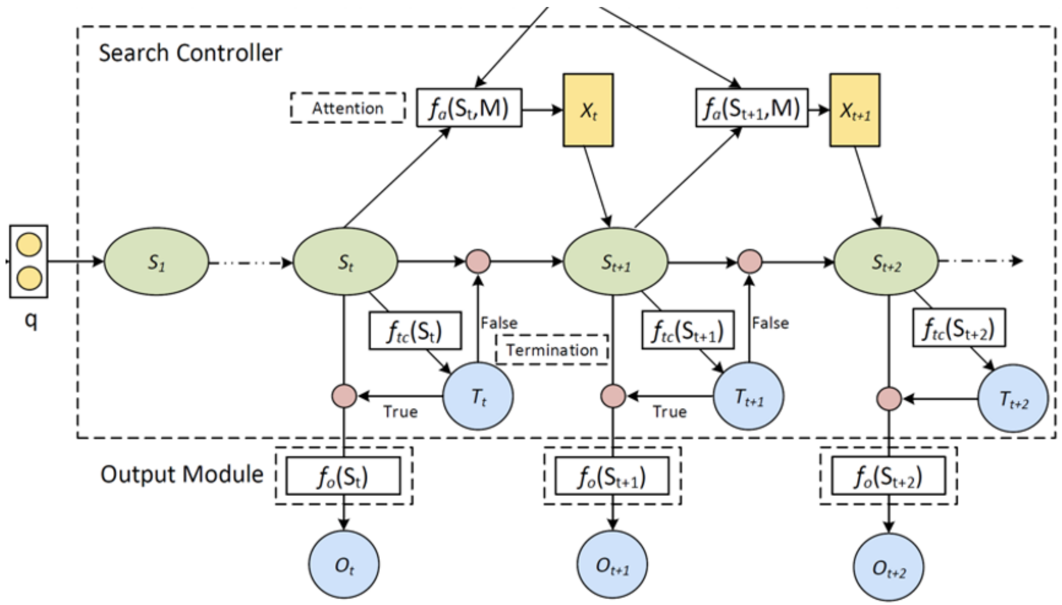


Step t	Termination Probability f_{tc}	Prob. Answer f_o
1	0.001	0.392
2	0.675	0.649
3	0.939	0.865

Training ReasoNet via reinforcement learning objectives

- Action: reading or termination/answer
- Reward: 1 if the answer is correct, 0 otherwise (Delay Reward)
- Expected total reward

$$J(\theta) = \mathbb{E}_{\pi(t_{1:T}, a_T; \theta)} \left[\sum_{t=1}^{T} r_t \right]$$

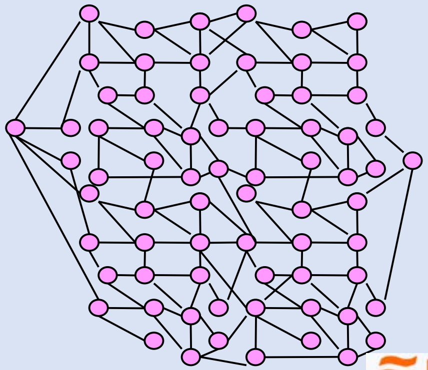


- REINFORCE algorithm

$$\nabla_\theta J(\theta) = \sum_{(t_{1:T}, a_T) \in \mathbb{A}^\dagger} \pi(t_{1:T}, a_T; \theta) \left[\nabla_\theta \log \pi(t_{1:T}, a_T; \theta)(r_T - b) \right]$$

$b = \sum_{(t_{1:T}, a_T) \in \mathbb{A}^\dagger} \pi(t_{1:T}, a_T; \theta) r_T$ **Instance-based baseline**

ReasoNet: joint learning of Shared Memory and Search Controller

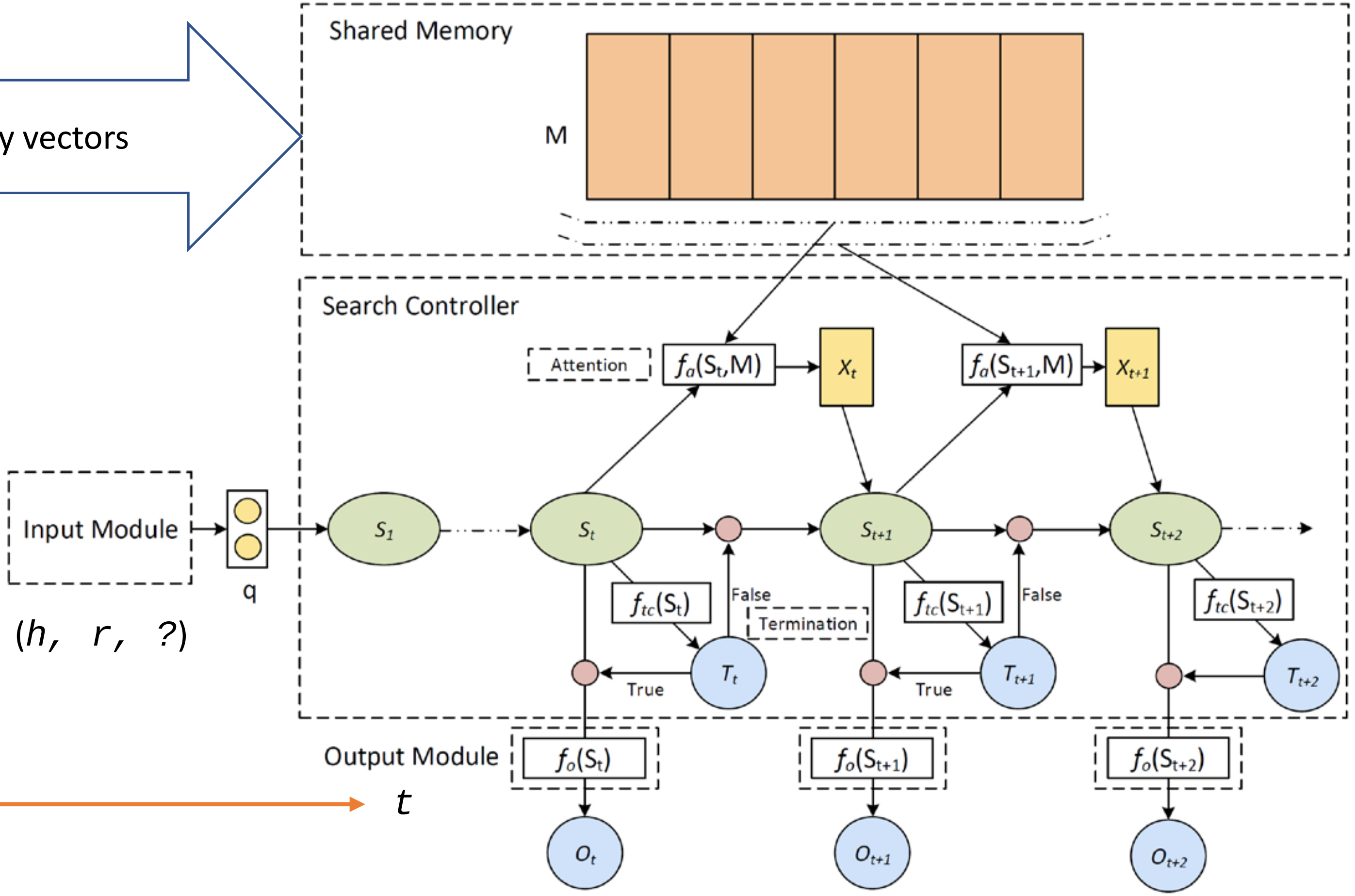


Freebase

Embed KG to memory vectors

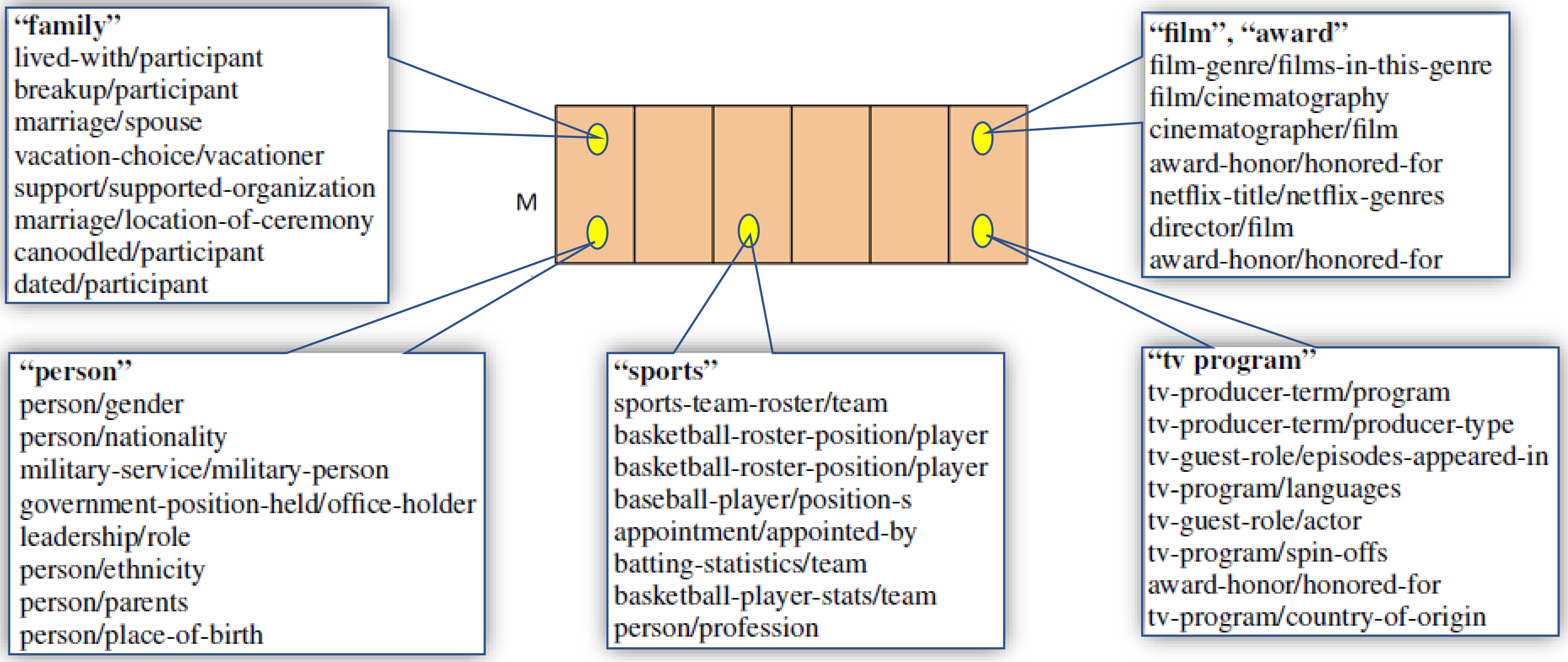
Training samples from KG:

(Obama, BornIn, Hawaii)
(Hawaii, PartOf, USA)
…
(h, r, t)
…
(Obama, Citizenship,?)->(USA)



Shared Memory: long-term memory to store learned knowledge, like human brain

- Knowledge is learned via performing tasks, e.g., update memory to answer new questions
- New knowledge is *implicitly* stored in memory cells via gradient update
- Semantically relevant relations/entities can be compactly represented using similar vectors.



Interim summary

- Symbolic approaches to QA
  - Knowledge representation and search in a symbolic space
  - A case study of MSR MindNet
- Neural approaches to MRC and QA
  - Knowledge representation and search in a neural space
  - A case study of ReasoNet
  - Learn more at [Deep Learning for Machine Reading Comprehension]
- Ongoing research
  - Neural approaches to symbolic reasoning
  - Interpret or visualize the reasoning process in neural space

Tutorial Outline

- Part 1: Background
- Part 2: Deep Semantic Similarity Models for text processing
- Part 3: Recurrent neural networks for text generation
- Part 4: Neural machine reading models for question answering
- **Deep reinforcement learning for goal-oriented dialogue**
  - **What kinds of Problems?**
  - **Reinforcement learning (RL) vs. supervised learning (SL)**
  - **Deep RL for dialogues**
  - **Three case studies**

What kinds of problems?

“I am smart”	Turing Test (“I” talk like a human)
“I have a question”	Information consumption
“I need to get this done”	Task completion
“What should I do?”	Decision support

What kinds of problems?

"I am smart"	Turing Test
"I have a question"	Information consumption
"I need to get this done"	Task completion
"What should I do?"	Decision support

- *What is the employee review schedule?*
- *What room is the project review meeting in?*
- *When is the ACL 2017 conference?*
- *What does AGI stand for?*

What kinds of problems?

"I am smart"	Turing Test
"I have a question"	Information consumption
"I need to get this done"	Task completion
"What should I do?"	Decision support

- *Book me the biz trip to San Francisco*
- *Reserve a table at Kisaku for 5 people, 7PM tonight*
- *Brief me on people in my Thursday 9:00 am meeting*
- *Schedule a meeting with Bill at 10:00 tomorrow.*

What kinds of problems?

"I am smart"	Turing Test
"I have a question"	Information consumption
"I need to get this done"	Task completion
"What should I do?"	Decision support

- *Why are sales in China so far behind forecast?*

What kinds of problems?

"I am smart"	Turing Test ("I" talk like a human)
"I have a question"	Information consumption
"I need to get this done"	Task completion
"What should I do?"	Decision support

Goal-oriented dialogues

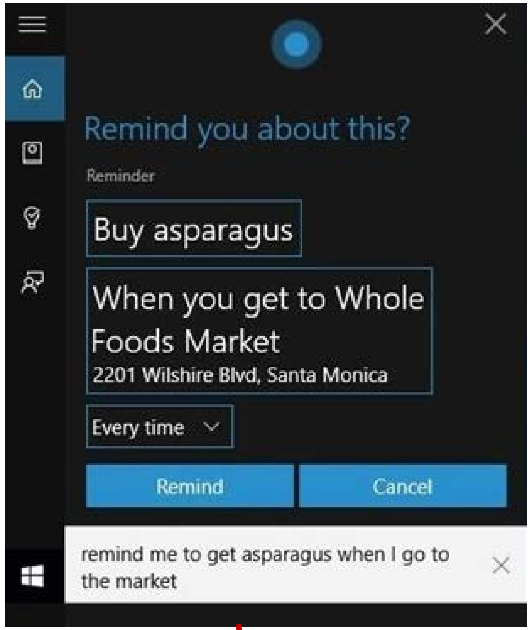
Personal assistants today

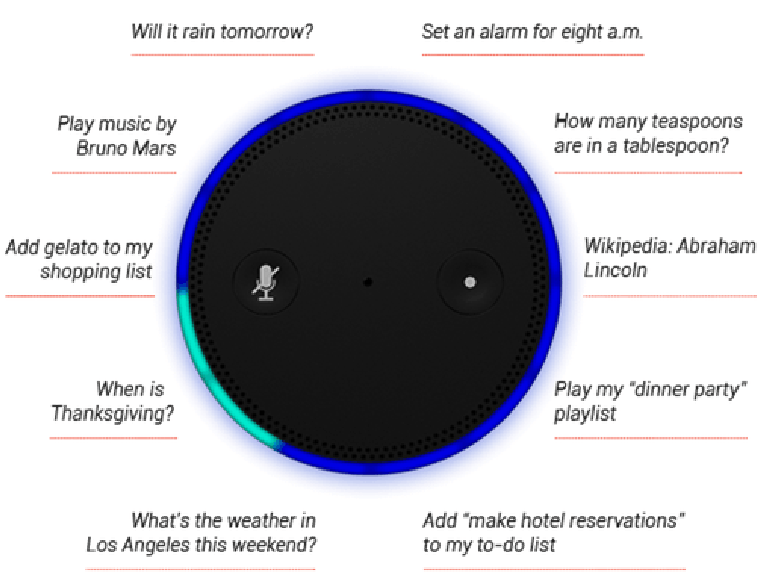
Google Now

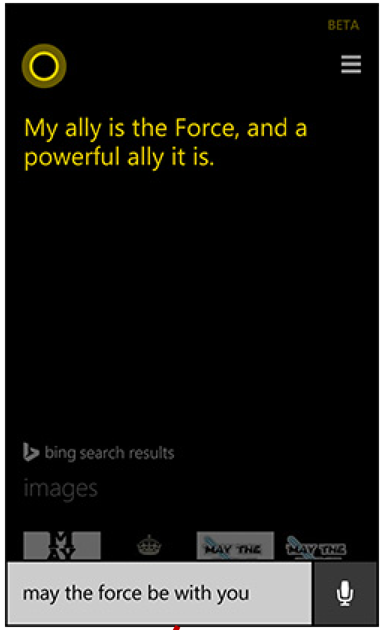
Siri

Cortana

amazon echo







goal oriented

Engaging (social bots)

Aspirational Goal: Enterprise Assistant

Where are sales lagging behind our forecast?

Task Completion

The worst region is [country], where sales are XX% below projections

Do you know why?

Decision Support

The forecast for [product] growth was overly optimistic

How can we turn this around?

Here are the 10 customers in [country] with the most growth potential, per our CRM model

Can you set up a meeting with the CTO of [company]?

Info Consumption

Task Completion

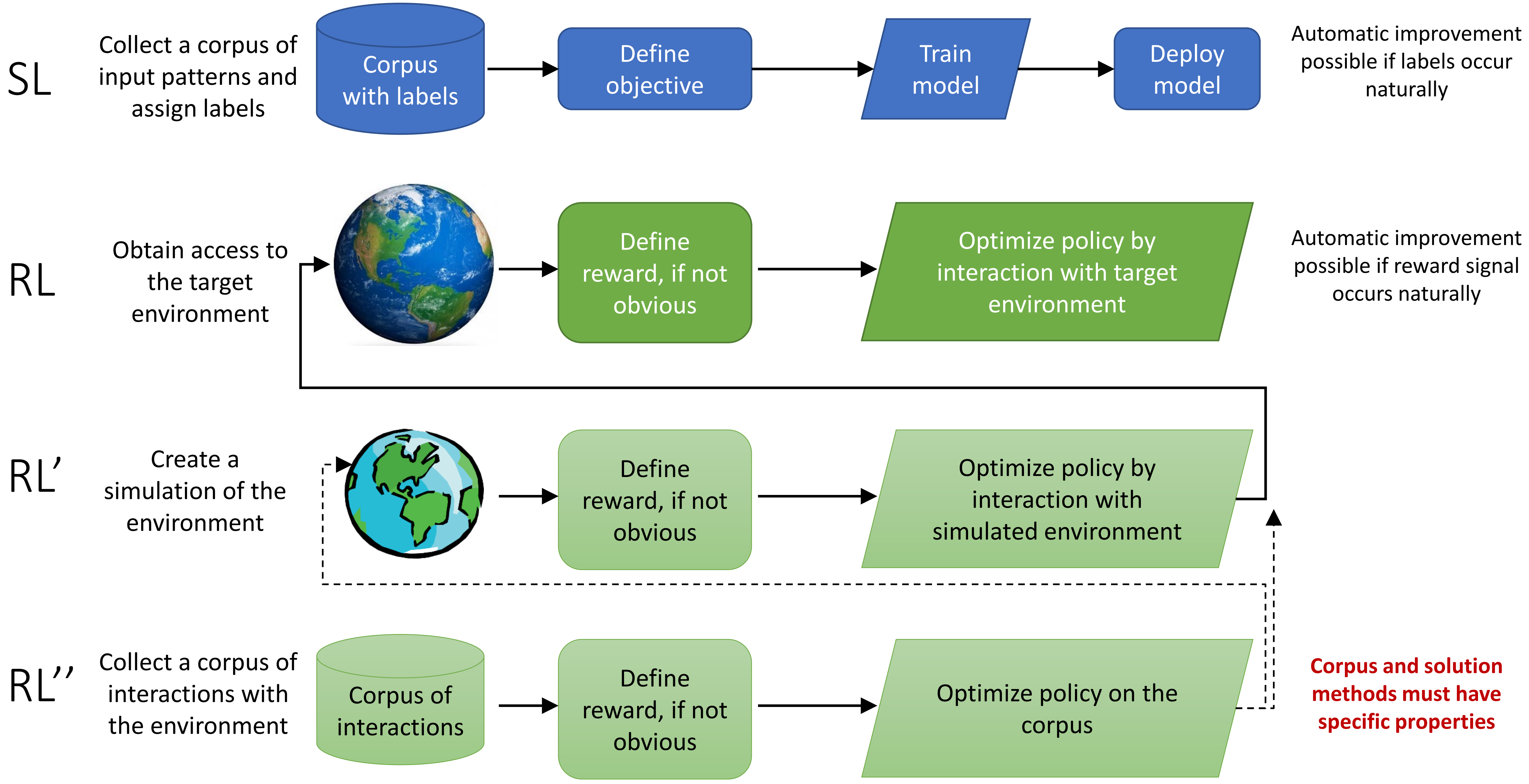
Yes, I’ve set up a meeting with [person name] for next month when you’re in [location]

Thanks

Supervised Learning (SL) vs. Reinforcement Learning (RL)

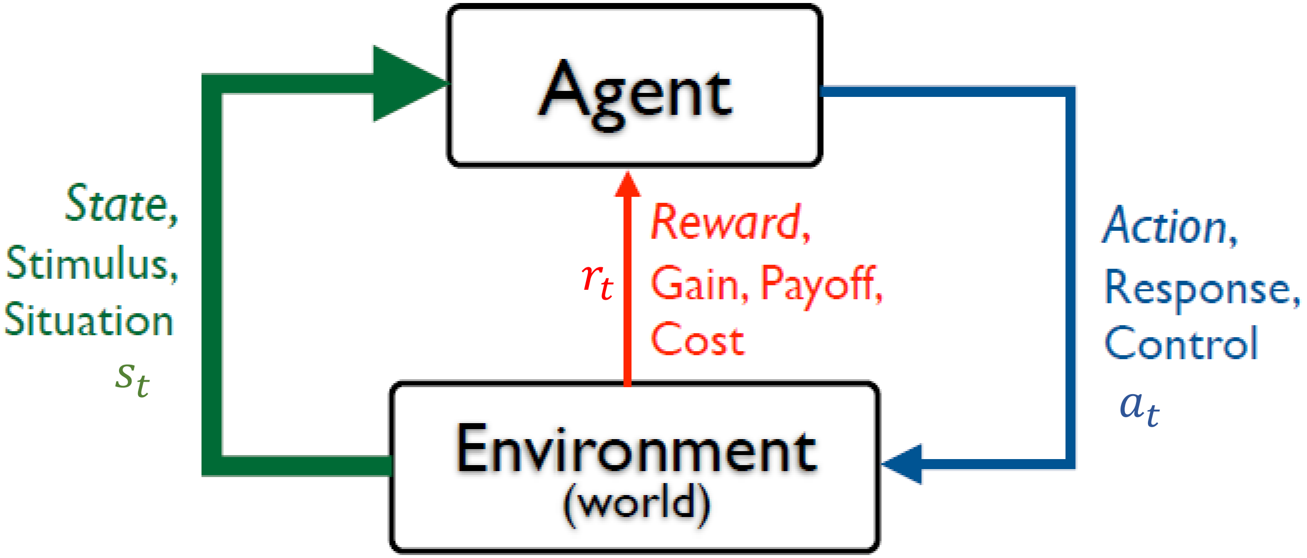
- Distinct methods of learning from experience
- SL – learning from previous experience
  - Learning a model on collected input-output pairs (training data),
  - by minimizing some loss functions
  - No explicit dependence on how training data is collected
- RL – learning by experiencing
  - An agent learned by interacting with an environment to achieve a goal
  - Learning by trial and error (exploration) with only delayed reward
  - Can tell for itself when it is right or wrong
- RL is more realistic, natural and ambitious than SL

SL vs. RL



[Slide from Lihong Li and Jason Williams]

The RL Interface



- Environment may be unknown, nonlinear, stochastic and complex
  - Learning to best represent the state-action space
- Agent learns a policy mapping states to actions, $a = \pi(s)$
  - So as to maximize its cumulative reward in the long run, $R = \sum_t \gamma^{t-1} r_t$

[Sutton & Barto 98]

Challenges of RL: Learning via Experiencing

- Complex, (unbounded) state-action space
- Evaluation feedback, (delayed) reward
- Non-stationarity
- Need for trial and error, to explore as well as exploit
  - how an agent can learn from success and failure, from reward and punishment
  - *one constantly has to decide btw continuing in a comfortable existence and striking out into unknown in the hopes of discovering a new and better life.*

A Finite Markov Decision Process (MDP)

- Discrete time $t = 1,2,3, \ldots$
- A finite set of states, s
- A finite set of actions, a
- A finite set of rewards, r
- Life is a trajectory: $\ldots s_t, a_t, r_{t+1}, s_{t+1}, a_{t+1}, r_{t+2}, s_{t+2} \ldots$

- RL task: search for a policy π,
  - according to which the agent chooses each action $a_t = \pi(s_t)$,
  - so as to maximize the long-term rewards (discounted sum of future rewards):
    $\pi^* = \underset{\pi}{\operatorname{argmax}} \mathbf{E}[\sum_t \gamma^{t-1} r_t \,|\pi]$

[Howard 60; Sutton & Barto 98]

RL algorithms: A simplified overview

	Model is known (aka "planning")	**Model is unknown (Access to environment/corpus, but not P or R)**
Value-function based (learn Q^* first, then π^*)	Value iteration Policy iteration Linear programming Monte Carlo tree search	Approx. value iteration (**Q-learning**, Sarsa, …) Approx. policy iteration (LSPI, …) Monte Carlo estimation …
Direct policy search (learn π^* directly)		**Policy gradient** …

[Slide from Lihong Li and Jason Williams]

Action-value function, $Q(s,a)$

- $Q(s,a)$: the value of taking action a in state s
- Given the optimal $Q^*(s,a)$ for all (s,a), the optimal policy is

$$\pi^*(s) = \operatorname*{argmax}_a Q^*(s,a)$$

- Bellman expectation equation:

$$\begin{aligned} Q(s,a) &= E[r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots | s_t = s, a_t = a] \\ &= E[r_{t+1} + \gamma Q(s_{t+1}, a_{t+1}) | s_t = s, a_t = a] \end{aligned}$$

- **Bellman optimality equation:**

$$\boldsymbol{Q^*(s,a) = E[r_{t+1} + \underbrace{\gamma \max_{a'} Q^*(s_{t+1}, a')}_{\text{Q-learning's target for } Q(s,a)} | s_t = s, a_t = a]}$$

Q-Learning

- Assume $Q(s, a)$ for all (s, a) can be represented in a table

1. Initialize an array $Q(s, a)$ arbitrarily
2. Choose actions based on Q such that all actions are taken in all states (infinitely often in the limit)
3. On each time step, update one element of the array:

$$\Delta Q(s_t, a_t) = \alpha \left(\underbrace{r_{t+1} + \gamma \left(\max_{a'} Q(s_{t+1}, a') \right)}_{\text{Q-learning's target for } Q(s,a)} - Q(s_t, a_t) \right)$$

- Model-free learning:
  - Learning long-term optimal behavior without model of the environment
  - All we need is the sample set of $(s_t, a_t, r_{t+1}, s_{t+1})$

[Sutton & Barto 98]

Function Approximation

- In many tasks, (s, a) is too large for tabular representation
- Estimate the action-value function approximately as $Q(s,a;\theta) \approx Q^*(s,a)$
- θ: a linear function (baseline)
- θ: a DNN, aka Deep Q-Network (DQN)
  - Optimize θ using SGD w.r.t loss

$$L_i(\theta_i) = \mathbb{E}_{s,a\sim\rho(\cdot)}\left[(y_i - Q(s,a;\theta_i))^2\right]$$

$$y_i = \mathbb{E}_{s'\sim\mathcal{E}}\left[r + \gamma \max_{a'} Q(s',a';\theta_{i-1}) | s,a\right]$$

$$\nabla_{\theta_i} L_i(\theta_i) = \mathbb{E}_{s,a\sim\rho(\cdot);s'\sim\mathcal{E}}\left[\left(r + \gamma \max_{a'} Q(s',a';\theta_{i-1}) - Q(s,a;\theta_i)\right)\nabla_{\theta_i} Q(s,a;\theta_i)\right]$$

Q-Learning for Deep Q-Network

- Issue: learning becomes unstable
  - Correlations present in the sequence of states
  - Small updates to Q leads to significant change of policy and data distribution
  - Correlations btw the to-be-learned Q and the target value $r + \max_{a'} Q(s, a')$
- Solution
  - Experience replay: randomize training samples (s, a, r, s')
  - Use a separate Q function to compute targets y

[DeepMind 15]

Policy gradient: RL as optimization

- Let $J(\theta)$ be any policy objective function
- Search for a local maximum in $J(\theta)$ by ascending the gradient of the policy w.r.t. parameters θ

$$\Delta\theta = \alpha \nabla_\theta J(\theta)$$

- Where $\nabla_\theta J(\theta)$ is the policy gradient

$$\nabla_\theta J(\theta) = \begin{cases} \dfrac{\partial J(\theta)}{\partial \theta_1} \\ \cdots \\ \cdots \\ \dfrac{\partial J(\theta)}{\partial \theta_n} \end{cases}$$

- and α is the learning rate.

Estimating gradient of a stochastic policy

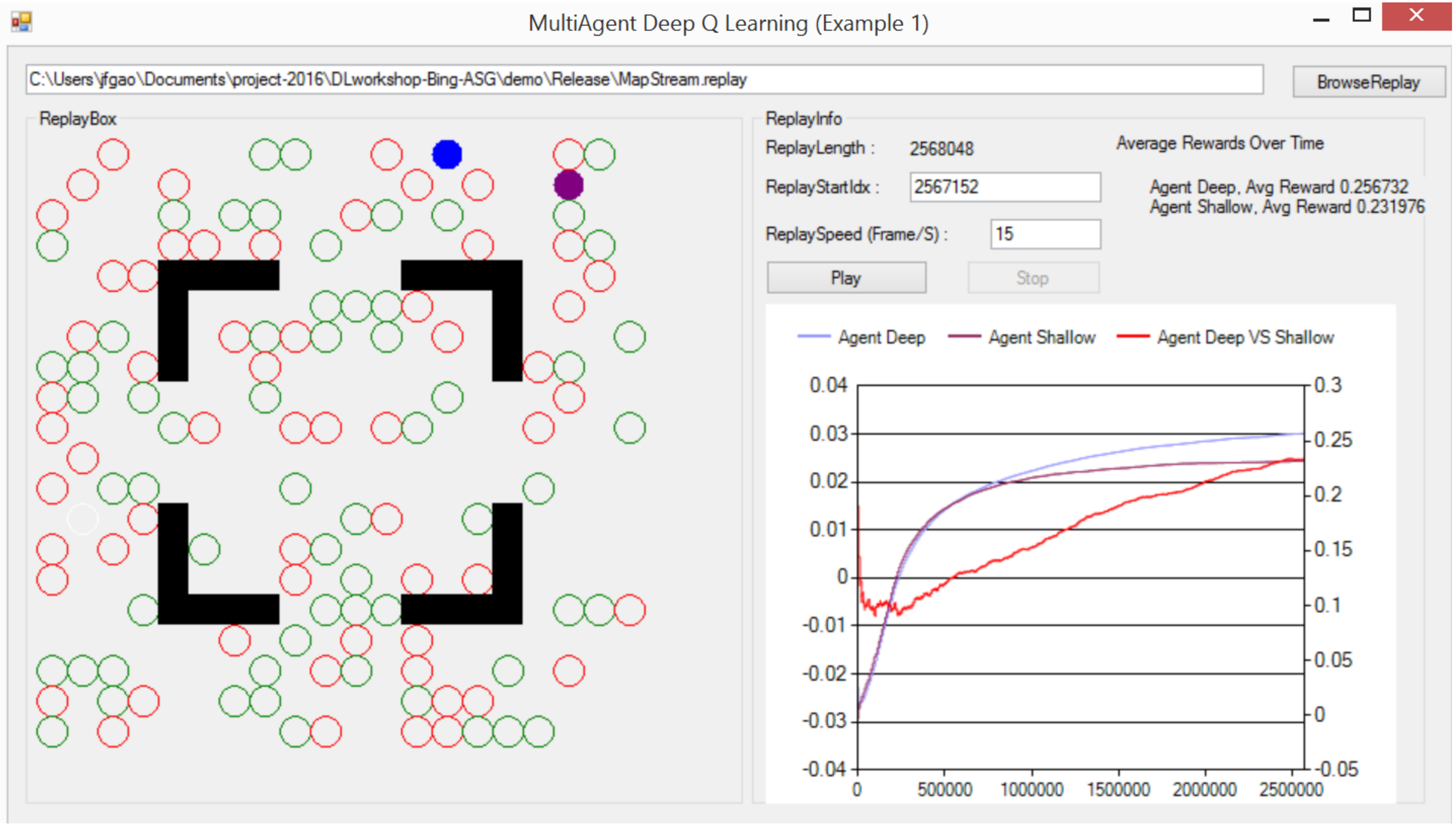
- Let $\pi(s, a|\theta)$ be a stochastic policy, e.g., $\pi(s, a|\theta) = \frac{\exp Q(s,a|\theta)}{\sum_{a'} \exp Q(s,a'|\theta)}$
- Consider the optimization of

$$\max_{\theta} J(\theta) = \max_{\theta} \mathbf{E}_{a\sim\pi(s,a|\theta)}[r(a)]$$

- Gradient of $J(\theta)$ can be estimated by

$$\begin{aligned}\nabla J(\theta) &= \nabla\left[\int \pi(s,a|\theta) r(a)\right] = \int \nabla\pi(s,a|\theta) r(a) \\ &= \int \pi(s,a|\theta)\nabla \log \pi(s,a|\theta)\, r(a) \\ &= E_{a\sim\pi(s,a|\theta)}\left[\nabla \log \pi(s,a|\theta)\, r(a)\right] \\ &\boxed{\approx \frac{1}{m}\sum_{i=1}^{m} \nabla \log \pi(s,a_i|\theta)\, r(a_i)}\end{aligned}$$

Approximating expectation by Monto Carlo sampling



Shrimp (as dots) have short life span.
In order to survive, fish need to eat enough shrimp when they are alive.
The two fish are competing for territories and food.

Purple fish: deep neural net
Blue fish: linear model

Green shrimps are *toxic* (-0.6)
Red shrimps are *healthy* (+0.8)

Demo: Natural Selection and Deep RL [Shen+ 16]

Three types of dialogue systems

- Social bot (see part 4 of this tutorial)
  - Microsoft XiaoIce, MSR neural conversation engine
- Task-completion bot
  - Movie ticket booking
  - Hotels booking
  - Travel assistant
- Info bot
  - Find the closest Starbucks with drive-thru
  - Find a family-friendly movie directed by Andrew Stanton near Redmond for upcoming weekend afternoons

Our focus:
goal-oriented
slot-filling
dialogues

An example dialogue of MovieBot

```
Turn 0 usr: can i get 2 tickets for race
Turn 1 sys: What date would you like to watch it?
Turn 2 usr: tomorrow
Turn 3 sys: Which theater would you like?
Turn 4 usr: amc pacific place 11 theater
Turn 5 sys: Which city you would like?
Turn 6 usr: seattle
Turn 7 sys: What time would you like to see it?
Turn 8 usr: 10:00 pm
 2 tickets for you
cific place 11 theater
```

Some of our dialogues can be more complex:

- Natural language understanding errors
  → reason under uncertainty
- Constraint violation
  → revise information collected earlier

Source code available on https://github.com/MiuLab/TC-Bot

Slot-filling dialogues

- Slot: information to be filled in before completing a task
  - For movie-bot: movie-name, theater, number-of-tickets, price, …
- Dialog act (intent)
  - Inspired by speech act theory (communication as action)

    request, confirm, inform, thank-you, …
  - Some may take parameters:

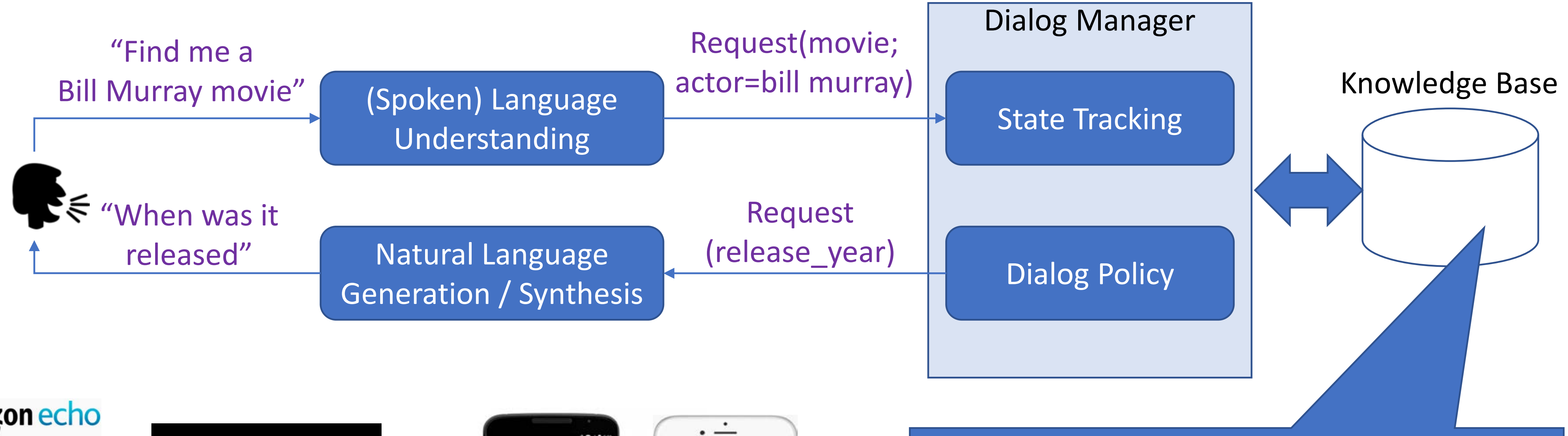
    request(price)

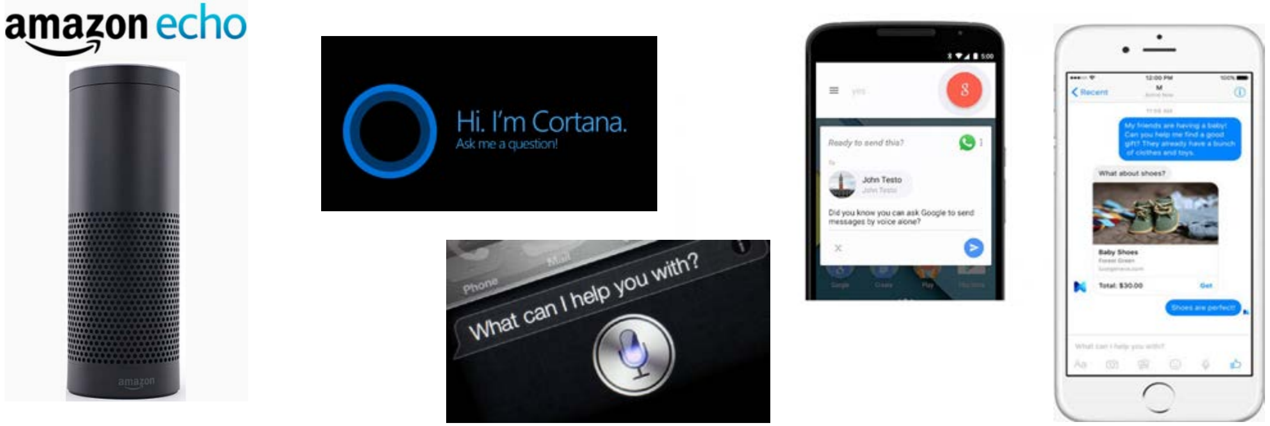
    confirm(moviename="kungfu panda")

    inform(price=$10)

    thank-you()

Multi-turn (goal-oriented) conversation

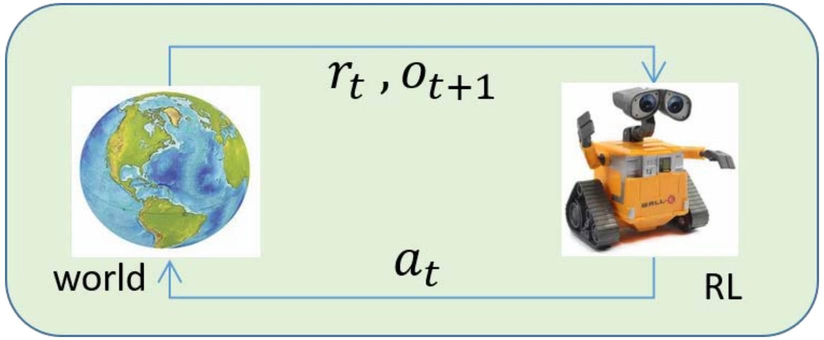


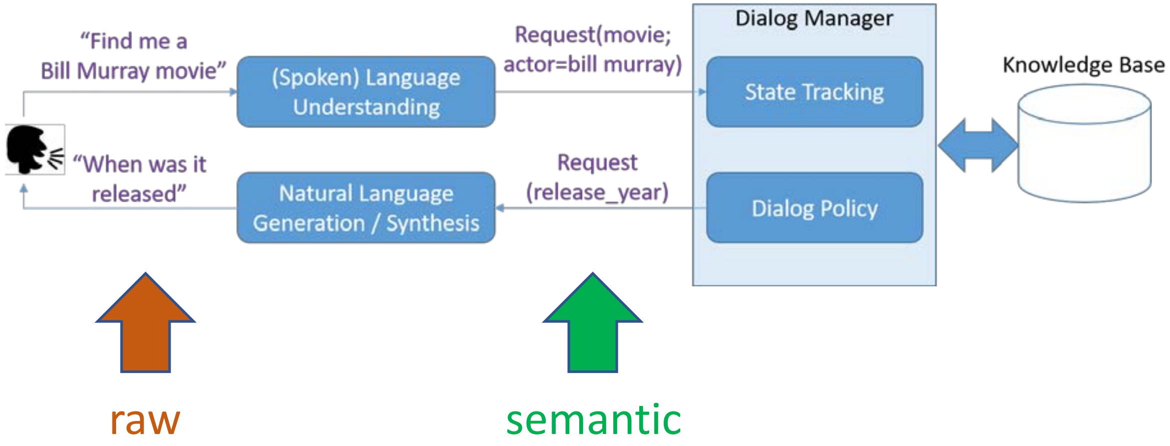


Entity-Centric Knowledge Base

Movie	Actor	Release Year
Groundhog Day	Bill Murray	1993
Australia	Nicole Kidman	X
Mad Max: Fury Road	X	2015

Conversation as RL





- Observation / action
  - o Raw utterance (natural language form)
  - o Semantic representation (dialog-acts)
- Reward
  - o $+10$ upon termination if succeeded
  - o -10 upon termination if failed
  - o -1 per turn
- State
  - o Explicitly defined (POMDP-based, …)
  - o Implicitly defined (RNNs)

Earlier examples: [Levin+ 00; Singh+ 02; Williams & Young 07]

Dialogue policy learning and evaluation

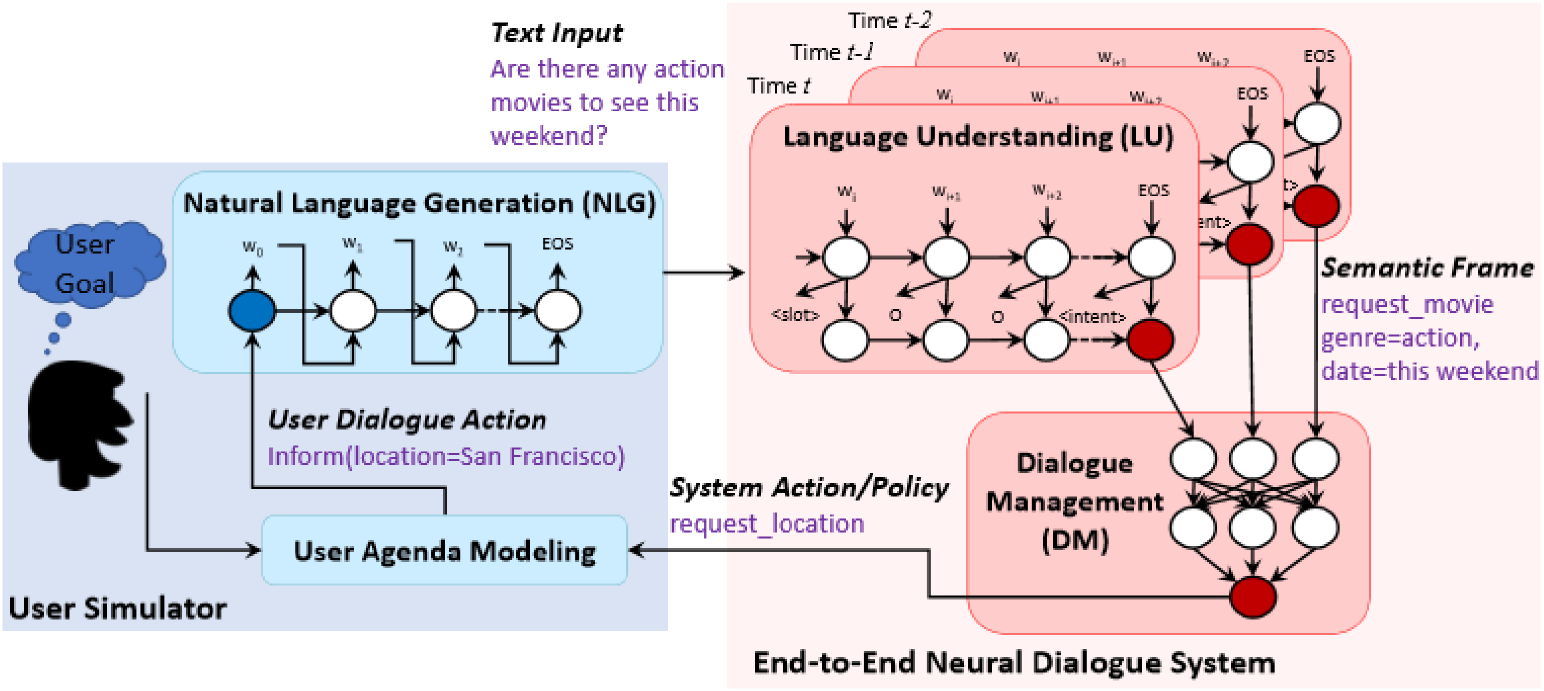
- Common metrics (reflected by reward function)
  - Task completion rate
  - Average #turns per dialogue
- But online learning on humans is too expensive
- Offline evaluation is very difficult [Liu+ 16]

Our approach:

1. Build a user simulator

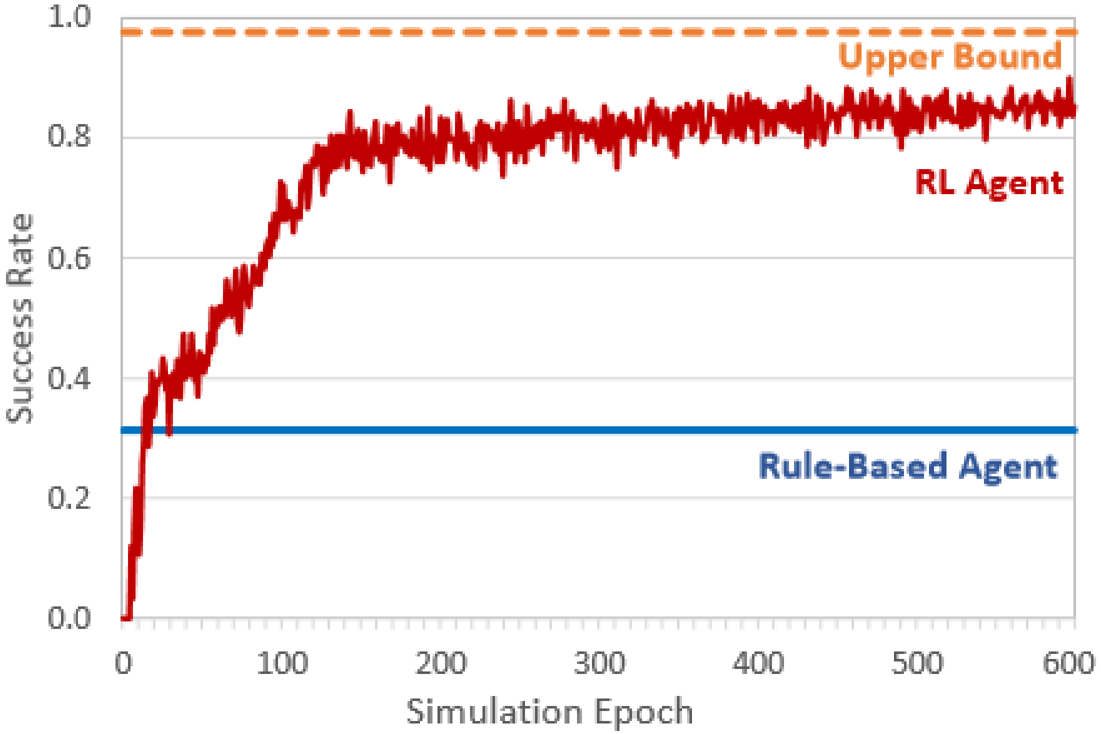
   E.g., agenda-based simulator [Schatzmann & Young 09]
2. Policy learning against the simulator
3. Policy metric evaluation against humans (e.g., on M. Turk)
4. Online incremental policy learning after deployment to product

A user simulator

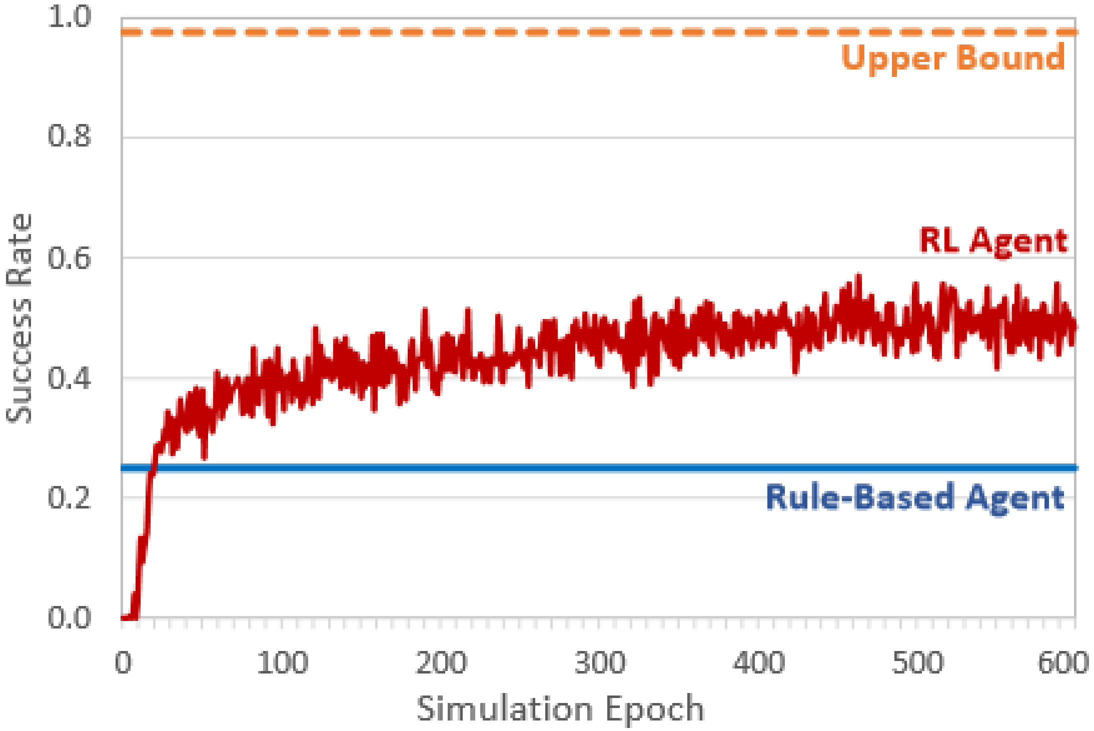


- Robustness: automatic action selection based on uncertainty by RL
- Flexibility: allow user-initiated behaviors
- Reproducibility: a R&D setting that allows consistent comparisons of competing methods

[Li+ 17] https://github.com/MiuLab/TC-Bot



(a) Frame-level semantics for training



(b) Natural language for end-to-end training

New episode, user goal:	**Rule-based agent**	**RL agent**
{ "request_slots": { "ticket": "UNK" }, "diaact": "request", "inform_slots": { "city": "seattle", "numberofpeople": "2", "theater": "regal meridian 16", "starttime": "9:25 pm", "date": "tomorrow", "moviename": "zoolander 2" } }	User: Can I get some tickets for zoolander 2? Agent: What movie are you interested in? User: I want to watch zoolander 2. Agent: What time would you like to see it? User: I want to watch at 9:25 pm. Agent: What city you would like? User: I want to watch at seattle. Agent: What date would you like to watch it? User: I want to set it up tomorrow. Agent: Which theater would you like? User: I want to watch at regal meridian 16. Agent: How many tickets do you need? User: I want 2 tickets please! Agent: Great - I was able to purchase 2 tickets for you to see zoolander 2 tomorrow at regal meridian 16 theater in seattle at 9:25 pm. User: Thank you. Agent: Thank you.	User: Can I get tickets for zoolander 2 tomorrow? Agent: Which city would you like? User: I want to watch at seattle. Agent: How many tickets do you need? User: I want 2 tickets please! Agent: 9:25 pm is available. User: I want to watch at regal meridian 16. Agent: Great - I was able to purchase 2 tickets for you to see zoolander 2 tomorrow at regal meridian 16 theater in seattle at 9:25 pm. User: Thank you. Agent: Thank you.
	Success	**Success**

RL agent learns to get information more efficiently by asking right questions at the right time.

New episode, user goal:	**Rule-based Agent**	**RL Agent**

{
"request_slots": {
"ticket": "UNK"
"theater": "UNK"
"starttime": "UNK"
},
"diaact": "request",
"inform_slots": {
"numberofpeople": "3",
"date": "tomorrow",
"moviename": "10 cloverfield lane"
}
}
	User: What is the start time for 10 cloverfield lane?
	Agent: What movie are you interested in?
	User: I want to watch at 10 cloverfield lane.
	Agent:What time would you like to see it?
	User: hat time is it playing tomorrow? I need 3 tickets.
	What city you would like?
	User:I do not care.
	What date would you like to watch it?
	User: I want to set it up tomorrow.
	Agent: Which theater would you like?
	User: Which theater is available?
	Agent: How many tickets do you need?
	User: I want 3 tickets please!
	Agent: Okay, I have booked 3 tickets for 10 cloverfield lane tomorrow.
	User: Thank you.
	Agent: Thank you.
		User: Which theater can I book 3 tickets for 10 cloverfield lane?
		Agent: What time would you like to see it?
		User: Which theater and start time are available tomorrow?
		Agent: 11:45am is available.
		User: Which theater is available?
		Agent: regal la live stadium 14 is available.
		User: Could you help me to book the tickets?
		Agent: Okay - I was able to book 3 tickets for you to see 10 cloverfield lane at regal la live stadium 14 at 11:45am tomorrow.
		User:Thank you
		Agent: Thank you

	Failure: Agent does not answer all the questions (starttime, theater) from user.	**Success**

RL agent learns to answer user questions

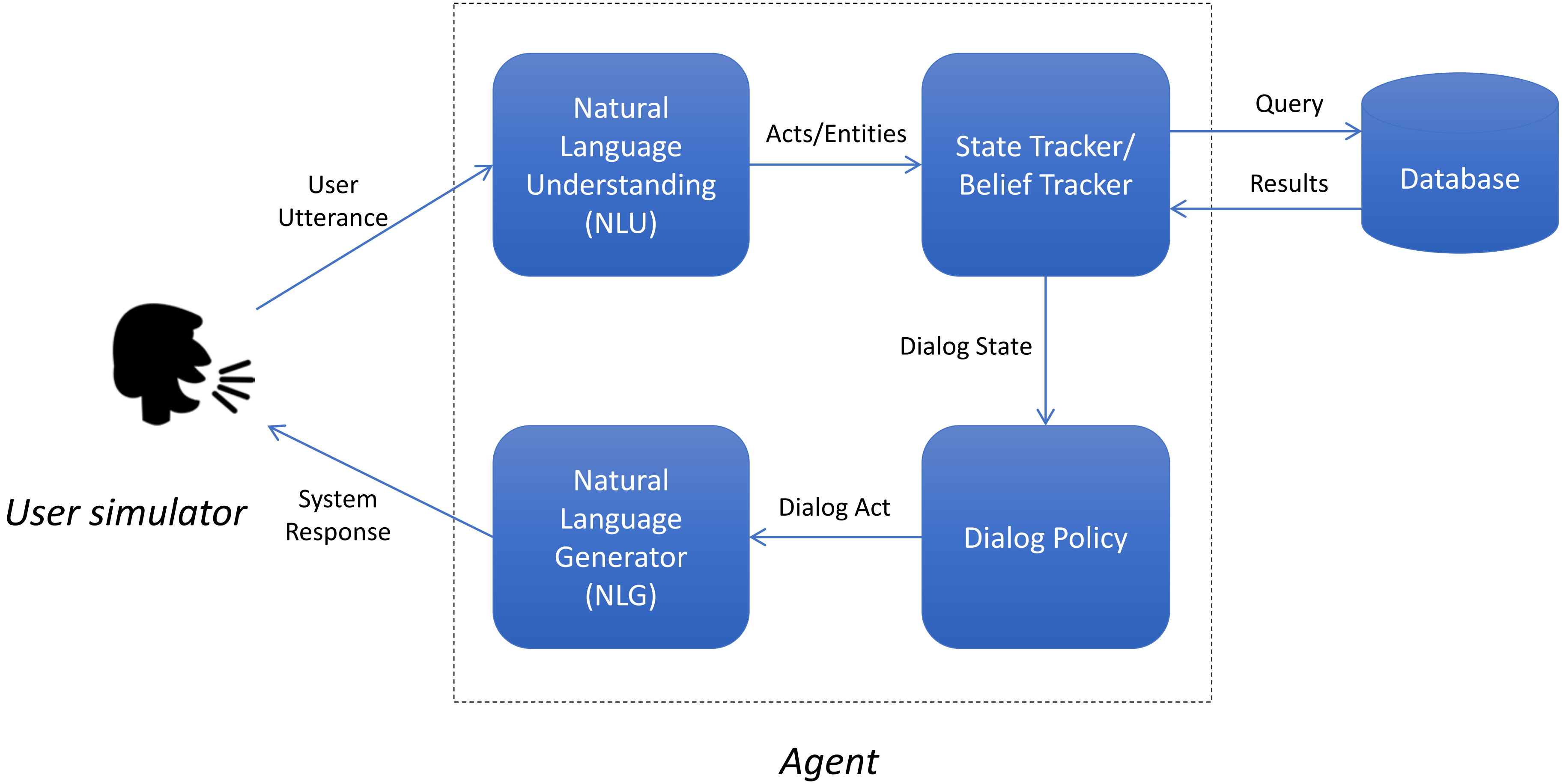
Three case studies

- Info bot: end-to-end training with non-differentiable knowledge base [Dhuwan+ 17]
- Task-completion bot: efficient exploration for domain extension [Zachary+ 17]
- Composite task completion bot with Hierarchical RL [Peng+ 17]

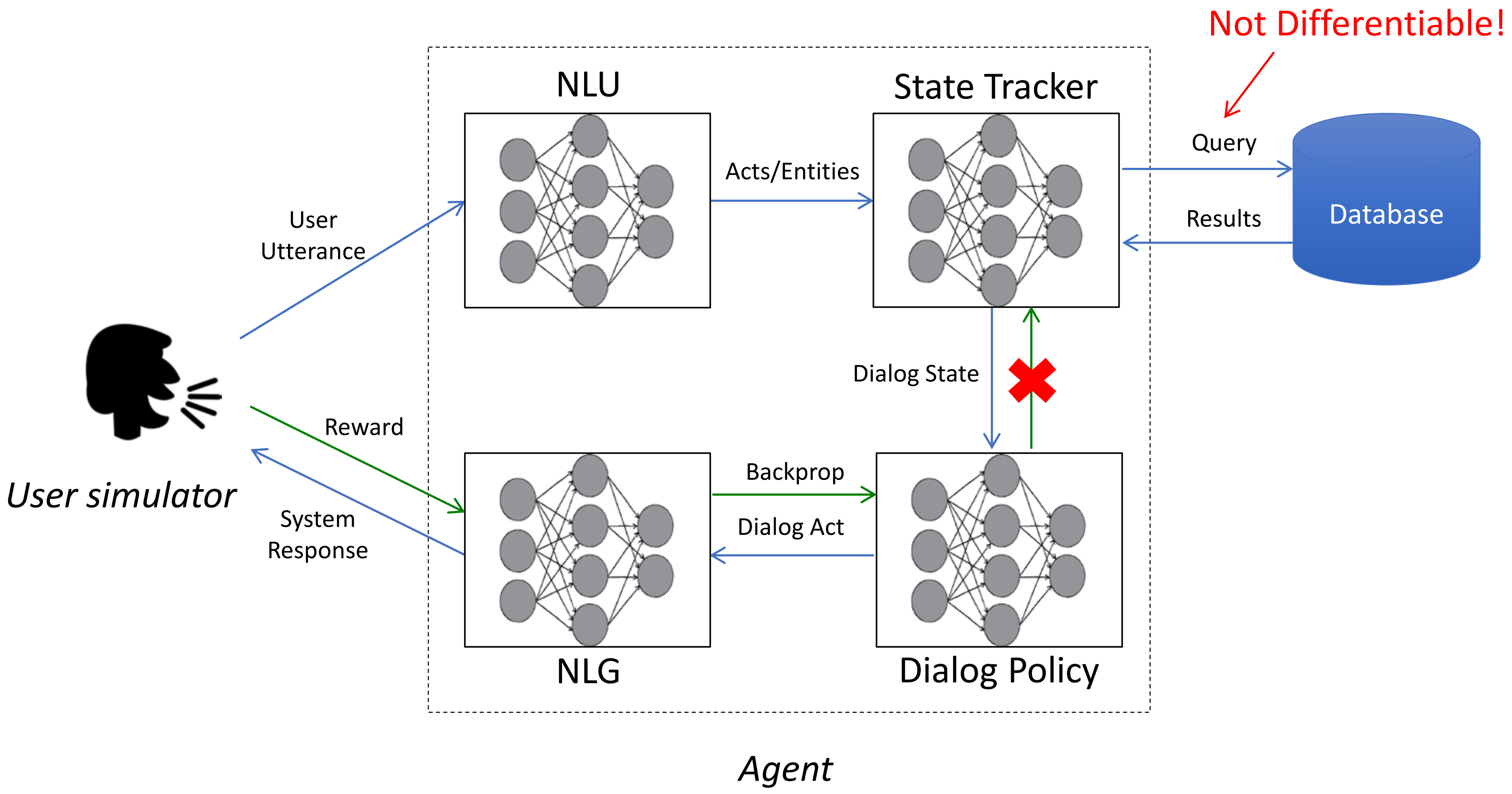
InfoBot as an interactive search engine

- Problem setting
  - User is looking for a piece of information from one or more tables/KBs
  - System must iteratively ask for user constraints (“slots”) to retrieve the answer
- A general rule-based approach
  - Given current beliefs, ask for slot with maximum uncertainty
  - Works well in most cases but,
    - Has no notion of what the user is likely to be looking for or likely to know
    - No principled way to deal with errors/uncertainty in language understanding

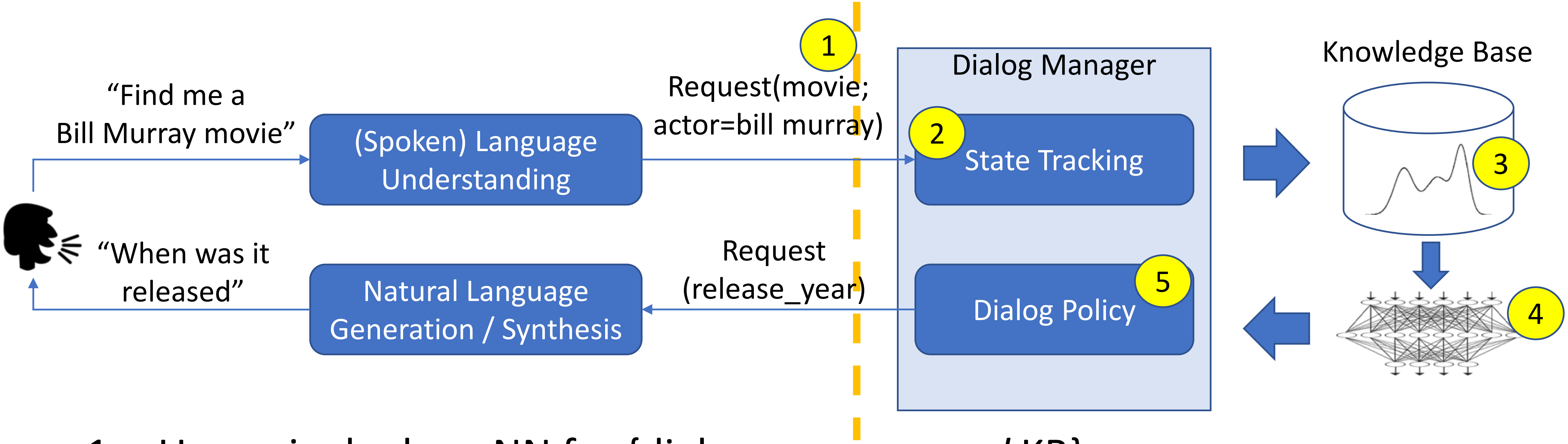
InfoBot as an interactive search engine



Deep Reinforcement Learning



Our end-to-end approach



1. Use a single deep NN for {dialog manager *and* KB}
2. Recurrent network to track states of conversation
3. **Maintain (implicitly) a distribution over entities in KB**
4. A summary network to “summarize” distribution information
5. Multilayer perceptron policy network

Whole network can be end-to-end trained by BP/SGD!

Soft attention for KB-lookup

Entity-Centric Knowledge Base

Movie	Actor	Release Year
Groundhog Day	Bill Murray	1993
Australia	Nicole Kidman	X
Mad Max: Fury Road	X	2015

- Posterior computation:

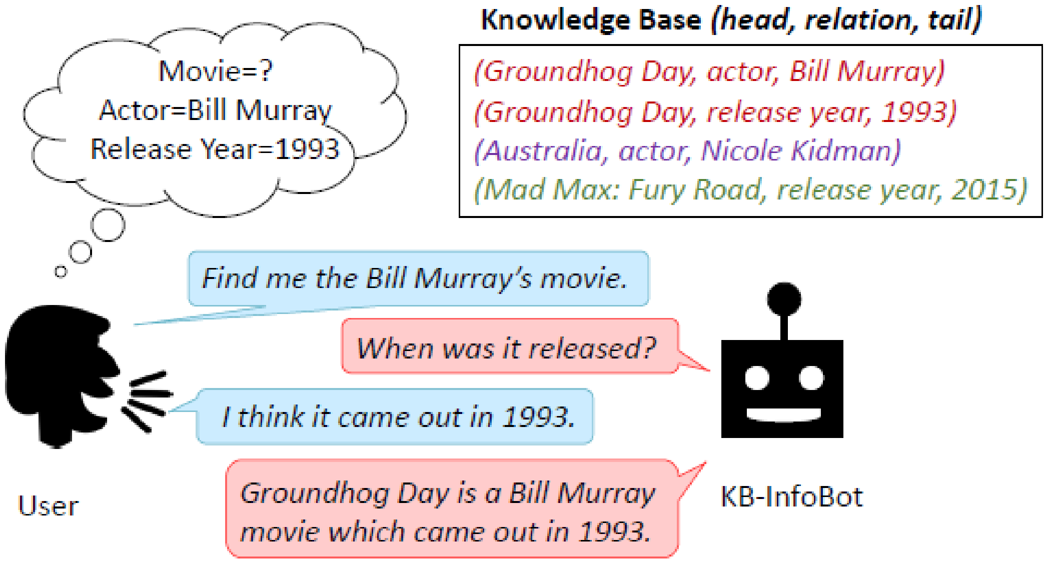
$$\mathrm{Pr}(\text{"GroundhogDay"}) \propto \mathrm{Pr}(\text{Actor} = \text{"Bill Murray"}) \cdot \mathrm{Pr}(\text{ReleaseYear} = \text{"1993"}) \cdots$$

Each $\mathrm{Pr}(\text{slot} = \text{value})$ is computed in terms of LU outputs

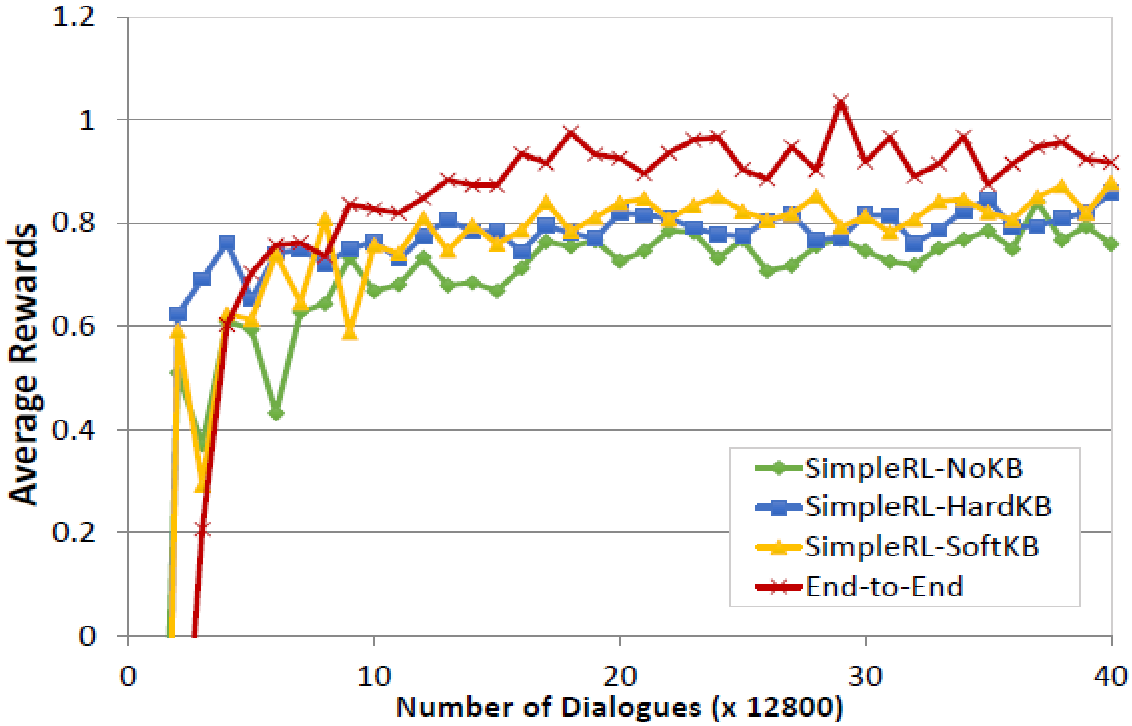
- Soft KB-lookup: sample a movie according to the posterior
  - Randomization results in differentiability (similar to policy gradient alg.)
  - As opposed to using SQL queries to look up results deterministically

Whole system can be trained using policy gradient & back-propagation

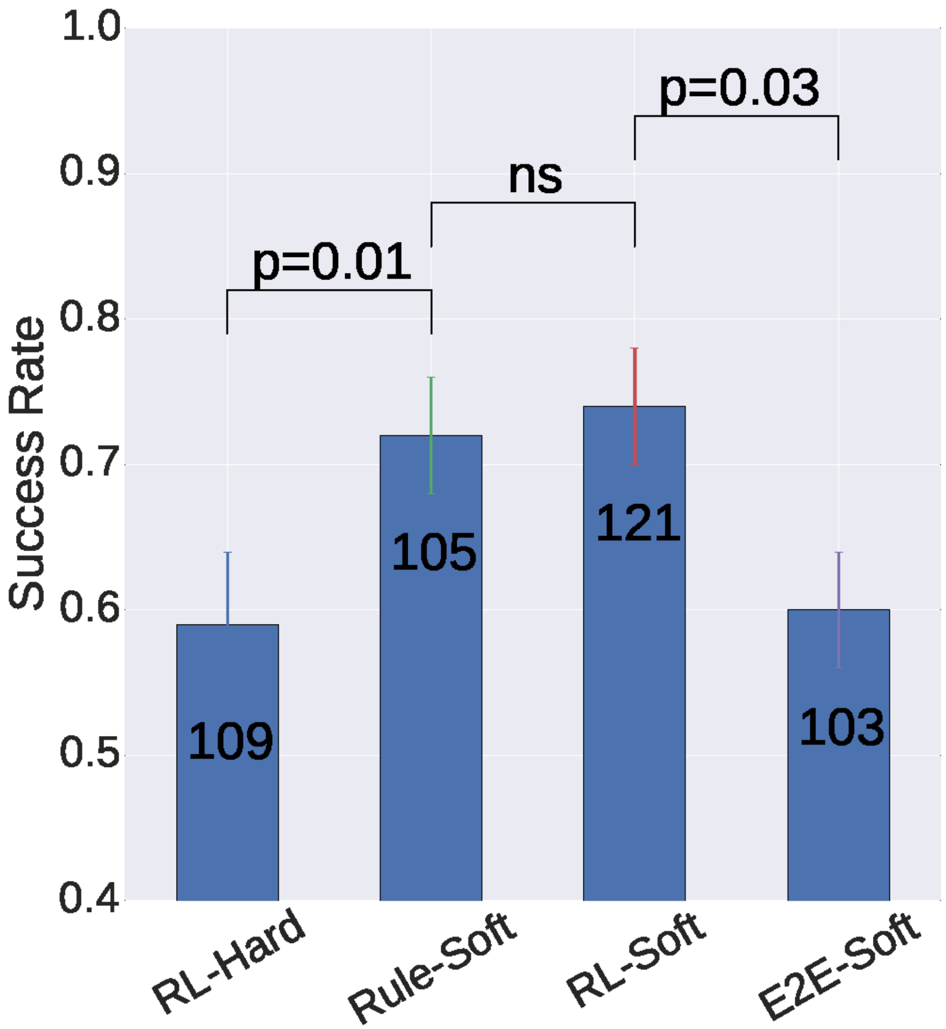
Result on IMDB using KB-InfoBot w/ simulated users

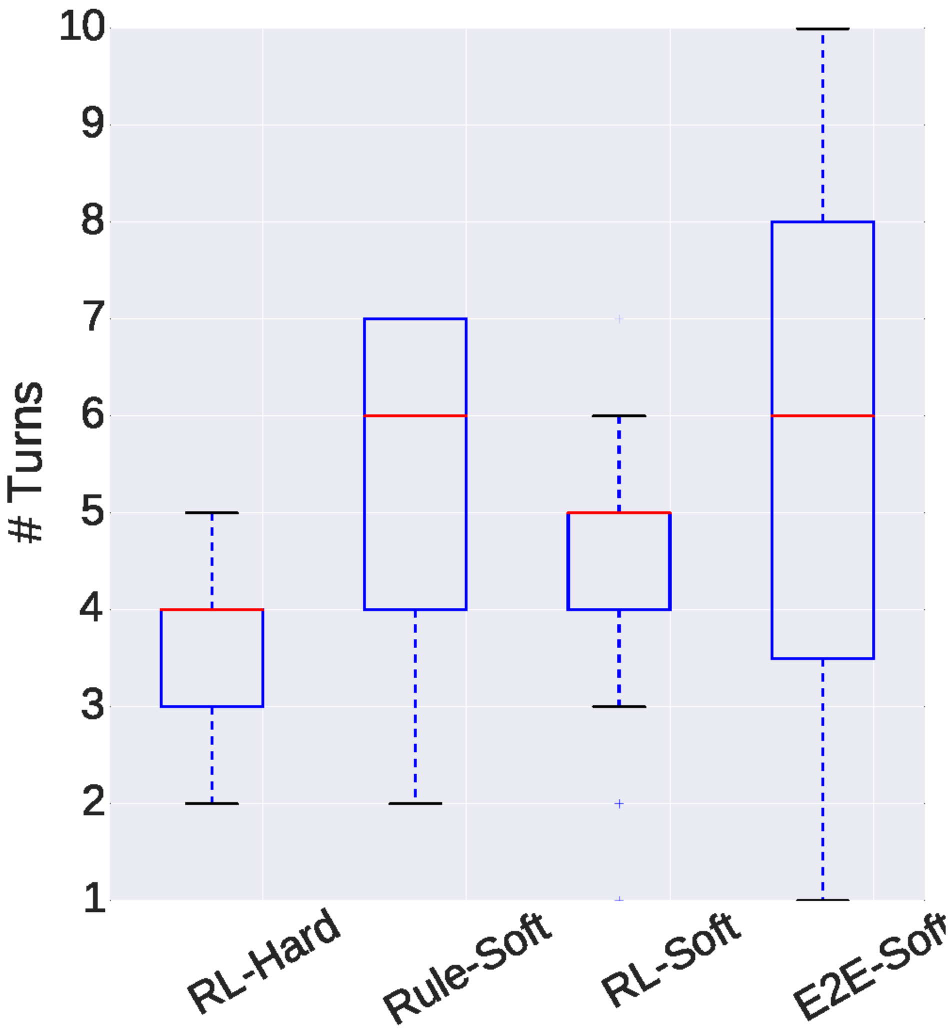


Agent	Success Rate	Avg Turns	Avg Reward
Rule-Soft	0.76	3.94	0.83
RL-Hard	0.75	**3.07**	0.86
RL-Soft	0.80	3.37	0.98
E2E-RL	**0.83**	3.27	**1.10**



Results on real users



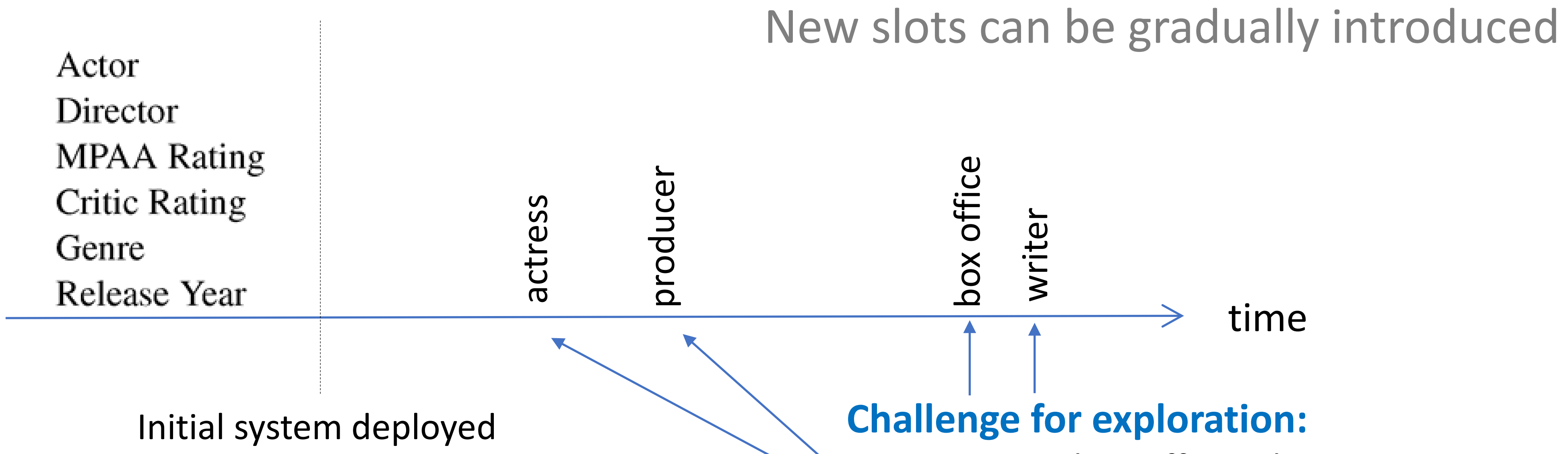


Three case studies

- Info bots: end-to-end training with non-differentiable knowledge base
- **Task-completion bots: efficient exploration for domain extension [Zachary+ 17]**
- Composite task completion bots with Hierarchical RL [Peng+ 17]

Domain extension

- Most goal-oriented dialogs require a closed and well-defined domain
- Hard to include all domain-specific information up-front

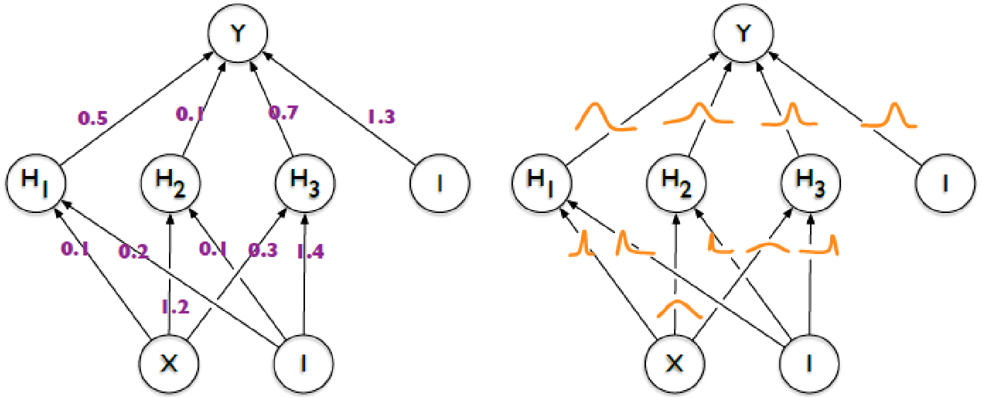


Challenge for exploration:

- How to explore efficiently
- to collect data for new slots
- When deep models are used

Efficient exploration for dialogue

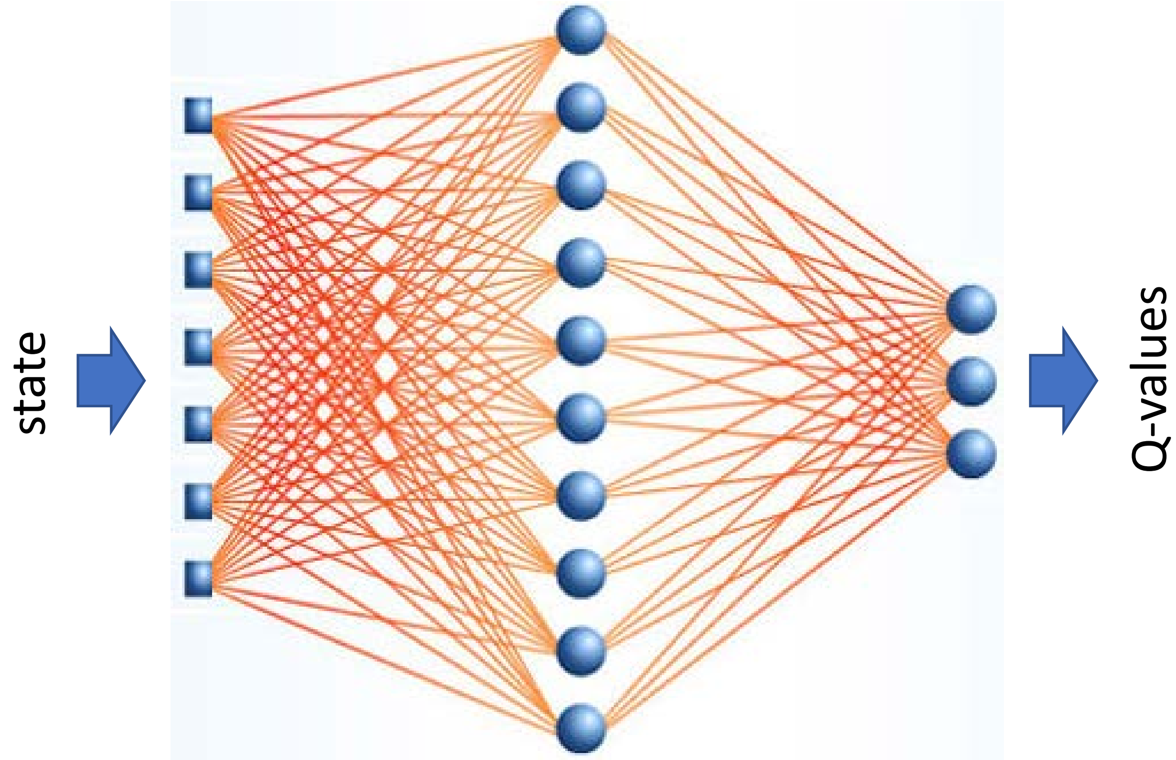
- ϵ-greedy can be slow & wasteful, frequently trying known bad moves
  - Compared to Atari/Go settings, failures in dialogue systems confer high economic costs
- Given uncertainty information, one can make smarter exploration decisions
  - DQNs give best estimates of value functions, but don't offer uncertainty information
- Our solution: get uncertainty info from Bayesian neural networks
  - Explore to area where the model is not confident



Deep Bayes-by-Backprop Q Network (Deep BBQ Networks)

- Construct a BBQN w. Gaussian variational dist. and Gaussian prior
- Explore by Thompson sampling, drawing Monte Carlo (MC) samples from a stochastic neural net
- At train time draw one MC sample from BBQN and update by BP, using the re-parameterization trick [Kingma & Welling 13]

Deep Q-network (DQN)



DQN-learning of network weights θ: apply SGD to solve

$$\hat{\theta} \leftarrow \arg\min_{\theta} \sum_{t} \left(r_{t+1} + \gamma \max_{a} Q_T(s_{t+1}, a) - Q_L(s_t, a_t) \right)^2$$

"Target network" to synthesize regression target

"Learning network" whose weights are to be updated

[DeepMind 15]

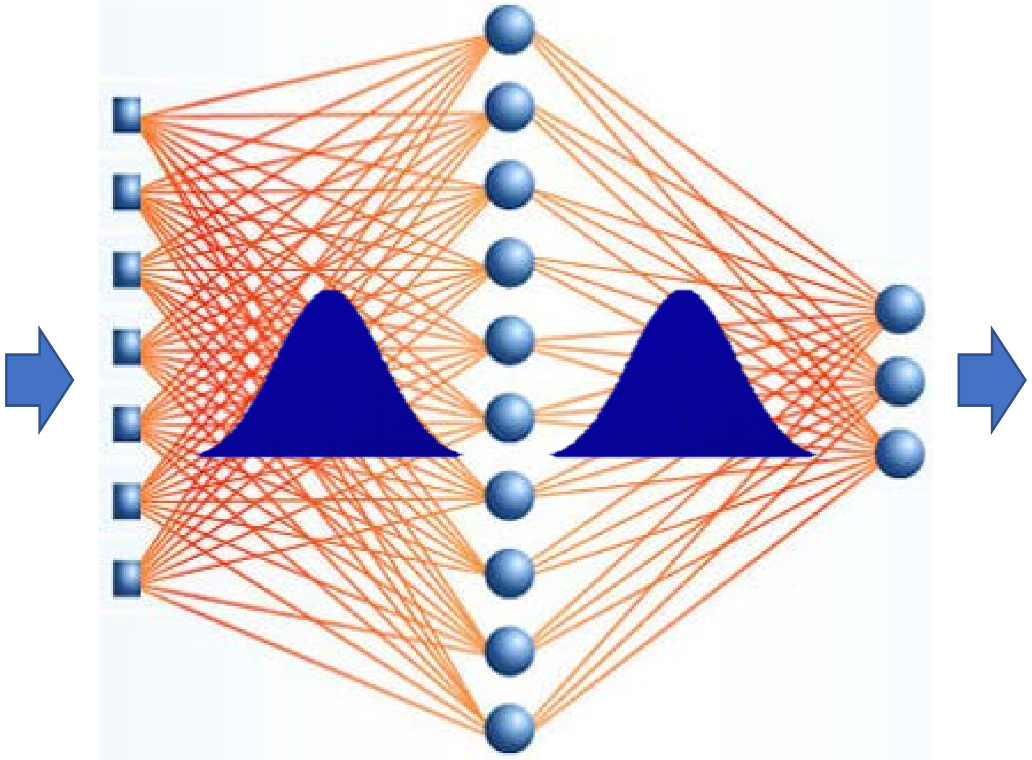
Bayes-by-Backprop Q (BBQ) network

BBQ-learning of network params $\theta = (\mu, \sigma^2)$:

$$\hat{\theta} = \arg\min_{\theta_L} \mathrm{KL}\big(q(\mathbf{w}|\theta_L) \;||\; p(\mathbf{w}|Data)\big)$$

Still use "target network" θ_T
to synthesize regression target

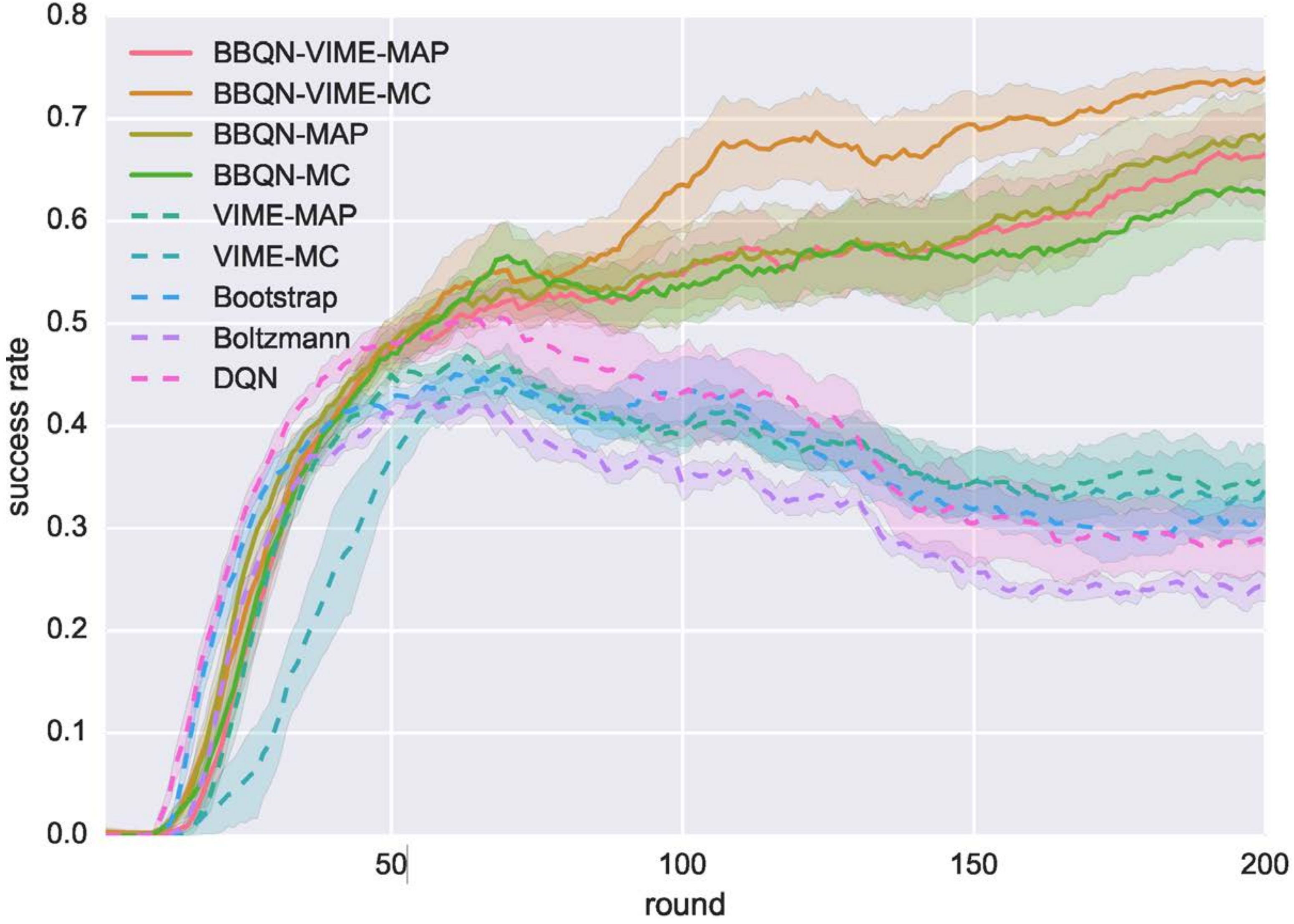
state



Q-values

- Parameter learning: solve for $\hat{\theta}$ with Bayes-by-backprop [Blundell+ 15]
- Params θ quantifies uncertainty in Q-values
- Action selection: use Thompson sampling for exploration

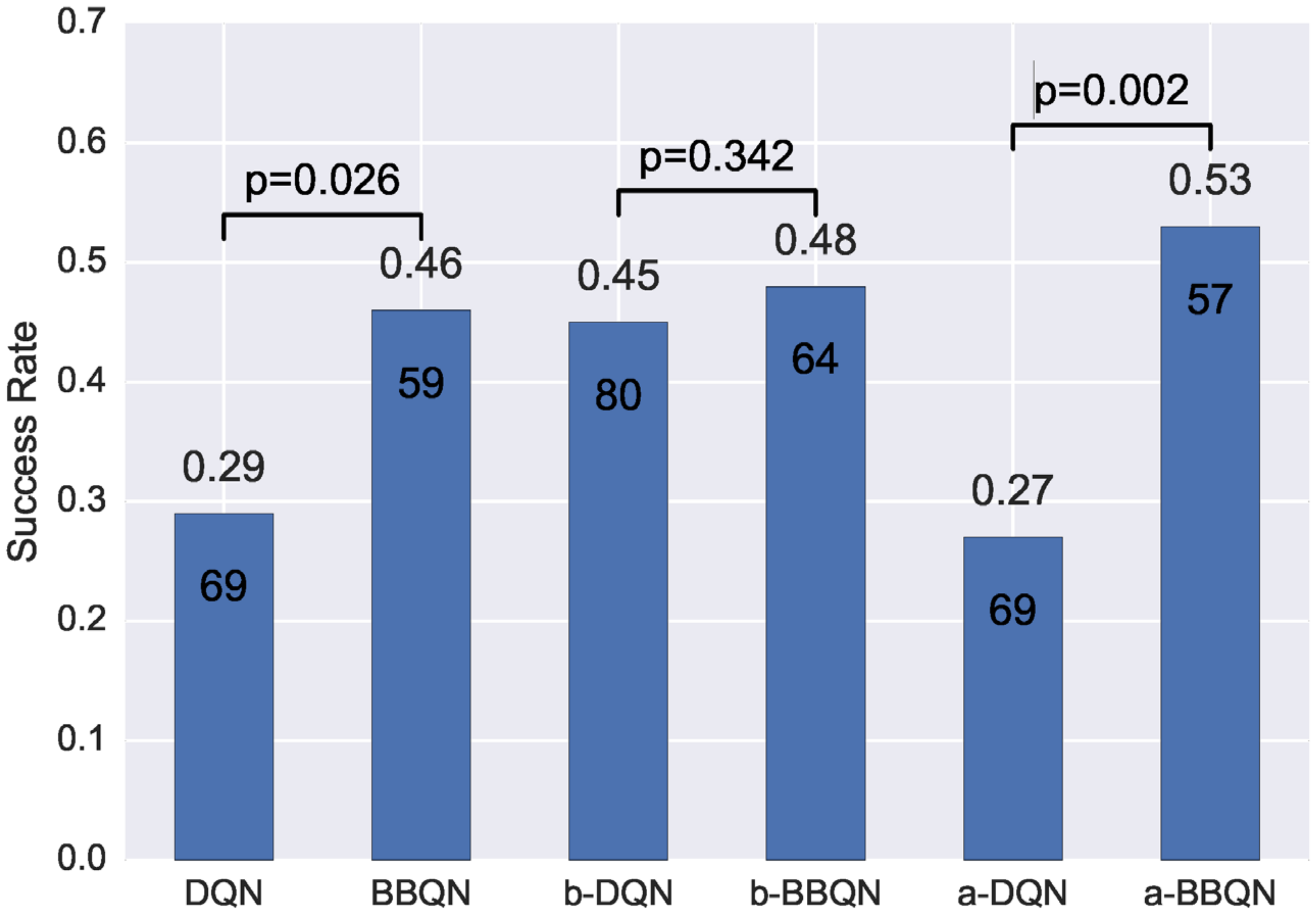
Results on simulated users



Our BBQ approach successfully explores to adapt to handle new slots.

It also works best in regular dialogue settings (with fixed/full domain)

BBQ results with real users

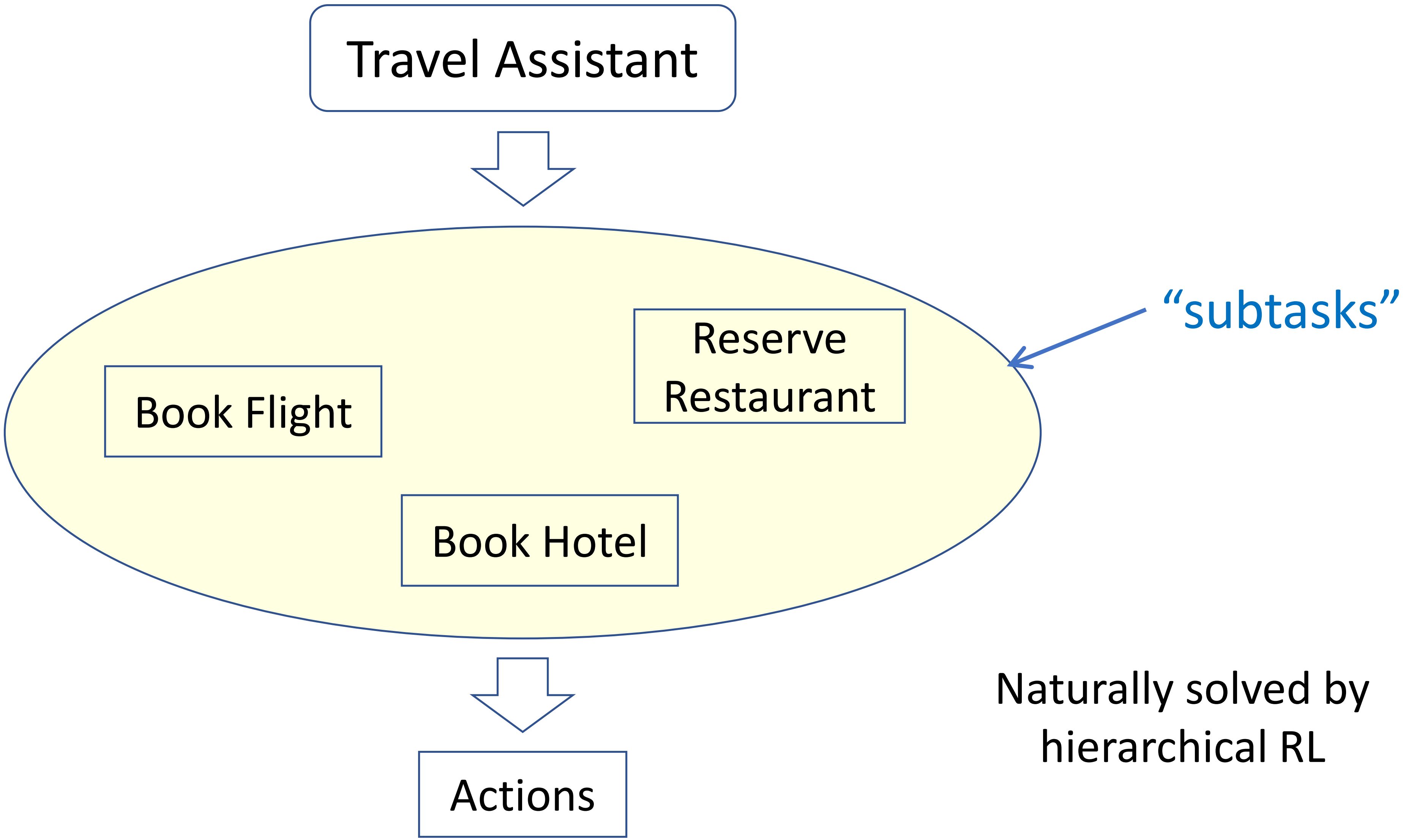


- **DQN/BBQN**: regular dialogue policy learning (with full/fixed domain)
- **b-***: model trained on smaller domain
- **a-***: models trained after domain extension

Three case studies

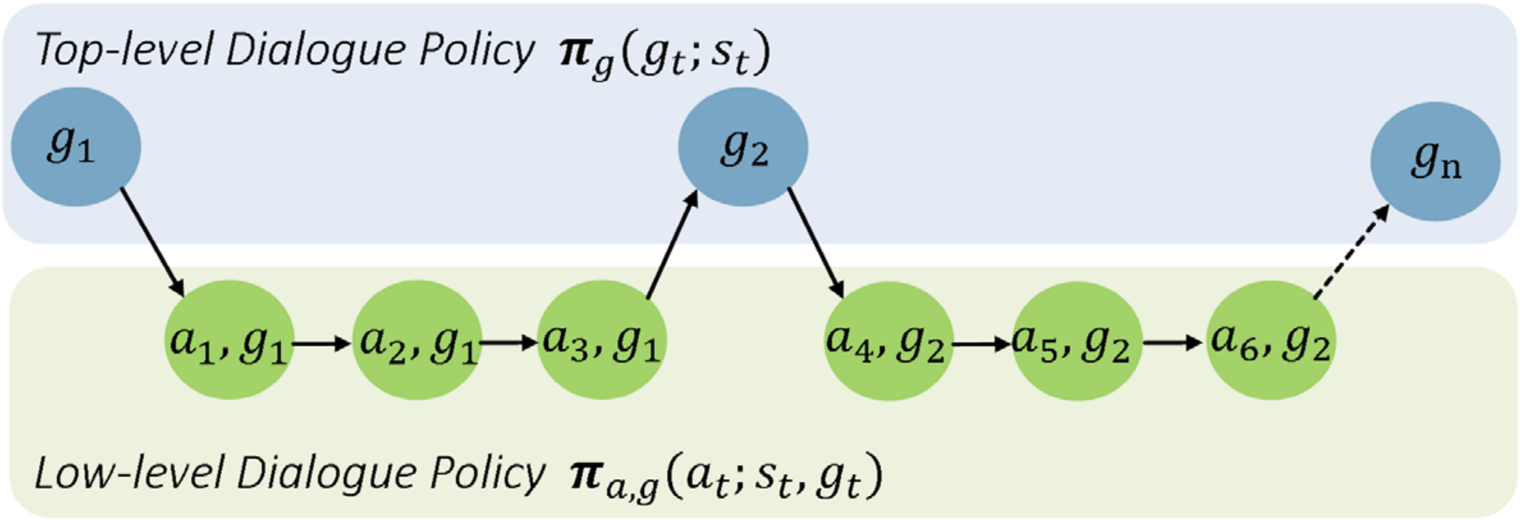
- Info bots: end-to-end training with non-differentiable knowledge base
- **Task-completion bots: efficient exploration for domain extension**
- **Composite task completion bots with Hierarchical RL [Peng+ 17]**

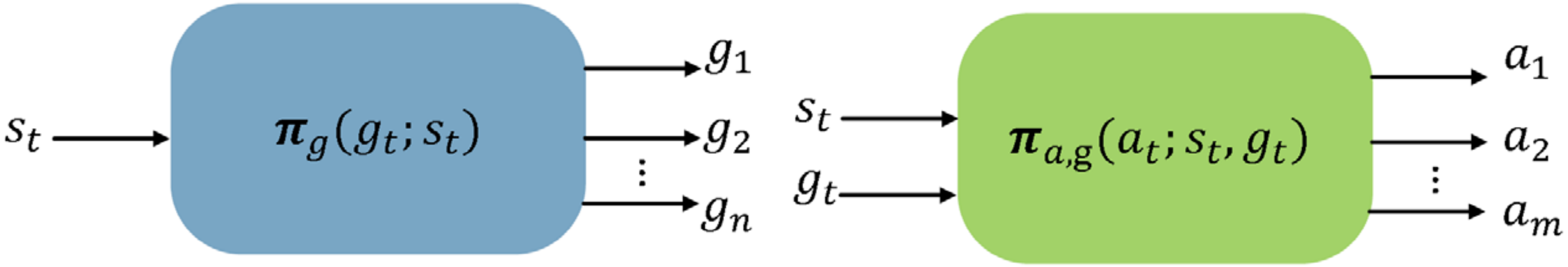
Composite tasks



Naturally solved by hierarchical RL

A hierarchical policy learner



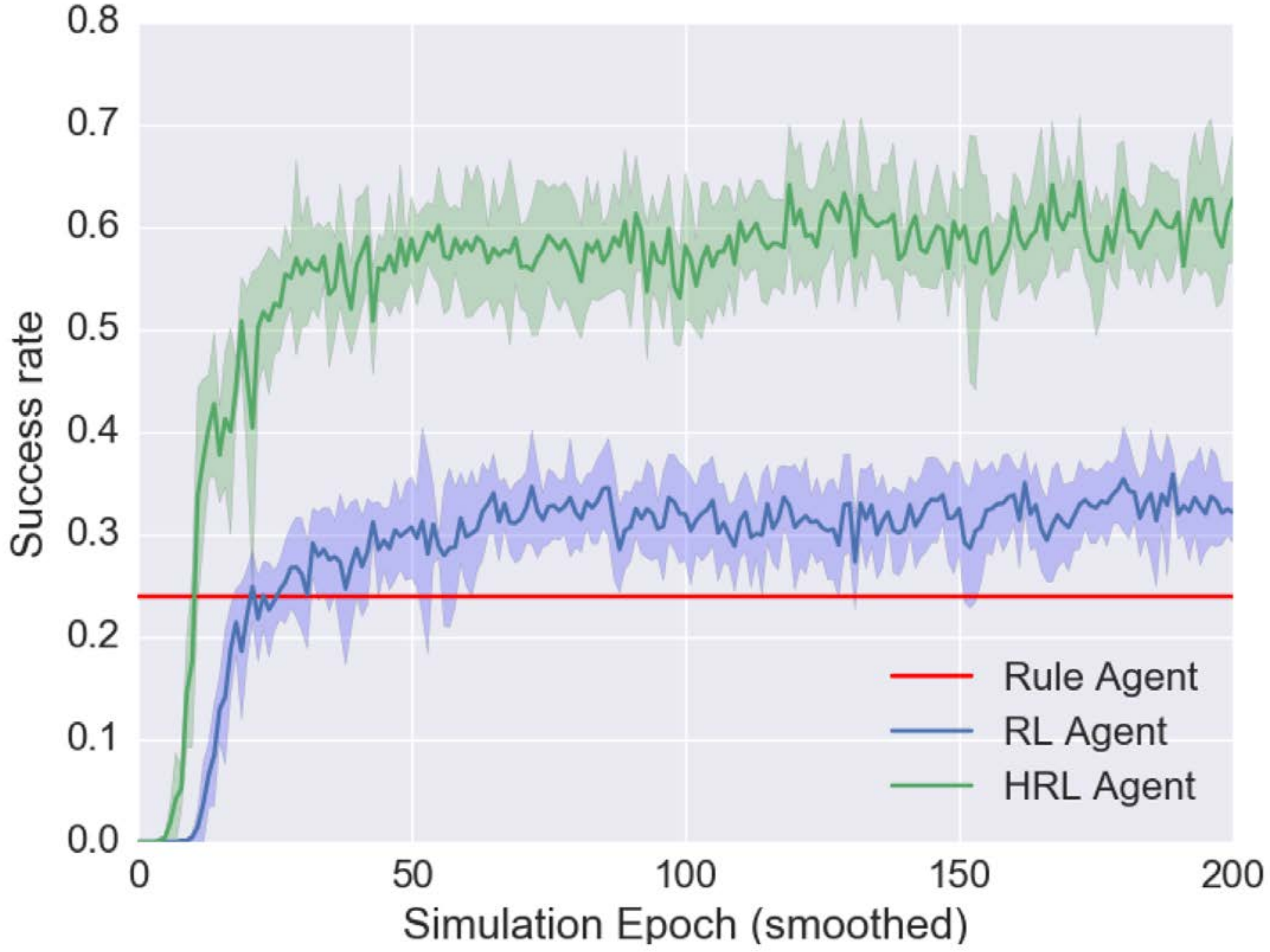


Similar to HAM [Parr & Russell 98] and hierarchical DQN [Kulkarni+ 16]

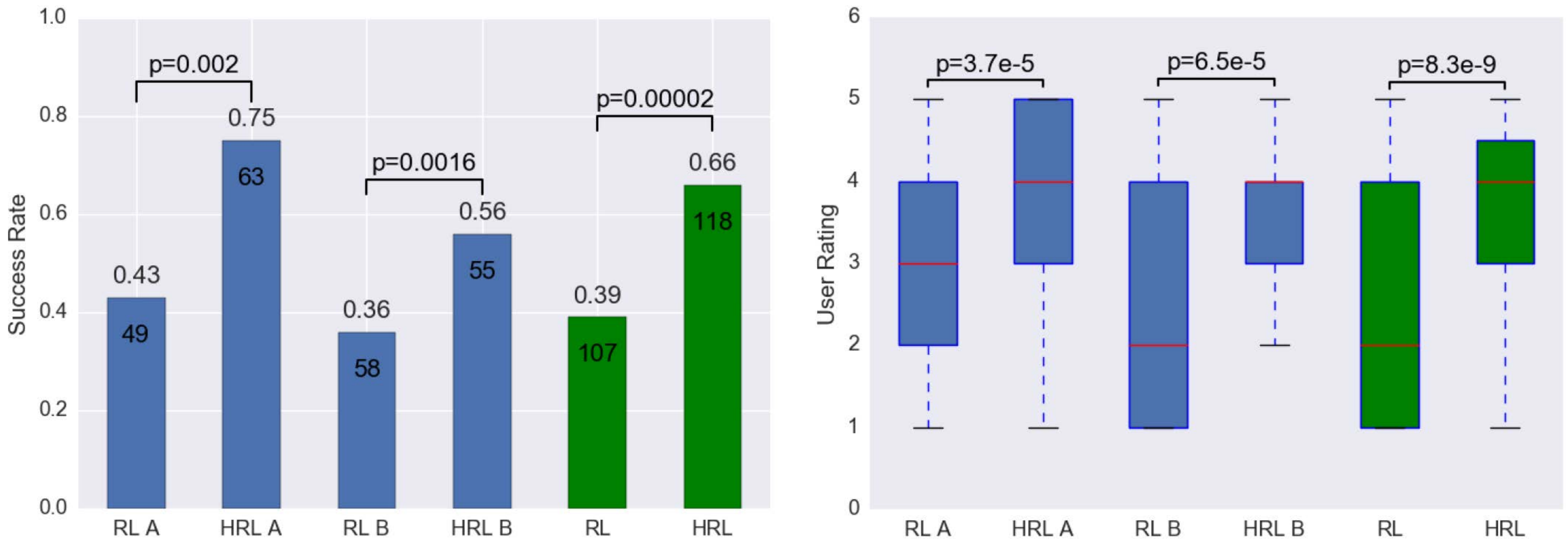
Experiment: Setup

- Travel assisting agent with two sub-tasks: hotel and flight
- Different types of users:
  - Some have more constraints with hotel booking
  - Some have more constraints with flight booking
  - The top-level policy should learn to personalize

Results on simulated users



Results on real users



Type A users: do not have any preference to subtask
Type B users: prefer to complete the book-flight-ticket subtask first

Interim summary

- Long-term ambition
  - An intelligent, human-like, open-domain conversational system
  - How to deal with commonsense knowledge?
  - How to handle open-domain dialogues
  - How to do better off-policy learning/evaluation?
  - …
- Deep RL plays a critical role
  - Learn more at deep RL for goal-oriented dialogues
- Interesting connections to other AI areas
- Lots of new research topics at the intersection between RL and NLP

Thank You!

- Part 1: Background
- Part 2: Deep semantic similarity models for text processing
- Part 3: Recurrent neural networks for text generation
- Part 4: Neural machine reading models for question answering
- Part 5: Deep reinforcement learning for task-completion dialogue

Contact Information:

www.microsoft.com/en-us/research/people/jfgao/