Deep Learning for Question Answering

Lei LI

Toutiao Lab

头条 今日头条

12/3/16

Goal

Enable machines to comprehend and converse

- Bots to write/tell a story
- Bots to chitchat
- Bots to organize knowledge

Major applications of QA

- Search engine
- Personal assistant
- Information platform

cortana

siri

助理来也

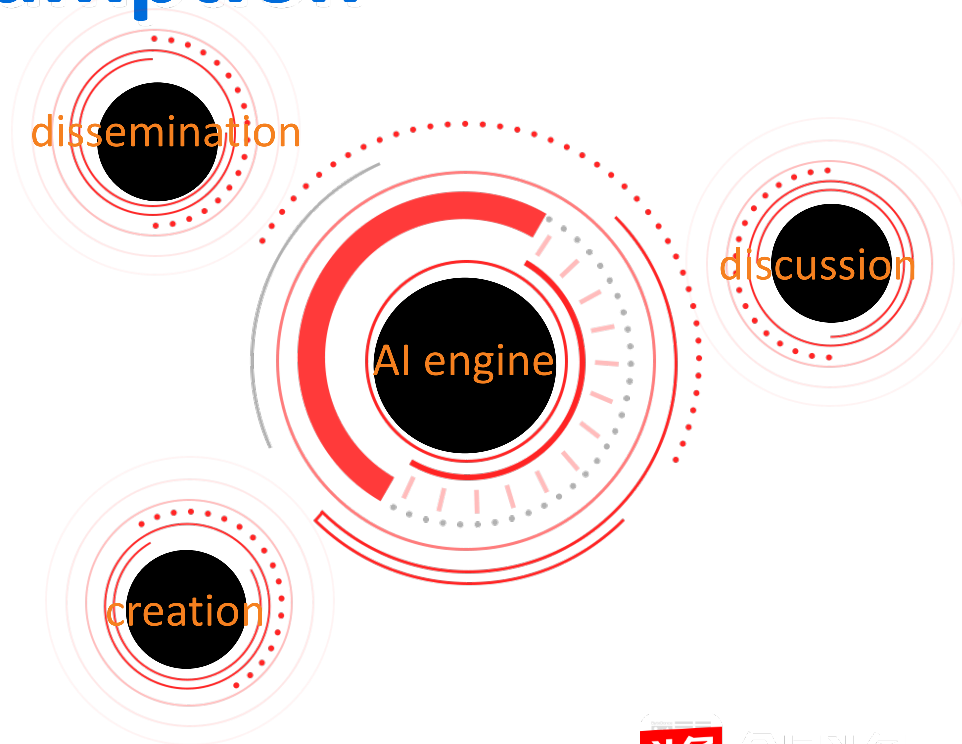
Information consumption platform

News article
Stories
Video
Community QA

Three key areas in effective information consumption

Creating intelligent machines that understand information in depth (text, images, videos, comments, etc.) to better serve our users with what they like

Developing large scale machine learning algorithms for personalized information recommendation



QA can be one genre of content

Outline

- Problem Setup, Knowledge graphs
- Basic DL techniques
- Word, entity and relation embeddings
- Recurrent neural nets for processing sequence
- Focused Pruning: parsing the subject mention
- Finding the right relation and subject entity
- Other approaches:
  - LTG+CNN
  - Memory network

Categories of Questions

- Factoid: who is the president of USA?
- Descriptive: what are characteristics of the new Mac Pro
- Procedural: how to install windows 10
- Calculation: how many Chinese won Turing awards?
- Causal: why is it dark at night
- Opinion: how do you think about Trump?

Factoid questions: Simple to Complex

Simple Question

- One that can be answered with single evidence
- E.g. Who wrote the book of Beijing Folding?

This tutorial

Multi-hop Question

- Requires with many facts
- E.g.

Aggregate Question

- Requires with many facts and calculation
- E.g. what is the longest Olympic opening before Beijing 2008

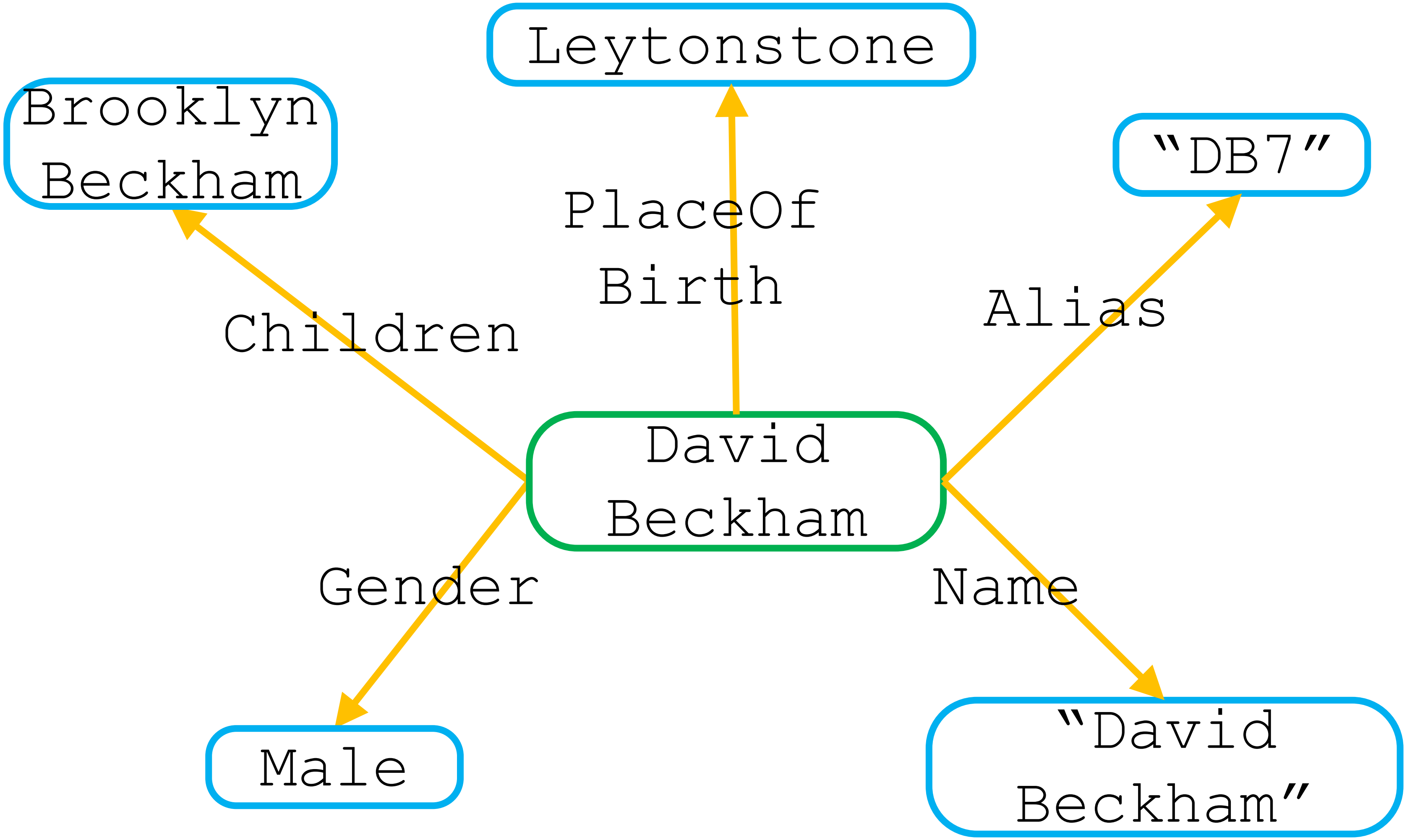
Q: Where was David Beckham born?

What do we need to answer questions?

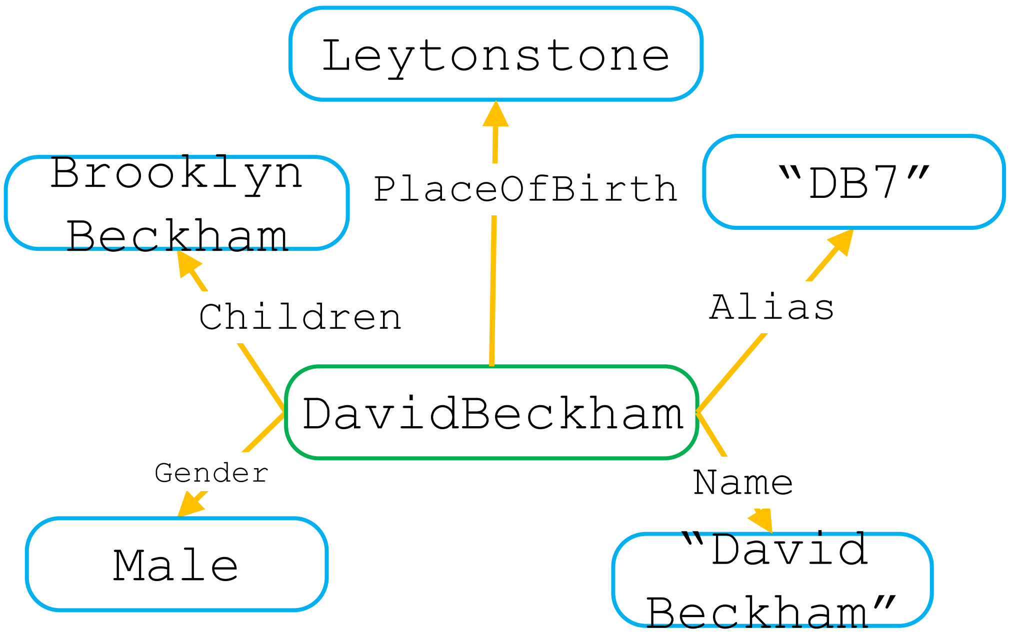
- Fact storage: knowledge graph
- Mapping from natural questions to structured queries executable on knowledge graph

Q: Where was David Beckham born?

Knowledge Graph



Knowledge Graph



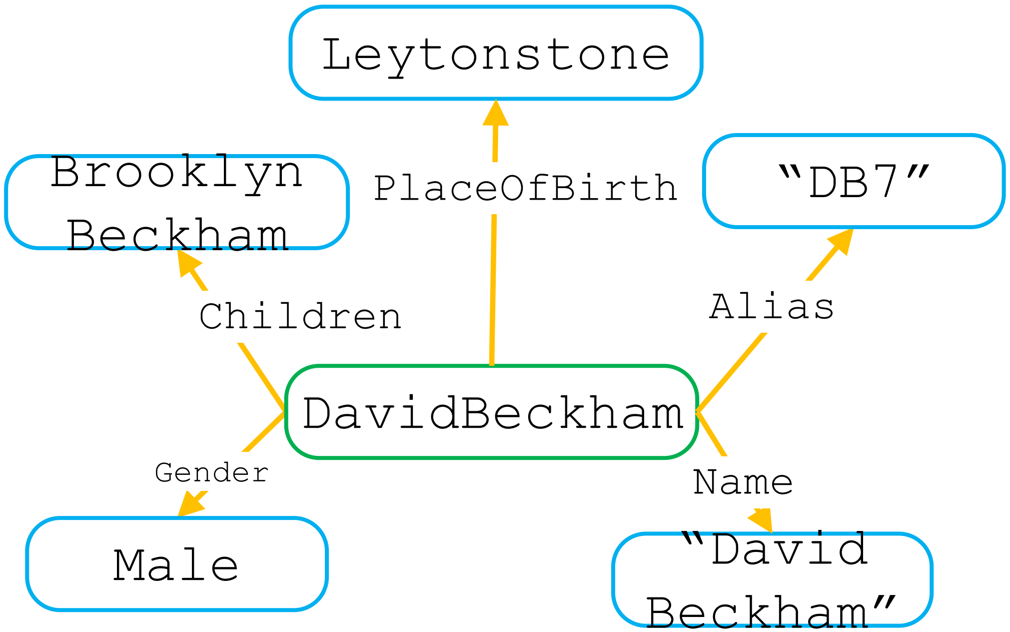
Knowledge as triples

<DavidBeckham, Name, "David Beckham">

<DavidBeckham, PlaceOfBirth, Leytonstone>

<Subject, relation, object>

Structure Query on KG



SPARQL

SELECT ?object
WHERE { <DavidBeckham> <PlaceOfBirth> ?object }

From natural language question to structured query

Question

Where was David Beckham born?

subject

relation

SPARQL query

```
SELECT ?object
WHERE { <DavidBeckham> <PlaceOfBirth> ?object }
```

related work
[Berant 2013]
[Yih 2014]
[Bordes 2015]

Why difficult for machines?

Language complexity

- 奥巴马总统在哪儿生的?
- 奥巴马总统出生地在哪里?
- What is the birthplace of Mr. Obama?
- Where was Mr. Obama born?

Ambiguity

- 麦克乔丹是谁?
- Who is Michael Jordan?

Sparse Label

- 2千万事实，十万标注问答对
- 22 million , 100k labeled QA pairs

Simple solutions: N-gram

- Rank and match all possible n-grams in the question
- Link them to entities in KG via alias matching

Where was David Beckham born?

N-gram candidates:
Uni-gram: Where, was, David, Beckham, born, ?
Bi-gram: Where was, was David, David Beckham, Beckham born, born ?
Tri-gram: Where was David, was David Beckham, David Beckham born, Beckham born ?
Four-gram: Where was David Beckham, was David Beckham born, David Beckham born ?

Improved simple solutions

- Rank and match all possible n-grams in the question
- Prune the n-grams with heuristics
- Link them to entities in KG via alias matching

Where was David Beckham born?

N-gram candidates:

Uni-gram: Where, was, David, Beckham, born, ?

Bi-gram: Where was, was David, David Beckham, Beckham born, born ?

Tri-gram: Where was David, was David Beckham, David Beckham born, Beckham born ?

Four-gram: Where was David Beckham, was David Beckham born, David Beckham born ?

Challenges

1. Insufficient Knowledge Representation

- Where is San Francisco?
- What is Columbus famous for?

- **MORE** than **400** entities
- **City, County, Person, Movie,** etc

2. Too Much Noise from N-Grams

- What theme is the book the armies of memory?

- the book: 73
- theme: 252
- memory: 553
- ……

Outline

- Problem Setup, Knowledge graphs
- Basic DL techniques
- Word, entity and relation embeddings
- Recurrent neural nets for processing sequence
- Focused Pruning: parsing the subject mention
- Finding the right relation and subject entity
- Other approaches:
  - LTG+CNN
  - Memory network

DL algorithms work well for

Supervised learning

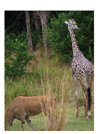
data — label

$$X \xrightarrow{\quad f(\cdot) \quad} Y$$

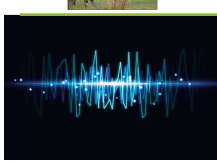
Cat/dog/...

“今天天气不错！”

“Today is a nice day”



A giraffe standing next to forest



“打车去故宫”

Handwriting Recognition

0
1
2
3
4
5
6
7
8
9

Inspired by a biological neuron

Neural networks: massively connected simple units

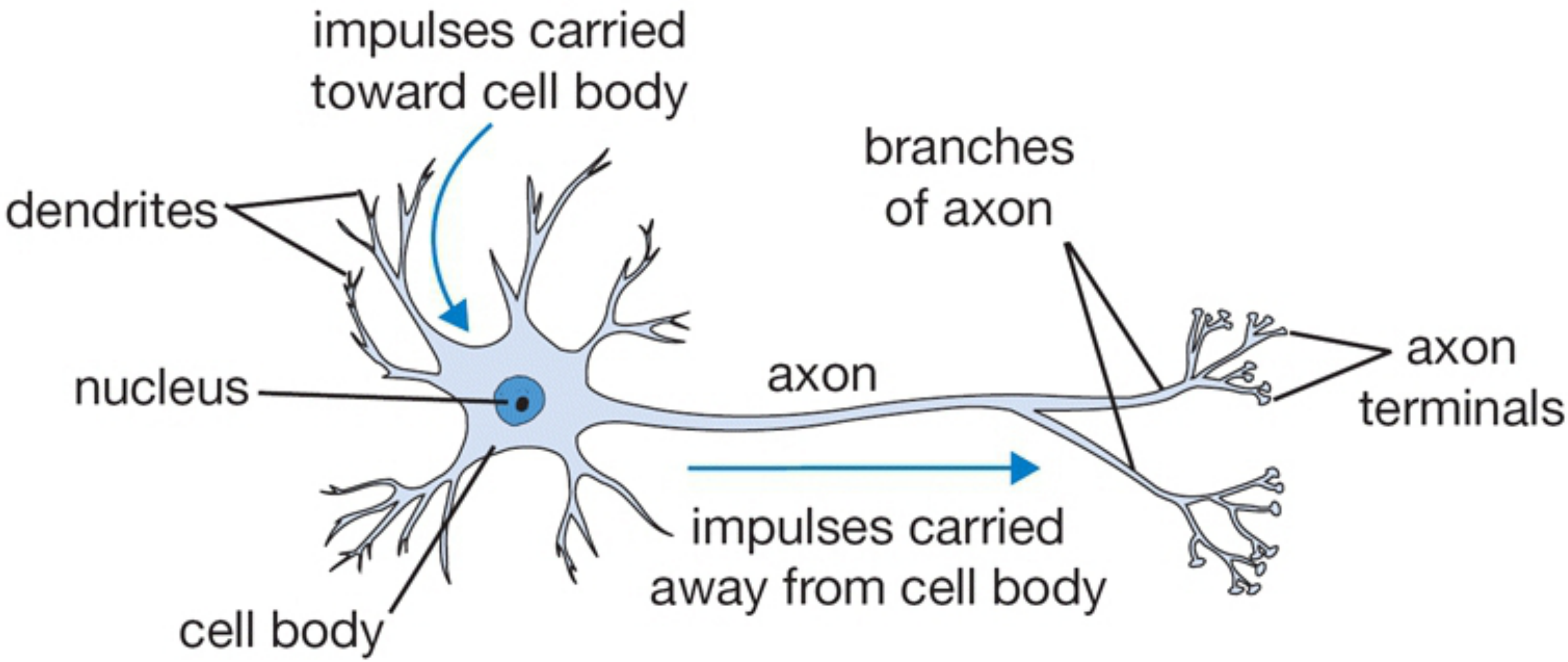
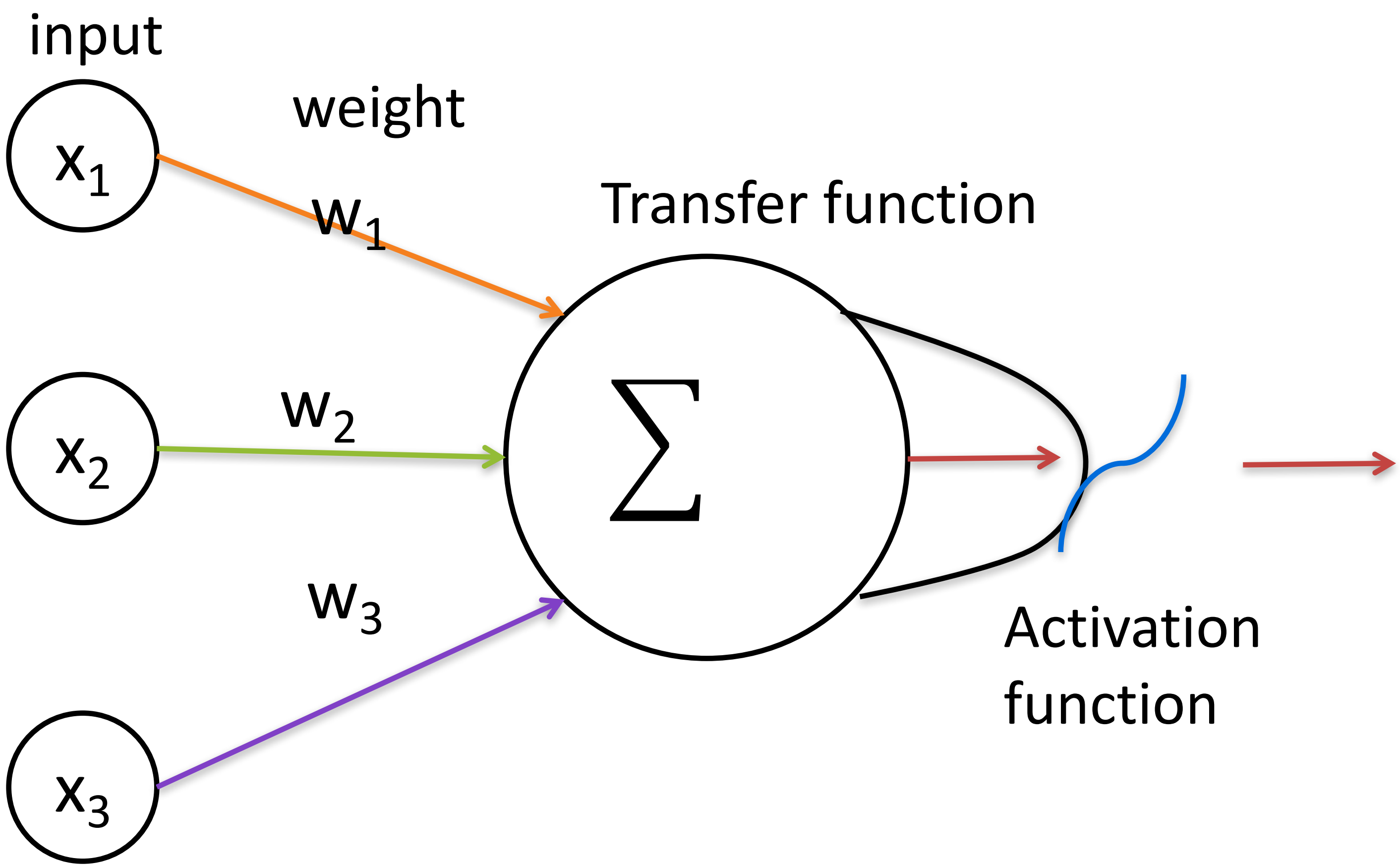


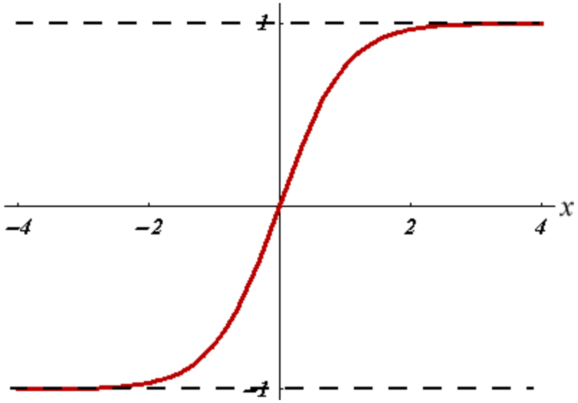
Image credit:
http://cs231n.github.io/neural-networks-1/

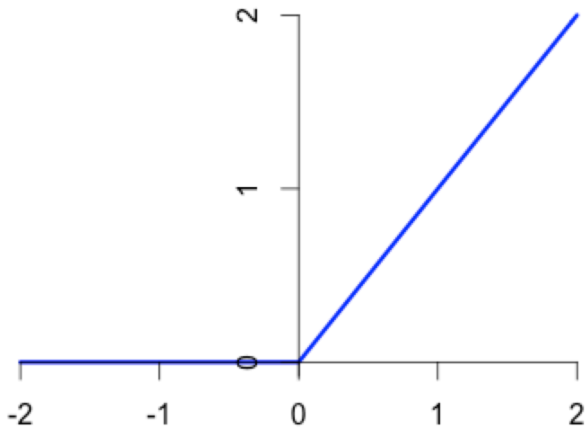
How to model a single artificial neuron?



Activation functions

$$\tanh(x) = \frac{e^{2x} - 1}{e^{2x} + 1} \qquad \mathrm{relu}(x) = \max(0, x)$$

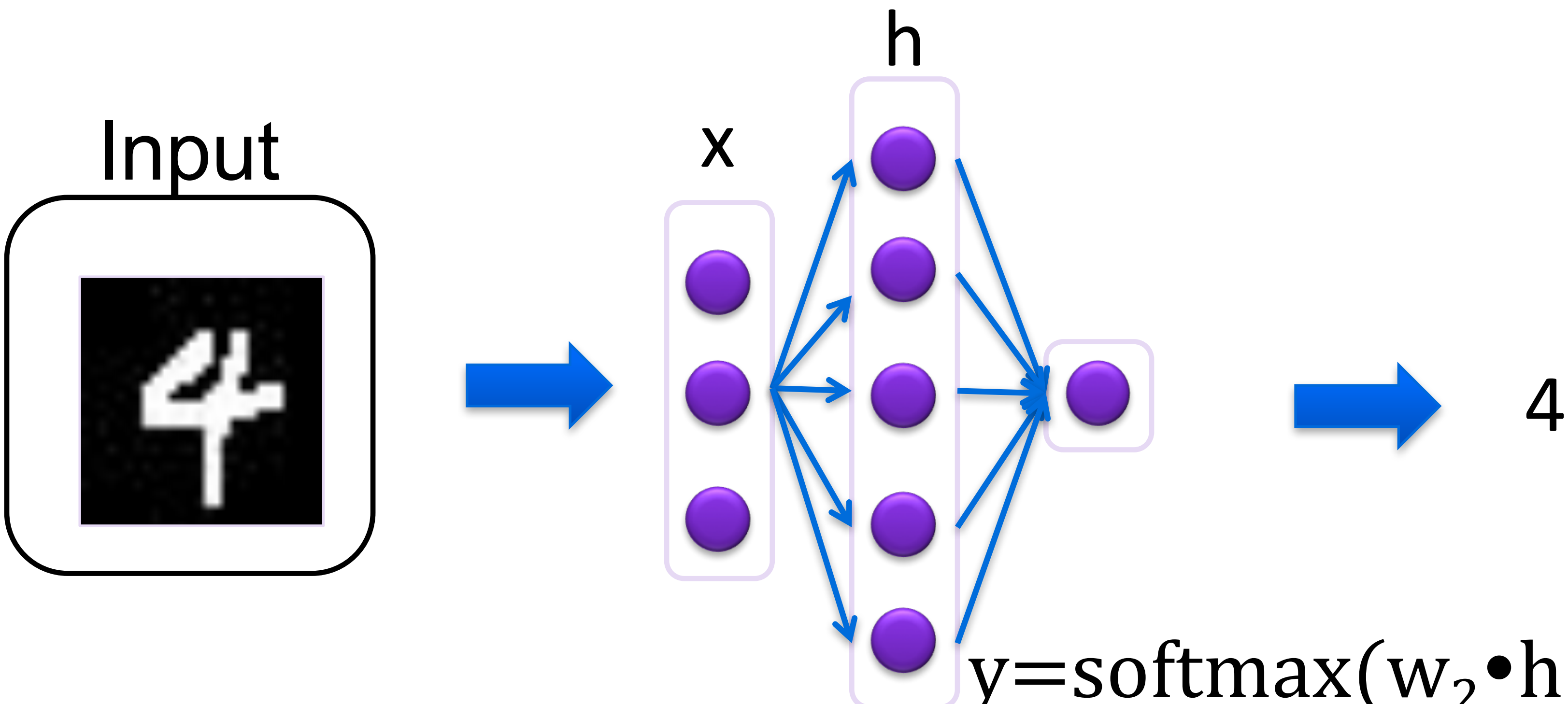




$$\mathrm{softmax}(x)_i = \frac{e^{x_i}}{\sum e^{x_i}}$$

Useful for modeling probability (in classification task)

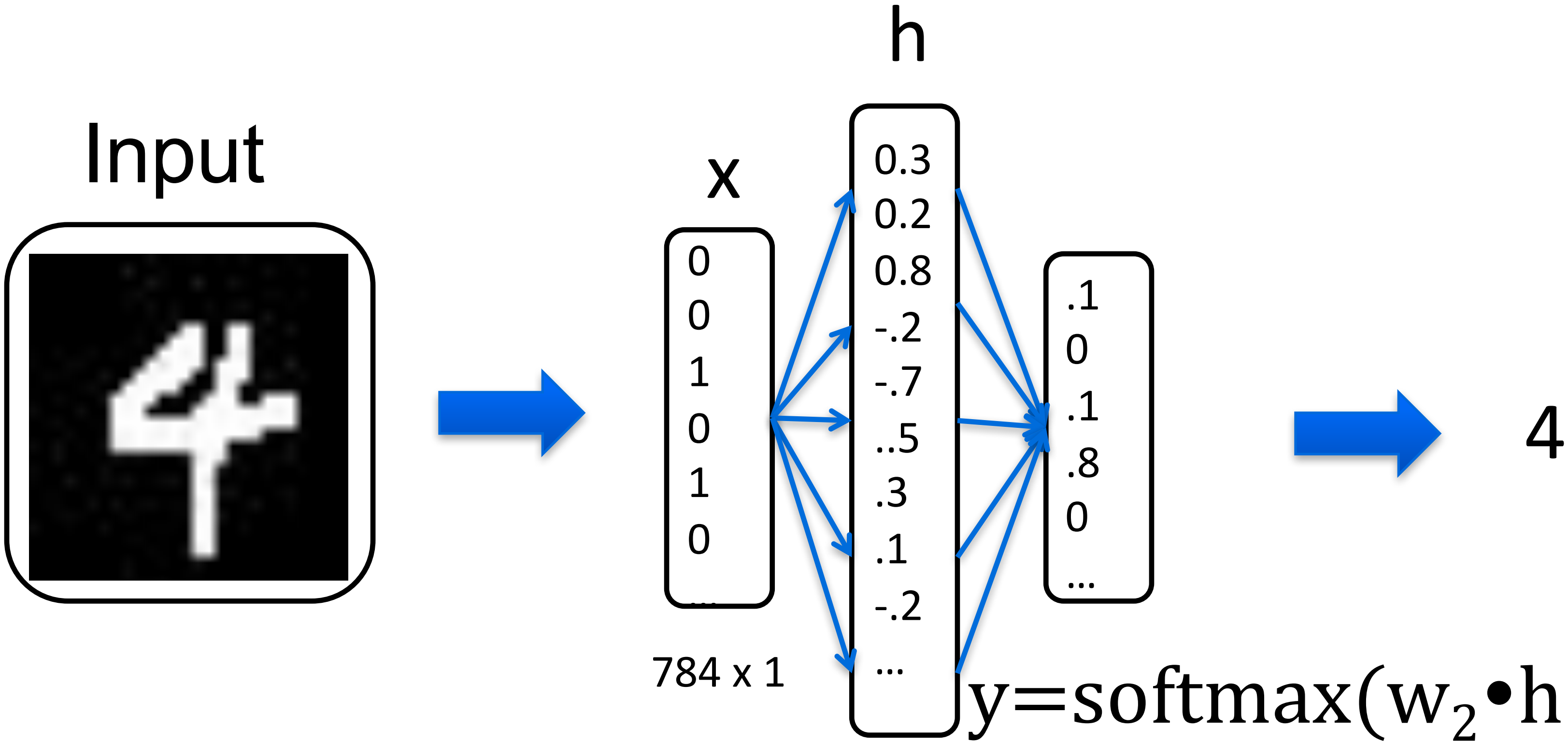
Supervised Learning with Neural Nets



$$y = \mathrm{softmax}(w_2 \bullet h_1 + b_2)$$

$$h_1 = \tanh(w_1 \bullet x + b_1)$$

Numerical Example

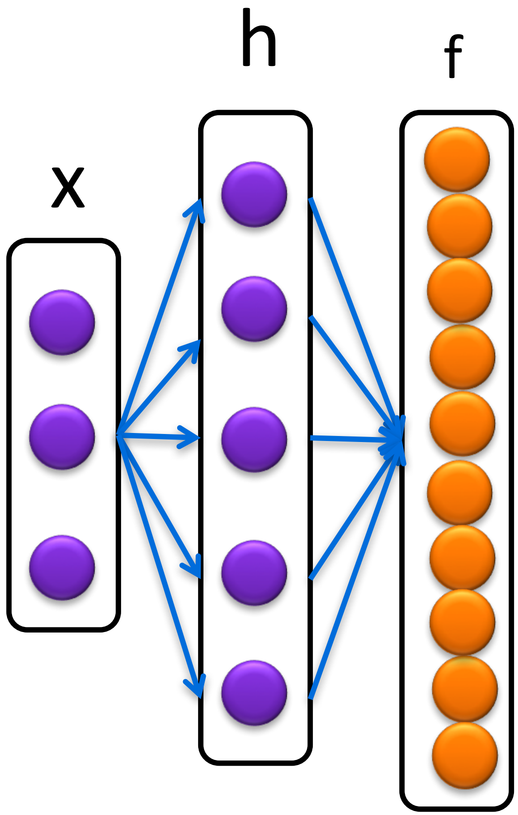


$$y = \text{softmax}(w_2 \bullet h_1 + b_2)$$

$$h_1 = \tanh(w_1 \bullet x + b_1)$$

w_1 : 256 x 784

Objective / Loss: cross-entropy



$$l(f(x_i), y_i) = -\log f(x_i)_{y_i}$$

$f(x_i)$ is a vector (e.g. $\in R^{10}$), representing predicted distribution

y_i is the ground-truth label, can be represented as an one-hot "distribution"
[0,…,0, 1, 0,…,0]

Cross-entropy

$$H(p,q) = -\sum_k p_k \log q_k$$

Cross-entropy

$$H(p,q) = -\sum_k p_k \log q_k$$

Average number of bits needed to represent message in q, while the actual message is distributed in p

OR. roughly
The information gap between p and q + (some const)

Minimizing cross-entropy == diminishing the information gap

$$H\big(y_i, f(x_i)\big) = -\sum_k y_{i,k} \log f(x_i)_k = -\log f(x_i)_{y_i}$$

Ideal case $f(x_i)_{y_i}$ ==> 1.0

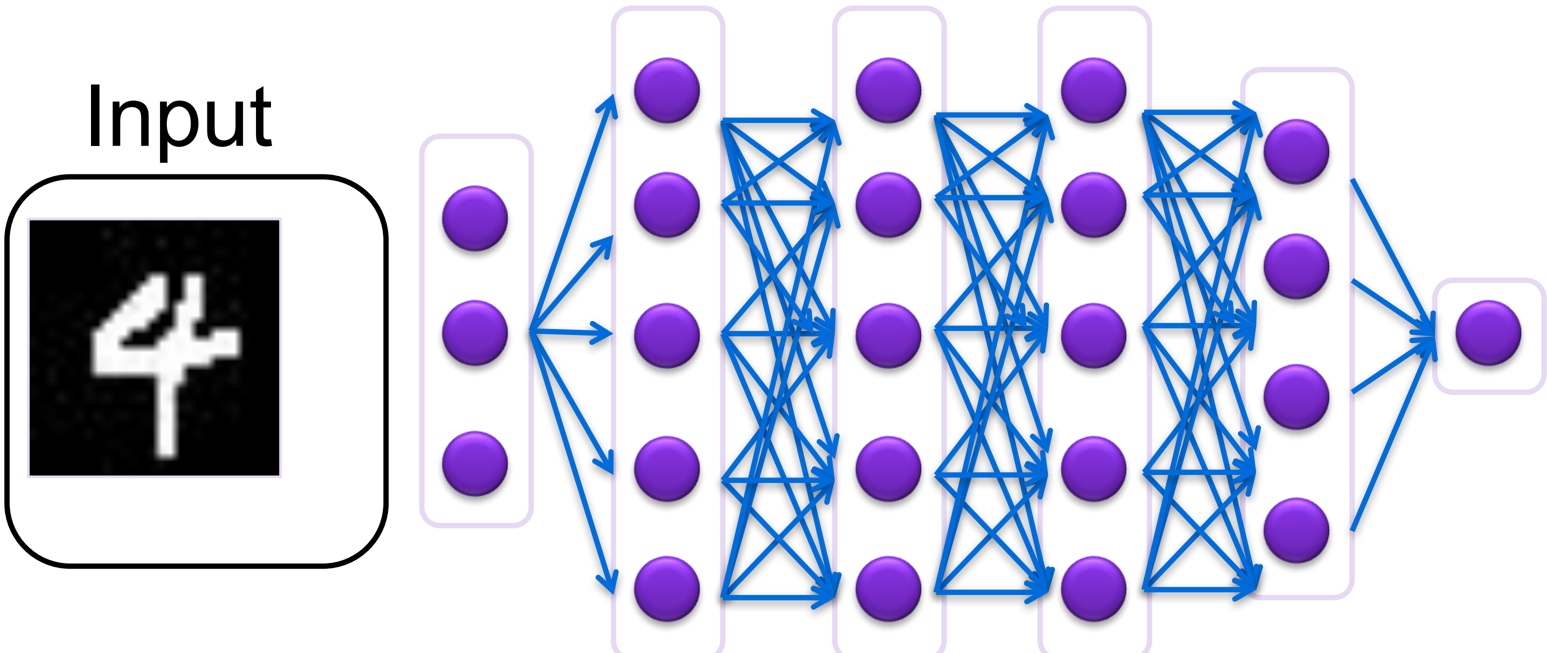
Alternative View: Max cond. log-likelihood

$$\boxed{\max\ \log p(y_i|x_i;w) = \sum_k y_{i,k} \log f(x_i)_k}$$

Or equivalently

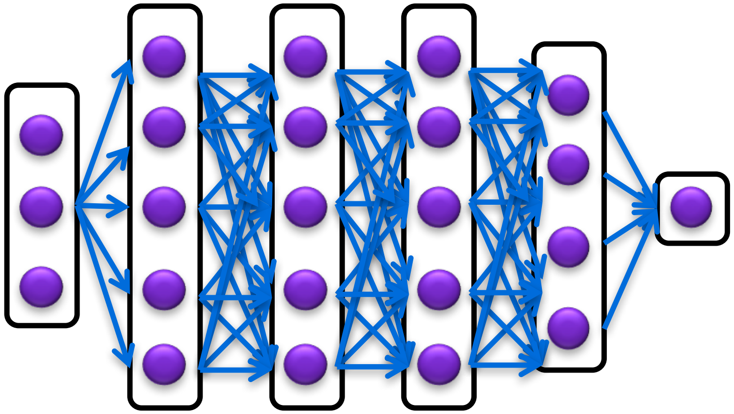
$$\min -\sum_k y_{i,k} \log f(x_i)_k$$

Deep Neural Nets



$$h_1 = \sigma_1(w_1 \bullet x + b_1) \quad h_2 = \sigma_2(w_2 \bullet h_1 + b_2)$$

Training DNN

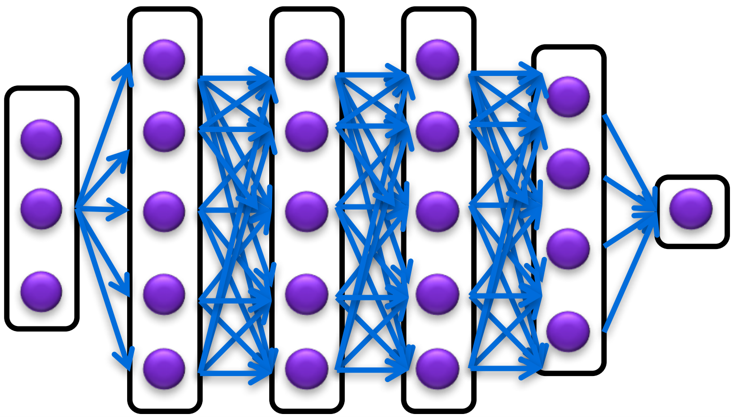


Given: N data points
$(x_1, y_1) \ldots (x_N, y_N)$

Goal: find the best model parameter w, to minimize cost

$$L(w) = \sum_{i=1}^{N} l(f(x_i, w), y_i)$$

Training deep neural nets



To improve efficiency:

Mini-Batch

Compute gradient and update parameters for every batch of k data samples.

Stochastic gradient descent algorithm

for iteration 1 to N (or until convergence)

compute g = $\partial/\partial w_j$

w = w – a • g

Step size

gradient

Advanced alg:

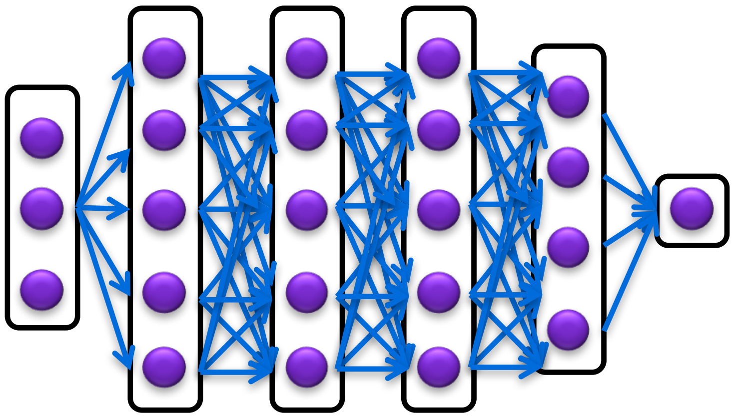
Momentum,

Adagrad,

Adam,

…

Forward and Backward propagation



Chain rule

forward pass: computing network prediction

$$h_i=\sigma_i(w_i \bullet h_{i-1})$$

backward prop: computing gradient from layer-wise error

$$\delta_{i-1}=w_i^T \bullet (\delta_i \odot \sigma_i')$$

$$\partial/\partial w_j=h_{i-1} \bullet \delta_i^T$$

More variation

- Optimization algorithms
  - Momentum
  - Adagrad
  - Adadelta
  - Adam
- Dropout
  - Randomly zeros the output neurons in each layer
- Regularization
  - L1,, L2, to improve generalization

Deep Learning platform

- Tensorflow (Google)
- Torch (NEC, FB)
- Caffe (ucb)
- Theano (U. Montreal)
- MXNet (DMLC, Li Mu et al)

- Provides easy language to construct network
- Rich set of layers, with forward and backward steps
- Library of optimization algorithms
- Many research papers build models based on these

Outline

- Problem Setup, Knowledge graphs
- Basic DL techniques
- Word, entity and relation embeddings
- Recurrent neural nets for processing sequence
- Focused Pruning: parsing the subject mention
- Finding the right relation and subject entity
- Other approaches:
  - LTG+CNN
  - Memory network

How to represent characters and words

Where was David Beckham born ?

Well-known methods: word2vec, Glove, etc.

Basic DL technology for language understanding

- Neural Language Model
  - Single layer NN for bigram, [Wei Xu and Alex Rudnicky, 2000]
  - Concatenated Word Embedding to predict next word [Yoshua Bengio, Réjean Ducharme, Pascal Vincent, Christian Jauvin, 2003]
  - RNN Language Model, [Mikolov et al, 2011]
- Basic NLP technology
  - NLP from scratch [Ronan Collerbert, Jason Weston et al 2011]
  - WSJ POS 97.29% acc; CoNLL NER 89.59% F1; CoNLL Chunking 94.32% F1

How to represent entities?

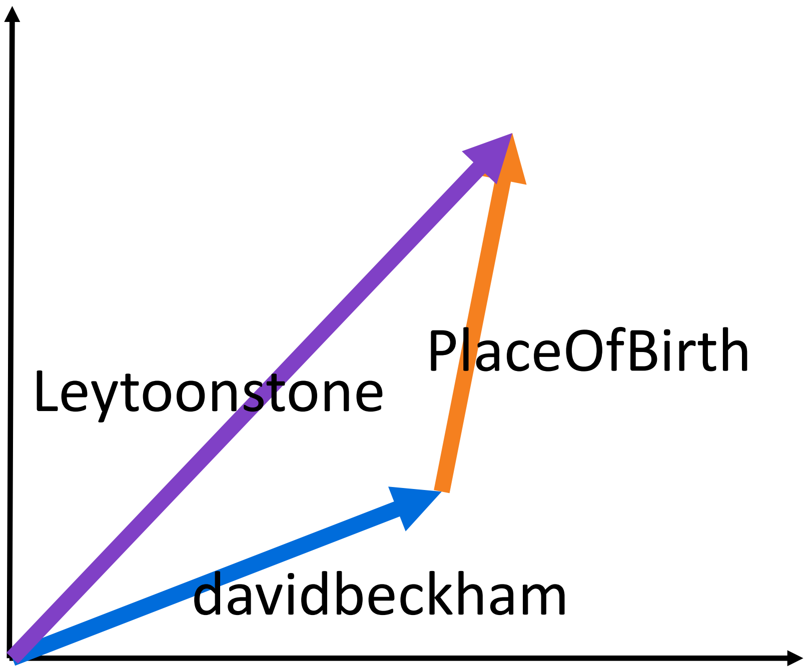
5 million entities in cleaned freebase

<DavidBeckham>

1. Random embedding
2. TransE trained embedding
3. zero-training embedding?

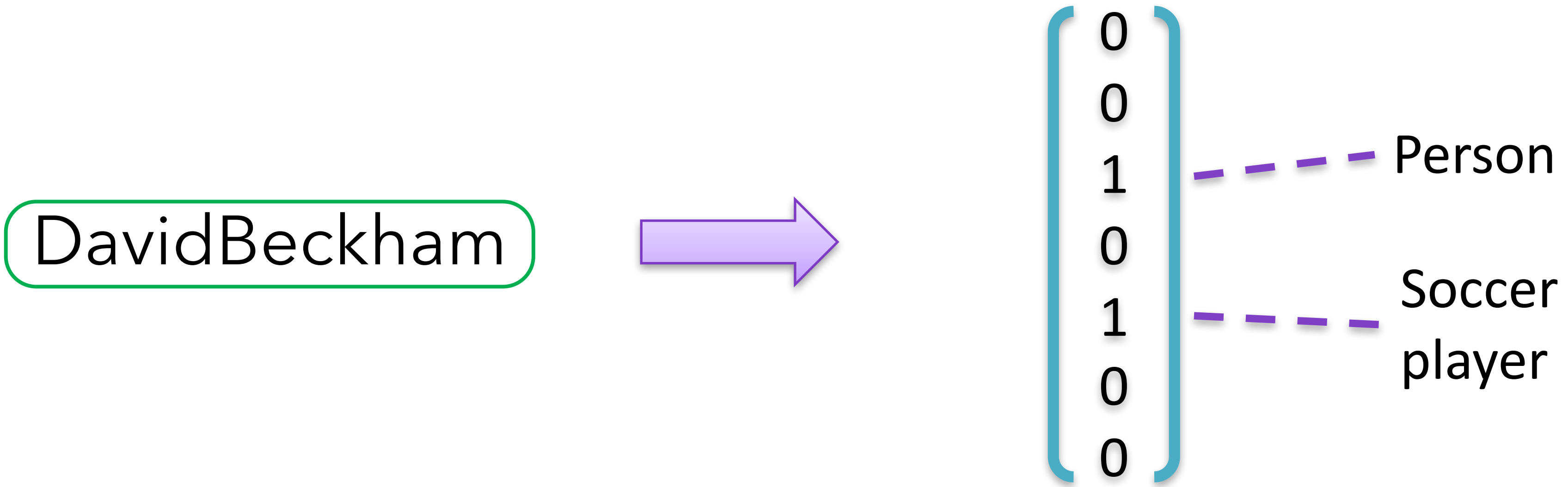
Learning entity embedding w/ TransE

<DavidBeckham, PlaceOfBirth, Leytonstone>



Zero-training embedding: Type-vector

- Benefits: no need to train, robust



Outline

- Problem Setup, Knowledge graphs
- Basic DL techniques
- Word, entity and relation embeddings
- Recurrent neural nets for processing sequence
- Focused Pruning: parsing the subject mention
- Finding the right relation and subject entity
- Other approaches:
  - LTG+CNN
  - Memory network

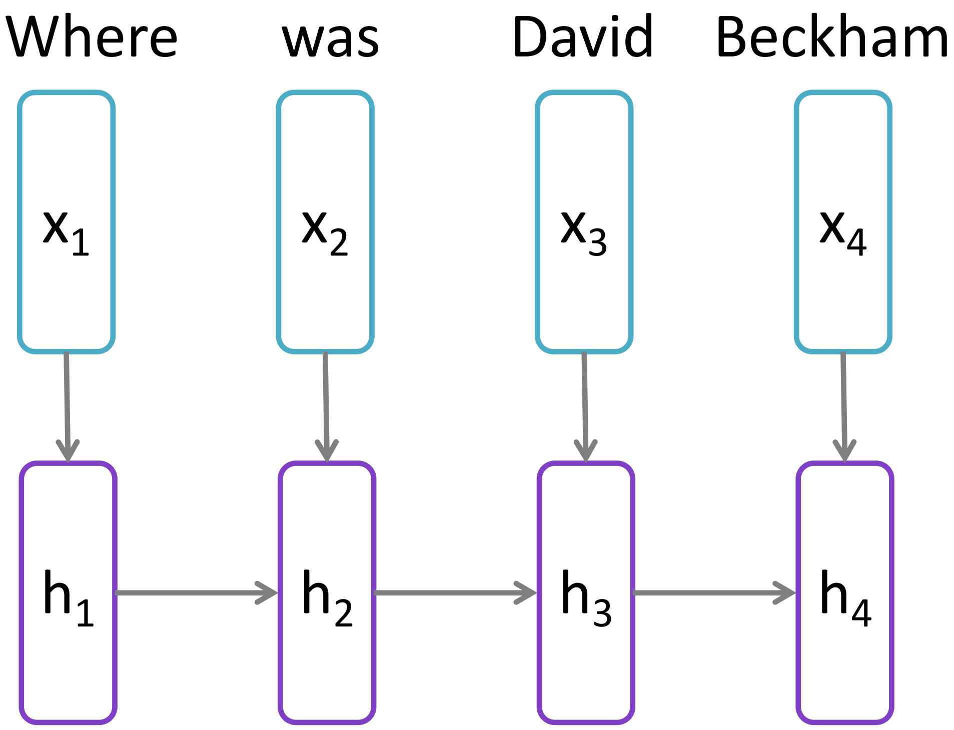
Challenge in processing language

- How to handle variable length of text sequences?
- Solution:
  - Adding Memory to Computation

Recurrent Neural Networks

Basic version: 1 fixed vector memory

- Remember previous state

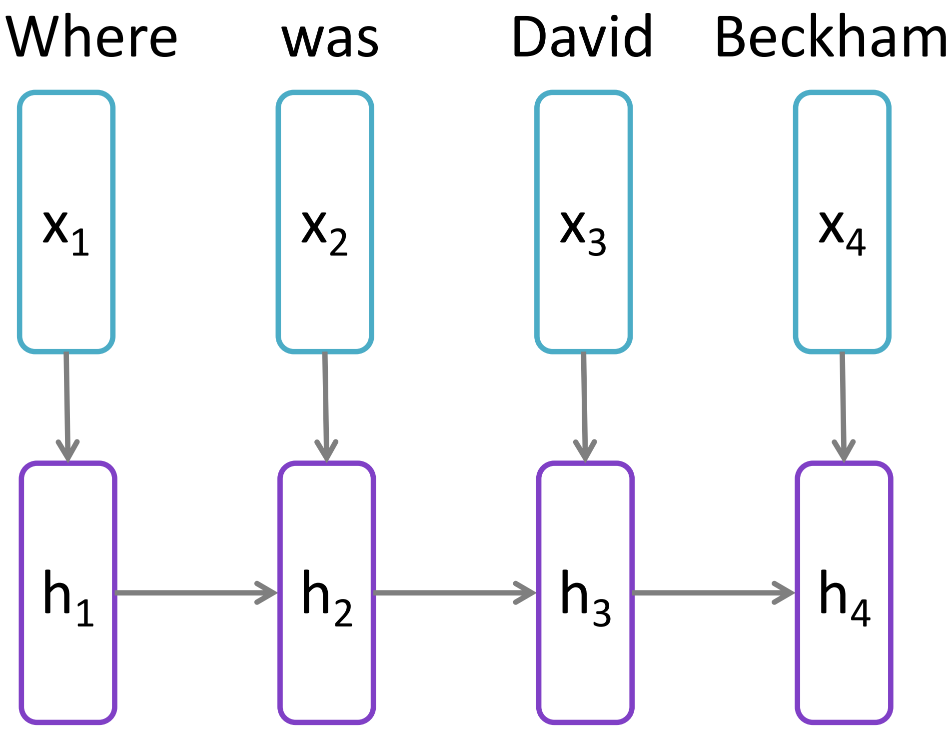


$h_t = f(W \bullet h_{t-1} + U \bullet x_t)$

f = sigmoid, tanh, relu

Recurrent Neural Networks

- Remember previous state

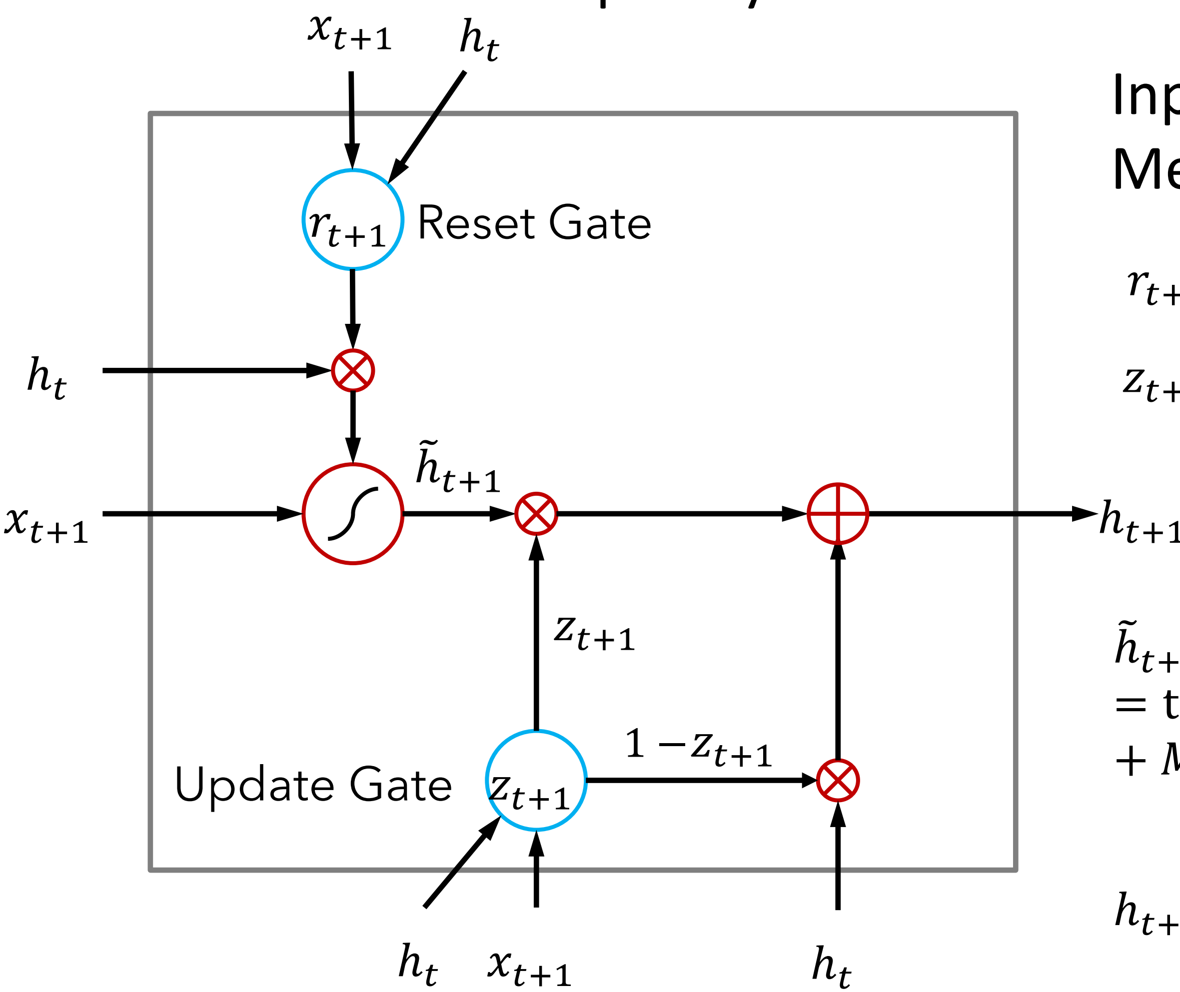


$$h_t = f(W \bullet h_{t-1} + U \bullet x_t)$$

f = sigmoid, tanh, relu

Gated recurrent unit

Adaptively memorize short and long term information



Input: x_t

Memory: h_t

$$r_{t+1} = \sigma(M_{rx} x_{t+1} + M_{rh} h_t + b_r)$$

$$z_{t+1} = \sigma(M_{zx} x_{t+1} + M_{zh} h_t + b_z)$$

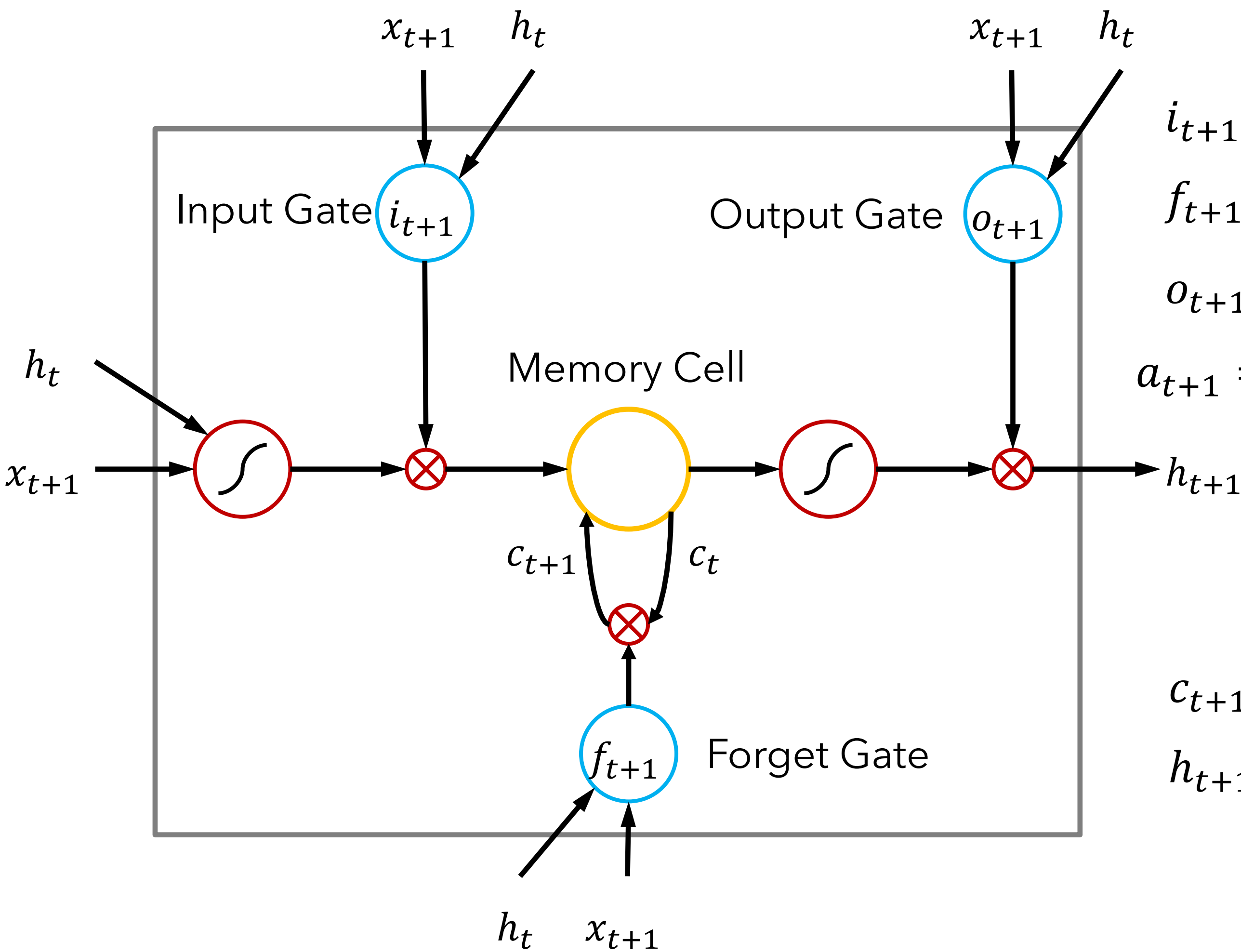
$$\tilde{h}_{t+1} = \tanh(M_{hx} x_{t+1} + M_{hh}(r_{t+1} \otimes h_t) + b_h)$$

$$h_{t+1} = z_{t+1} \otimes \tilde{h}_{t+1} + \boldsymbol{(1 - z_{t+1}) \otimes h_t}$$

[Chung et al 2014]

Long-Short Term Memory (LSTM)

Adaptively memorize short and long term information



$$i_{t+1} = \sigma(M_{ix}x_{t+1} + M_{ih}h_t + b_i)$$

$$f_{t+1} = \sigma(M_{fx}x_{t+1} + M_{fh}h_t + b_f)$$

$$o_{t+1} = \sigma(M_{ox}x_{t+1} + M_{oh}h_t + b_o)$$

$$a_{t+1} = \tanh(M_{cx}x_{t+1} + M_{ch}h_t + b_a)$$

$$c_{t+1} = \boldsymbol{f_{t+1} \otimes c_t} + i_{t+1} \otimes a_{t+1}$$

$$h_{t+1} = o_{t+1} \otimes \tanh(c_{t+1})$$

Outline

- Problem Setup, Knowledge graphs
- Basic DL techniques
- Word, entity and relation embeddings
- Recurrent neural nets for processing sequence
- Focused Pruning: parsing the subject mention
- Finding the right relation and subject entity
- Other approaches:
  - LTG+CNN
  - Memory network

From natural language question to structured query

Question

Where was David Beckham born?

subject

relation

SPARQL query

```
SELECT ?object
WHERE { <DavidBeckham> <PlaceOfBirth> ?object }
```

Finding subject mention: simple heuristics fails

"What theme is the book the armies of memory?"

- the book: 73
- theme: 252
- memory: 553
- ……

Finding subject mention with focused pruning

$$p(s, r \mid q) = \sum_{f \in \mathcal{F}(q)} p(s, r, f \mid q)$$

Using RNN sequence labelling model

Focus: Where was David Beckham born?

Prob.: 0.05 0.85 0.01

Finding focus ~ sequence labelling

Wuhan Tech University's nearby handmade noodle house

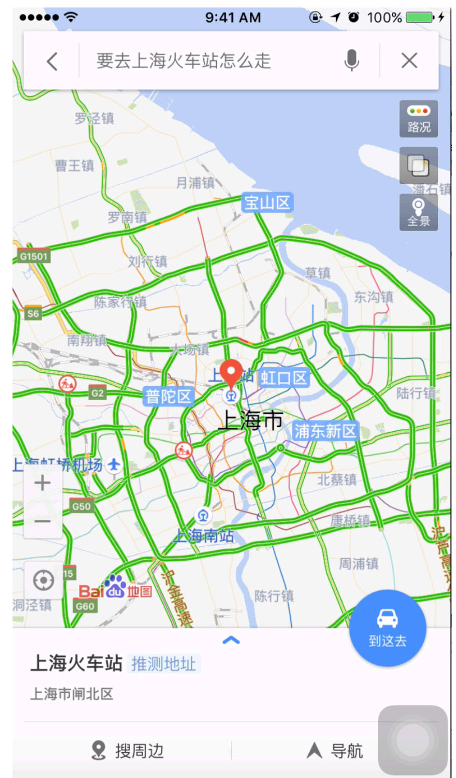
武汉理工大学附近的拉面馆

center keywords

how to go from shanghai to hangzhou

上海到杭州开车怎么走

origin destination



A Sequence Labelling Task
Named entity recognition

date Location

In **April 1775** fighting broke out between **Massachusetts** militia units and **British** regulars at **Lexington** and **Concord** .

Geo-Political

Named entity recognition

三	藩	市	长	李	孟	贤	…
1640	897	45	1890	78	943	3521	
B-GPE	I-GPE	O	O	B-PER	I-PER	I-PER	

Entity chunking scheme: B-I-O Begin of entity chunk, In-middle-of entity chunk, Other (not entity)

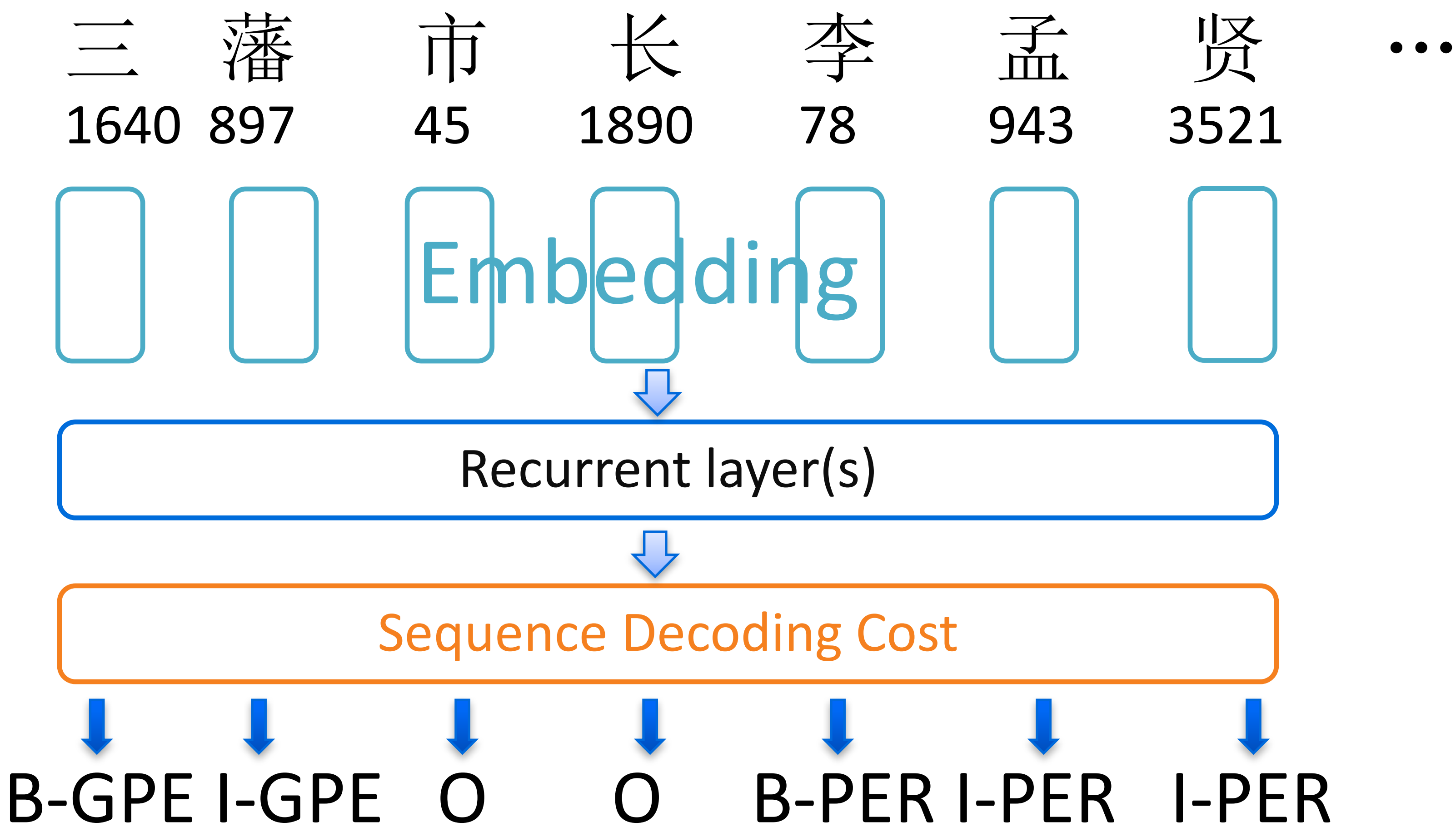
Traditional approach

- Conditional random fields with rich expert created features.

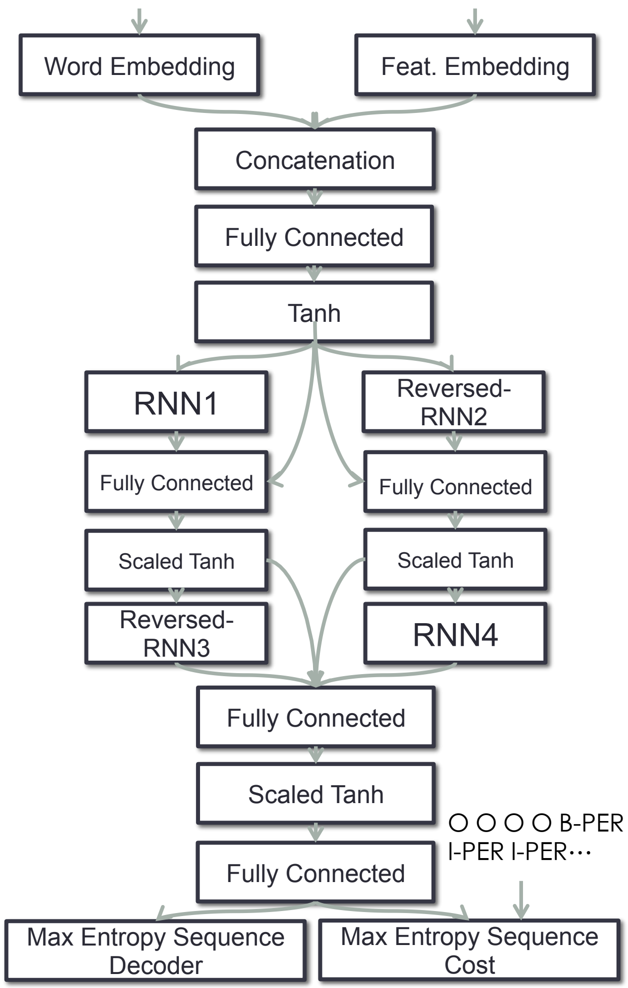
三	藩	市	长	李	孟	贤	…
1640	897	45	1890	78	943	3521	

Features: neighboring words,
POS of current word and neighboring words,
Lexical features etc.

End-to-end training with minimal linguistic features



Complete NER Model



Chinese NER

OntoNotes Data 4-class:

Model	P	R	F1
Bi-NER-WA* Wang et al.	84.42	76.34	80.18
RNN-2b with WS ours	84.75	77.85	81.15

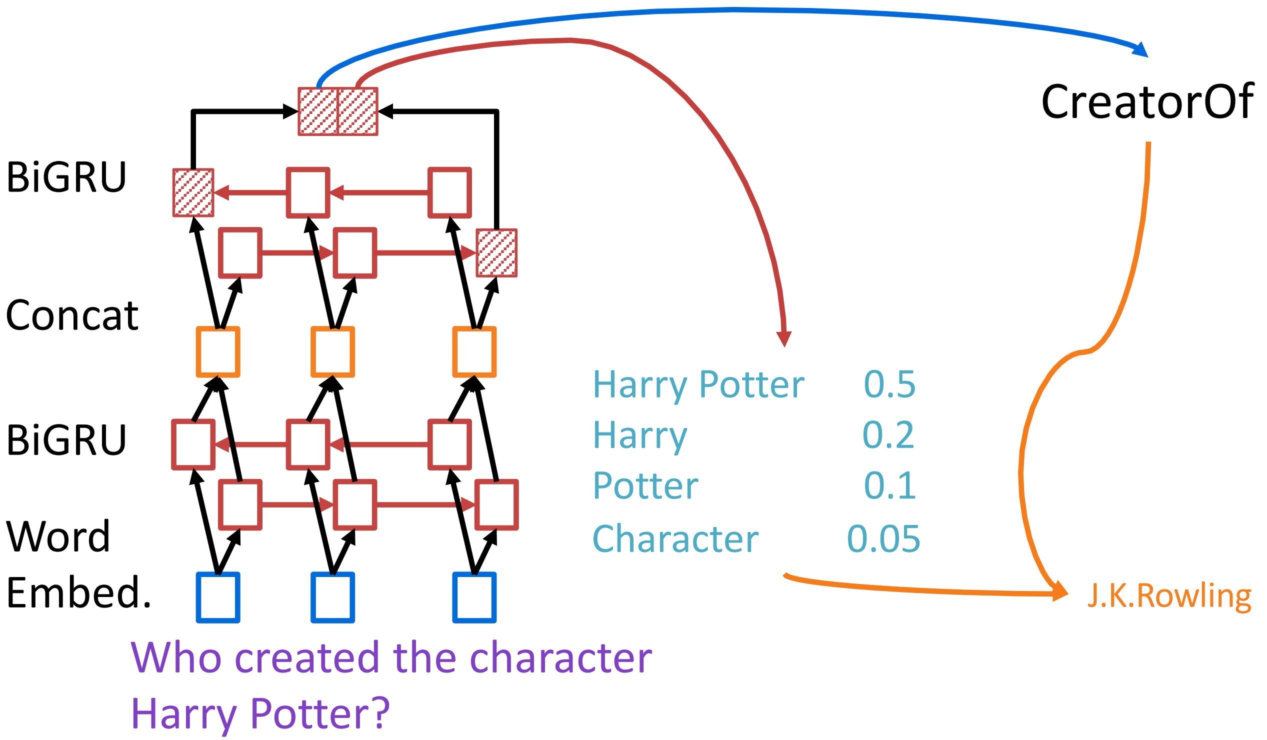
* Wang et al used bilingual data

OntoNotes Data 18-class:

Model	P	R	F1
Sameer Pradhan et al.	78.20	66.45	71.85
RNN-2b with WS ours	78.69	70.54	74.39

[Zefu Lu, Lei Li, Wei Xu, 2015]

Stacked bi-directional GRU for sentence embedding



CFO: Conditional Focused Neural Question Answering with Large-scale Knowledge Bases
[Zihang Dai, Lei Li, Wei Xu, ACL 2016]

Answers by our CFO system:

哈利波特在哪儿上的学？ Which school did Harry Potter attend?

霍格沃兹魔法学校 Hogwarts School of Witchcraft and Wizardry

格罗格里小学 Gregory Primary school

哈利波特是谁写的？ Who created Harry Potter?

罗琳女士 J.K. Rowling

罗琳的写作风格受谁影响？ Who influenced J.K. Rowling?

乔治艾略特 George Eliot

史蒂文金 Stephen King

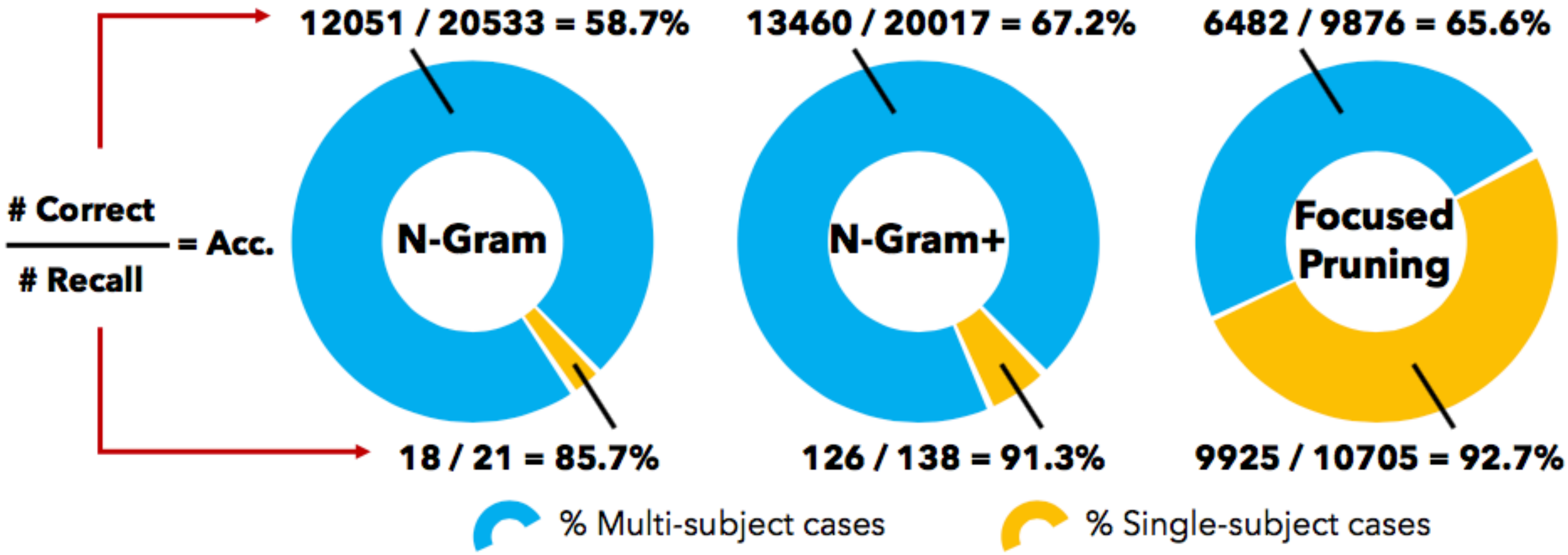
史蒂文金写了什么小说? What books did Stephen King write?

Las cuatro estaciones/different seasons

肖生克的救赎

[Dai, Li, Xu, 2016]

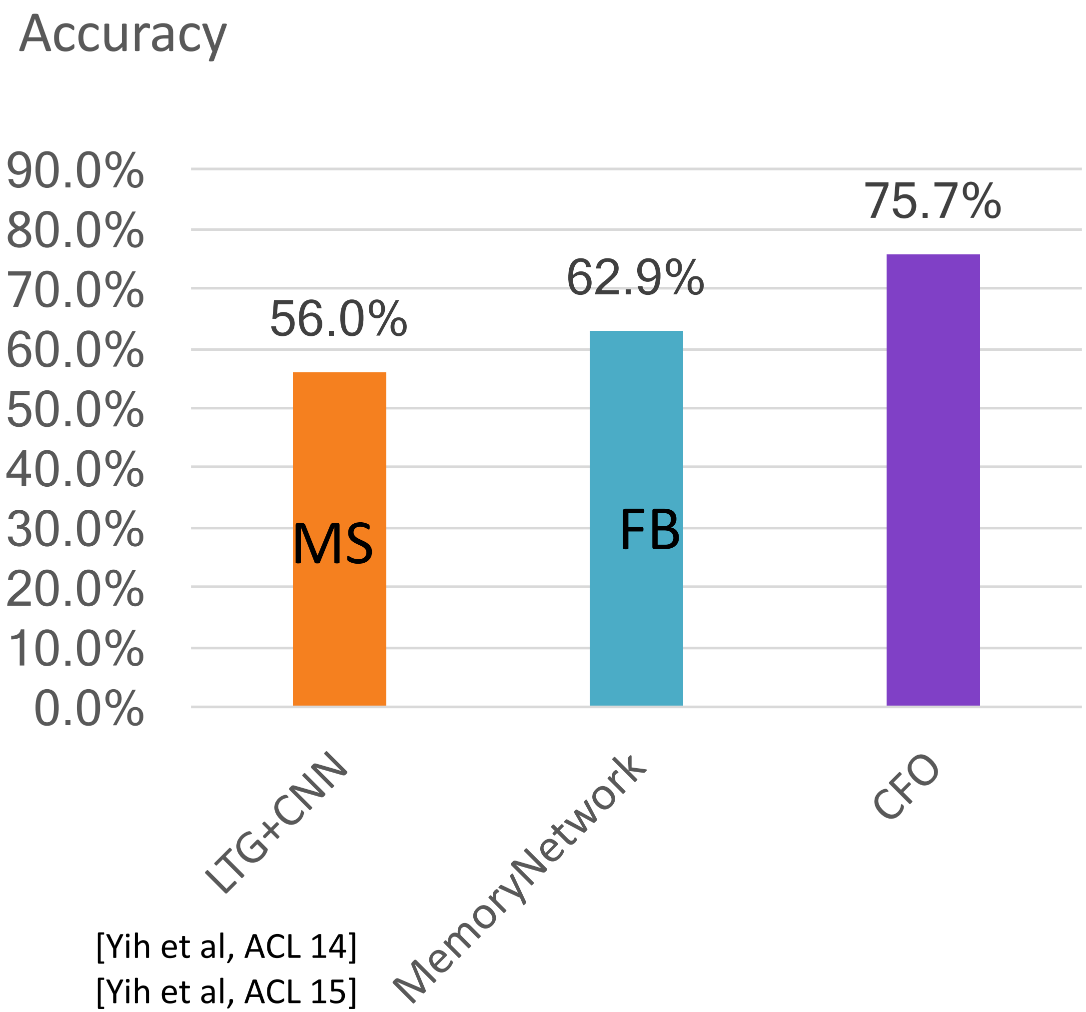
Does focus help?



Evaluation Results

Pruning Method	**Relation Network**	**Entity Representation**		
		Random	Pretrain	Type Vec
Memory Network [3]		62.9	63.9*	
N-Gram	Embed-AVG	39.4	42.2	50.9
	LTG-CNN	32.8	36.8	45.6
	BiGRU	43.7	46.7	55.7
N-Gram+	Embed-AVG	53.8	57.0	58.7
	LTG-CNN [1,2]	46.3	50.9	56.0
	BiGRU	58.3	61.6	62.6
Focused Pruning	Embed-AVG	71.4	71.7	72.1
	LTG-CNN	67.6	67.9	68.6
	LTG-CNN+	70.2	70.4	71.1
	BiGRU	75.2	75.5	**75.7**

Comparison



[Yih et al, ACL 14]
[Yih et al, ACL 15]

[Bordes et al 2015]

Conclusion

- Problem Setup, Knowledge graphs
- Basic DL techniques
- Word, entity and relation embeddings
- Recurrent neural nets for processing sequence
- Focused Pruning: parsing the subject mention
- Finding the right relation and subject entity
- Other approaches:
  - LTG+CNN
  - Memory network

Thanks!

Contact： Lei Li (lileilab@toutiao.com)

Joint work with

Zihang Dai (CMU): QA
Wei Xu (Baidu IDL)

Toutiao Lab is Hiring!

头条 今日头条

Research Scientist and Software Engineer in

Machine Learning

Natural Language Understanding

Computer Vision

http://www.toutiao.com/lab

lab-hr@toutiao.com

Reference

Parsing & Sequence labelling

- Collobert et al, Natural language processing almost from scratch
- Lu et al, Twisted recurrent network for named entity recognition
- Huang et al, Bidirectional LSTM-CRF models for sequence tagging
- Zhou et al, End-to-end learning of semantic role labeling using recurrent neural networks.

Question Answering

- Jonathan Berant, Andrew Chou, Roy Frostig, Percy Liang Semantic Parsing on Freebase from Question-Answer Pairs, EMNLP 2013.
- W. Yih, X. He & C. Meek. Semantic Parsing for Single-Relation Question Answering. In ACL-14.
- W. Yih, M. Chang, X. He & J. Gao. Semantic Parsing via Staged Query Graph Generation: Question Answering with Knowledge Base. In ACL-IJCNLP-2015
- Antoine Bordes, Nicolas Usunier, Sumit Chopra, Jason Weston, Large-scale Simple Question Answering with Memory Networks, 2015
- Zihang Dai, Lei Li, Wei Xu, CFO: Conditional Focused Question Answering with Large Knowledge-bases. ACL 2016