Knowledge Acquisition

Ni Lao, Xipeng Qiu
8/28/2017

Everything presented here is publicly available.
The opinions stated here are my own, not those of Google.
Partial Slides in the section of "Information Extraction" are provided by Dr Kang Liu.
Where does knowledge come from?

- The human brain contains roughly 100 billion neurons each capable of making around 1,000 connections.
- Where do we get these 100 TB parameters?
- How many lines of code do I need to write if I want to achieve AI?
The Mind’s Eye

The world

a small machine which can copy large amount of complexity from the world to the brain

a suitable representation
Mandelbrot Set

the nature of complex numbers

\[ z_0 = 0 \]
\[ z_{n+1} = z_n^2 + c \]

\[ c \in M \iff \lim_{n \to \infty} |z_{n+1}| \leq 2 \]
ACL 2017

- The premier conference in natural language understanding
- Vancouver, Canada
- 1,318 submissions, 302 accepted
- 1700 attendees
(A)CL is booming!
Recent changes

- More IE, summarization/generation, and dialogue
In perspective

- Sequence models are the most advanced and impactful, which ML has ever offered
  - A lot of recent effort in adding memory, but no impact yet
Seq2seq models
Lapata’s scream

WHAT NOW? IS NLP DEAD?
Plan

- Information extraction
- Semantic parsing
- Semantic representation
Information Extraction

● Has its root in DARPA
  ○ An intelligent agent monitoring a news data feed requires IE to transform unstructured data into something that can be reasoned with, e.g., (PERSON, works_for, ORGANIZATION)
Information Extraction

● The result technologies can only be applied to restricted domains
  ○ Supervised training is limited by labeled data
    ■ (Zhou et al., 2005; Zhou et al., 2007; Sur-deanu and Ciaramita, 2007)
  ○ Unsupervised approaches can extract very large numbers of triple, but may not be easy to map to relations needed
    ■ (Shinyama and Sekine, 2006; Banko et al., 2007)
  ○ Distantly supervision is scalable, but still limited by the KB schema
    ■ (Mints et al., 2009)
Problem Formulation

Entity->Relation->Event

1. Named Entity Recognition (NER)
2. Relation Classification: Binary and N-ary
3. Event Extraction
Neural Architectures for Named Entity Recognition

Supervised training

Neural Architecture: LSTM+CRF

https://github.com/clab/stack-lstm-ner
Extract Relations from Unstructured Text

姚明

姚明 1980年生于上海，美国NBA及世界篮球巨星。中国篮球史上里程碑式人物。CBA上海队老板。曾效力于中国国家篮球队，NBA火箭队。2011年7月20日退役。获7次NBA“全明星”，被美国《时代周刊》列入“世界最具影响力100人”，被中国体育总局授予“体育运动荣誉奖章”“中国篮球杰出贡献奖”。姚明以高超球技，顽强进取精神，谦逊幽默气质与人格魅力，赢得了世界声誉。让世界对中国有了新的了解与认识；让更多的人关注、喜爱篮球。姚明成为东西方文化的桥梁，具有史无前例的个人影响力。姚明的意义在于，超越了篮球运动，超越了国界。

（姚明，国籍，中国）

实体1 关系名（属性名） 实体2
Extract Relations from Unstructured Text

Sentence Level

The $[\text{haft}]_{e1}$ of the $[\text{axe}]_{e2}$ is made of yew wood.

Component-Whole(e1,e2)

The $[\text{fire}]_{e1}$ inside WTC was caused by exploding $[\text{fuel}]_{e2}$.

Cause-Effect(e1,e2)
Extract Relations from Unstructured Text

Corpus Level

At-least-one Hypothesis
If two entities participate in a relation, at least one sentence that mentions these two entities might express that relation.

Evidence
Steve Jobs was the co-founder and CEO of Apple and formerly Pixar.
Steve Jobs passed away the day before Apple unveiled iPhone 4S.
......
“Steve Jobs was the co-founder and CEO of Apple and formerly Pixar.”
Matrix-Vector Recursive Neural Network for Relation Classification
Convolutional Neural Network
Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification

Zhou, ACL 2016
## Performances

**SemEval-2010 Task 8**

- # of training instance 8,000
- # of test instance 2,717
- # of relationships 19

<table>
<thead>
<tr>
<th>Model</th>
<th>Feature Set</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>POS, prefixes, morphological, WordNet, dependency parse, Levin classed, ProBank, FramNet, NomLex-Plus, Google n-gram, paraphrases, TextRunner</td>
<td>82.2</td>
</tr>
<tr>
<td>CNN</td>
<td>WV (Turian et al., 2010) (dim=50) + PF + WordNet</td>
<td>69.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>82.7</td>
</tr>
<tr>
<td>RNN</td>
<td>WV (Turian et al., 2010) (dim=50) + PI</td>
<td>80.0</td>
</tr>
<tr>
<td></td>
<td>WV (Mikolov et al., 2013) (dim=300) + PI</td>
<td>82.5</td>
</tr>
<tr>
<td>SDP-LSTM</td>
<td>WV (pretrained by word2vec) (dim=200), syntactic parse + POS + WordNet + grammar relation embeddings</td>
<td>82.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>83.7</td>
</tr>
<tr>
<td>BLSTM</td>
<td>WV (Pennington et al., 2014) (dim=100) + PF + POS + NER + WNSYN + DEP</td>
<td>82.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>84.3</td>
</tr>
<tr>
<td>BLSTM</td>
<td>WV (Turian et al., 2010) (dim=50) + PI</td>
<td>80.7</td>
</tr>
<tr>
<td>Att-BLSTM</td>
<td>WV (Turian et al., 2010) (dim=50) + PI</td>
<td>82.5</td>
</tr>
<tr>
<td>BLSTM</td>
<td>WV (Pennington et al., 2014) (dim=100) + PI</td>
<td>82.7</td>
</tr>
<tr>
<td>Att-BLSTM</td>
<td>WV (Pennington et al., 2014) (dim=100) + PI</td>
<td>84.0</td>
</tr>
</tbody>
</table>
Joint Extraction of Entities and Relations

Input Sentence: The United States President Trump will visit the Apple Inc founded by Steven Paul Jobs

Tags: 0 B-CP-1 E-CP-1 0 S-CP-2 0 0 B-CF-1 E-CF-1 0 0 B-CF-2 I-CF-2 E-CF-2

Final Results: {United States, Country-President, Trump} {Apple Inc, Company-Founder, Steven Paul Jobs}

Number of tags: $2 \times 4 \times |R| + 1$

$|R|$ is the number of relation, 4 means begin, end, single, inside
Distant Supervision for Relation Extraction

Distant supervision automatically generates amount of training data, overcome the manually-labeling problem.

Knowledge base

<table>
<thead>
<tr>
<th>Relation</th>
<th>Entity 1</th>
<th>Entity 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Founder</td>
<td>Steve Jobs</td>
<td>Apple</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Sentence

Steve Jobs was the co-founder and CEO of Apple and formerly Pixar.

Steve Jobs passed away the day before Apple unveiled iPhone 4S.

...
Multi-instance Learning

- T Bags
- i-th bag has \( q_i \) instances

\[
M_i = \{ m^1_i, m^2_i, \ldots m^{q_i}_i \}
\]

- Objective function:

\[
J(\theta) = \sum_{i=1}^{T} \log p(y_i | m^j_i; \theta)
\]

where

\[
j^* = \arg\max_j p(y_i | m^j_i; \theta) \quad 1 \leq j \leq q_i
\]

Algorithm 1 Multi-instance learning

1. Initialize \( \theta \). Partition the bags into mini-batches of size \( b_s \).
2. Randomly choose a mini-batch, and feed the bags into the network one by one.
3. Find the \( j \)-th instance \( m^j_i \) \((1 \leq i \leq b_s)\) in each bag according to Eq. (9).
4. Update \( \theta \) based on the gradients of \( m^j_i \) \((1 \leq i \leq b_s)\) via Adadelta.
5. Repeat steps 2-4 until either convergence or the maximum number of epochs is reached.
Piece-wise CNN Model

Vector representation

Convolution

Piecewise max pooling

Softmax classifier

Zeng, EMNLP 2015
Selective Attention over Instances

Selective Attention

$$\alpha_i = \frac{\exp(e_i)}{\sum_k \exp(e_k)}$$

$$e_i = x_i A r$$

$A$ is a weighted diagonal matrix

$r$ is the query vector associated with relation $r$

The final set vector $s$

$$s = \sum_i \alpha_i x_i$$
### Case Study

#### Employer of

<table>
<thead>
<tr>
<th>Relation</th>
<th>Subject</th>
<th>Object</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>When Howard Stern was preparing to take his talk show to <strong>Sirius Satellite Radio</strong>, following his former boss, <strong>Mel Karmazin</strong>, Mr. Hollander argued that ...</td>
<td></td>
</tr>
<tr>
<td>High</td>
<td><strong>Mel Karmazin</strong>, the chief executive of <strong>Sirius Satellite Radio</strong>, made a lot of phone calls ...</td>
<td></td>
</tr>
</tbody>
</table>

#### Place of Birth

<table>
<thead>
<tr>
<th>Relation</th>
<th>Subject</th>
<th>Place</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td><strong>Ernst Haefliger</strong>, a Swiss tenor who ... roles, died on Saturday in <strong>Davos</strong>, Switzerland, where he maintained a second home.</td>
<td></td>
</tr>
<tr>
<td>High</td>
<td><strong>Ernst Haefliger</strong> was born in <strong>Davos</strong> on July 6, 1919, and studied at the Wettinger Seminary ...</td>
<td></td>
</tr>
</tbody>
</table>
From Static Knowledge to Dynamic Knowledge

Dynamic Knowledge: Event-Centric Knowledge Graph
Barry Diller on Wednesday *quit* as chief of *Vivendi Universal Entertainment*.

<table>
<thead>
<tr>
<th>Trigger</th>
<th>Quit (a “Personnel/End-Position” event)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arguments</td>
<td>Role = Person</td>
</tr>
<tr>
<td></td>
<td>Role = Organization</td>
</tr>
<tr>
<td></td>
<td>Role = Position</td>
</tr>
<tr>
<td></td>
<td>Role = Time-within</td>
</tr>
</tbody>
</table>
Definition of Event Extraction

Definition:
Event trigger, Event Type, Event argument, Argument role

Example: Barry Diller on Wednesday quit as chief of Vivendi Universal Entertainment.

1. Event Identification (Trigger Words)
2. Event Type Identification
3. Argument Identification
4. Argument Role Identification

<table>
<thead>
<tr>
<th>Trigger</th>
<th>Role = Person</th>
<th>Role = Organization</th>
<th>Role = Position</th>
<th>Role = Time-within</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quit (a &quot;Personnel/End-Position&quot; event)</td>
<td>Barry Diller</td>
<td>Vivendi Universal Entertainment</td>
<td>Chief</td>
<td>Wednesday (2003-03-04)</td>
</tr>
</tbody>
</table>
**Event Extraction vs. Relation Extraction**

- **Relation Extraction**
  - Identify the relation between **two given entities**

  ```
  /business/company/founder

  Steve Jobs was the co-founder of Apple Inc.
  entity1
  entity2
  ```

- **Event Extraction**
  - Identify the relation between **an event and an entity**

  ```
  Person
  Time
  Organization
  Position

  Barry Diller on Wednesday quit as chief of Vivendi Universal Entertainment.
  Trigger Words
  Arguments Words
  ```
Type of Events

- Life: Be-born, Marry, Divorce, Die
- Movement: Transport
- Transaction: Transfer-ownership, Transfer-money
- Business: Start-org, Merge-org, Declare-bankruptcy, End-org
- Conflict
- Contact
- Personal
- Justice
Example

He has fired his air defense chief.

Position (End-Position)
Event Extraction via Dynamic Multi-Pooling Convolutional Neural Networks

Chen, ACL 2015
Experiments

Dataset: ACE 2005
Testing: 40 newswire articles
Development: 30 documents
Training: The rest (529) documents

<table>
<thead>
<tr>
<th>Methods</th>
<th>Trigger Identification (%)</th>
<th>Trigger Identification + Classification (%)</th>
<th>Argument Identification (%)</th>
<th>Argument Role (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P  R  F</td>
<td>P  R  F</td>
<td>P  R  F</td>
<td>P  R  F</td>
</tr>
<tr>
<td>Li’s baseline</td>
<td>76.2 60.5 67.4</td>
<td>74.5 59.1 65.9</td>
<td>74.1 37.4 49.7</td>
<td>65.4 33.1 43.9</td>
</tr>
<tr>
<td>Liao’s cross-event</td>
<td>N/A</td>
<td>68.7 68.9 68.8</td>
<td>50.9 49.7 50.3</td>
<td>45.1 44.1 44.6</td>
</tr>
<tr>
<td>Hong’s cross-entity</td>
<td>N/A</td>
<td>72.9 64.3 68.3</td>
<td>53.4 52.9 53.1</td>
<td>51.6 45.5 48.3</td>
</tr>
<tr>
<td>Li’s structure</td>
<td>76.9 65.0 70.4</td>
<td>73.7 62.3 67.5</td>
<td>69.8 47.9 56.8</td>
<td>64.7 44.4 52.7</td>
</tr>
<tr>
<td>DMCNN model</td>
<td>80.4 67.7 73.5</td>
<td>75.6 63.6 69.1</td>
<td>68.8 51.9 59.1</td>
<td>62.2 46.9 53.5</td>
</tr>
</tbody>
</table>
Joint Event Extraction via Recurrent Neural Networks
Brief Summary of IE

- Deep Learning
  - Sentence Representation (CNN/RNN)
  - Attention

- Data
  - Human Labeled
  - Distant Supervision
A representative IE domain

- Its conclusions can be applied to other domains

Natural Language Processing for Precision Medicine

Hoifung Poon, Chris Quirk, Kristina Toutanova, Scott Wen-tau Yih
"Biomedicine is an ocean that's one meter deep"
Medicine Today Is Imprecise

IMPRECISION MEDICINE
For every person they do help (blue), the ten highest-grossing drugs in the United States fail to improve the conditions of between 3 and 24 people (red).

1. ABILIFY (aripiprazole)
Schizophrenia

2. NEXIUM (esomeprazole)
Heartburn

3. HUMIRA (adalimumab)
Arthritis

4. CREATOR (rosuvastatin)
High cholesterol

5. CYMBALTA (duloxetine)
Depression

6. ADVAIR DISKUS (fluticasone propionate)
Asthma

7. ENBREL (etanercept)
Psoriasis

8. REMICADE (infliximab)
Crohn’s disease

9. COPAXONE (glatiramer acetate)
Multiple sclerosis

10. NEULASTA (pegfilgrastim)
Neutropenia

Top 20 drugs
80% non-responders

Wasted
1/3 health spending
$1 Trillion / year
Disruption: Big Data

Accenture study: 93% of US doctors using EMRs

2009 – 2013: 40% → 93%

The 1,000 Genome

The Revolution in DNA Sequencing and the New Era of Personalized Medicine

Kevin Davies

[Poon+ 2017]
Why We Haven’t Solved Precision Medicine?

Bottleneck #1: Knowledge

Bottleneck #2: Reasoning

AI is the key to overcome these bottlenecks
Key Scenario: Molecular Tumor Board

Problem: Hard to scale

U.S. 2016: 1.7 million new cases, 600K deaths
902 cancer hospitals

Knowledge bottleneck
E.g., given a tumor sequence, determine:
- What genes and mutations are important
- What drugs might be applicable
Can do manually but hard to scale

Reasoning bottleneck
E.g., personalize drug combinations
Can’t do manually, ever
PubMed

27 million abstracts
Two new abstracts every minute
Adds over one million every year

[Poon+ 2017]
Example: Personalize Drug Combos

Targeted drugs: 149

Pairs: 11,026

Tested: 102 (in two years)

Unknown: 10,924

Can we find good combos in months, not centuries?
Challenge: Cross-Sentence Relation Extraction

The deletion mutation on exon-19 of EGFR gene was present in 16 patients, while the L858E point mutation on exon-21 was noted in 10. All patients were treated with gefitinib and showed a partial response.

Gefitinib could be used to treat tumors w. EGFR mutation L858E. TREAT(Gefitinib, EGFR, L858E)
Generalize to N-ary Relations

The deletion mutation on exon-19 of EGFR gene was present in 16 patients, while the L858E point mutation on exon-21 was noted in 10. All patients were treated with gefitinib and showed a partial response.

Peng et al. “Cross-Sentence N-ary Relation Extraction with Graph LSTM”, TACL-17.
The p56Lck inhibitor Dasatinib was shown to enhance apoptosis induction in otherwise GC-resistant CLL cells.

This shows that Notch-mediated resistance of a mouse lymphoma cell line could be overcome by inhibiting p56Lck.
Asynchronous Update

All patients were treated with gefitinib and showed a partial response.
<table>
<thead>
<tr>
<th>Relations</th>
<th>Single-Sent.</th>
<th>Cross-Sent.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Candidates</td>
<td>169,168</td>
<td>332,969</td>
</tr>
<tr>
<td>$p \geq 0.5$</td>
<td>32,028</td>
<td>64,828</td>
</tr>
<tr>
<td>$p \geq 0.9$</td>
<td>17,349</td>
<td>32,775</td>
</tr>
<tr>
<td>GDKD</td>
<td>162</td>
<td></td>
</tr>
</tbody>
</table>

Orders of magnitude more knowledge by machine reading
So far: Relationships Directly Expressed in Text

Tumor suppressor P53 down-regulates the activity of BCL-2 proteins.

negative_regulation(P53,BCL-2)

Reasoning: combining several pieces of relevant information.
Genomics Knowledge Base (Network) [Poon+ 2017]

MAPK3 and MAPK1 are in the same family
MAPK1 up-regulates GRB2

Likely that MAPK3 up-regulates GRB2
Graphical models are expensive

Statistical relational learning [Getoor & Taskar, 2007]

- Modeling dependencies among the truth values of multiple possible relations

$F_1 : (x, parentOf, z) \land (y, parentOf, z) \Rightarrow (x, marriedTo, y)$

- Can be prohibitively expensive (e.g. marginal inference is exponential in the treewidth for Markov Random Fields)
Embeddings and random walks are more scalable

Knowledge base embedding
- Assumes truth values of facts are independent given latent features (embeddings) of entities and relations
- Can be very efficient (e.g. matrix multiplication for prediction)
- Has difficulty generalizing when graph has many small cliques

Path ranking methods (e.g., random walk) [e.g., Lao+ 2011]
- Assumes truth values of unknown facts are independent given observed facts
- Difficulty capturing dependencies through long relation paths
- Sparsity when number of relation types is large

Hybrid of path ranking and embedding methods
WideOpen: “Make Public Data Public”

NLP: Automate detection of overdue datasets
PubMed: Identify dataset mentions
Repo: Parse query output to determine if overdue

Enabled GEO to release 400 datasets in a week
Plan

- Information extraction
- **Semantic parsing**
- Semantic representation
Can this KG pattern be learned?

When reasoning is needed to understand text [Liang+ 2017]

Bart's father is Homer

Bart Simpson
1) Natural languages are programming languages to **control** human behavior

2) For machines to understand natural languages, they just need a **translation** model, which converts questions (statements) to **programs**

3) The programs find answers when "**executed**" against KB
Semantic Parsing: Language to Programs

Natural Language Question/Instruction → Program / Logical Form

Full supervision (hard to collect)

[Berant, et al 2013; Liang 2013]

Weak supervision (easy to collect)

[Liанг+ 2017]
Question Answering with Knowledge Base

Largest city in US?

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

NYC

1. Compositionality

2. Large Search Space

Freebase:
23K predicates, 82M entities, 417M triplets

Freebase, DBpedia, YAGO, NELL

[Image]
WebQuestionsSP Dataset

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

Grammatical error

- What do Michelle Obama do for a living?
- What character did Natalie Portman play in Star Wars?
- What currency do you use in Costa Rica?
- What did Obama study in school?
- What killed Sammy Davis Jr?

Multiple entities

writer, lawyer
Padme Amidala
Costa Rican colon
political science
throat cancer

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

[Liang+ 2017]
(Scalable) Neural Program Induction

- Impressive works to show NN can learn addition and sorting, but...
- The learned operations are not as scalable and precise.
- Why not use existing modules that are scalable, precise and interpretable?
Connectionism + Symbolism

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

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

[Source: Liang+ 2017]
Neural Symbolic Machines

Weak supervision
Manager

Neural
Programmer

Symbolic
Computer

Knowledge Base
Predefined Functions

Abstract
Scalable
Precise
Non-differentiable

HERE'S ANOTHER SHOVEL FULL OF ASSIGNMENTS.
Symbolic Machines in Brains

- 2014 Nobel Prize in Physiology or Medicine awarded for ‘inner GPS’ research
- Positions are represented as discrete numbers in animals' brains, which enable accurate and autonomous calculations

Mean grid spacing for all modules (M1–M4) in all animals (colour-coded)

[Stensola+ 2012]
Simple Seq2Seq model is not enough

1. Compositionality
   \[ v_2 \leftarrow \text{Argmax} \; v_1 \; \text{Population} \]
   \[ v_1 \leftarrow \text{Hop} \; v_0 \; \text{CityIn} \]
   \[ v_0 \leftarrow (\text{USA}) \; \text{CityIn} \]

2. Large Search Space
   23K predicates,
   82M entities,
   417M triplets

1. Key-Variable Memory
2. Code Assistance
3. Augmented REINFORCE

[Challenges Ahead]

[Liang+ 2017]
Key-Variable Memory for Compositionality

- A linearised bottom-up derivation of the recursive program.

[Liang+ 2017]
Key-Variable Memory: Save Intermediate Value

<table>
<thead>
<tr>
<th>Key (Embedding)</th>
<th>Variable (Symbol)</th>
<th>Value (Data in Computer)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$V_0$</td>
<td>R0</td>
<td>m.USA</td>
</tr>
<tr>
<td>$V_1$</td>
<td>R1</td>
<td>[m.SF, m.NYC, ...]</td>
</tr>
</tbody>
</table>

Expression is finished.

[Liag+ 2017]
Key-Variable Memory: Reuse Intermediate Value

![Diagram and table]

**Table:**

<table>
<thead>
<tr>
<th>Key (Embedding)</th>
<th>Variable (Symbol)</th>
<th>Value (Data in Computer)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$V_0$</td>
<td>R0</td>
<td>m.USA</td>
</tr>
<tr>
<td>$V_1$</td>
<td>R1</td>
<td>[m.SF, m.NYC, ...]</td>
</tr>
</tbody>
</table>

[Softmax]

[Neural]

[Symbolic]

[CityIn]
Pen and paper → IDE

Code Assistance: Prune Search Space

[Chen+ 2017]
Code Assistance: Syntactic Constraint

Decoder Vocab

- $V_0$: R0
- $V_1$: R1
- $E_0$: Hop
- $E_1$: Argmax
- $P_0$: CityIn
- $P_1$: BornIn

Variables: <10
Functions: <10
Predicates: 23K

[Liăng+ 2017]
Code Assistance: Syntactic Constraint

Decoder Vocab

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

Variables: <10
Functions: <10
Predicates: 23K

[Liag+ 2017]
Code Assistance: Semantic Constraint

Decoder Vocab

<table>
<thead>
<tr>
<th>V₀</th>
<th>R0</th>
</tr>
</thead>
<tbody>
<tr>
<td>V₁</td>
<td>R₁</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>E₀</td>
<td>Hop</td>
</tr>
<tr>
<td>E₁</td>
<td>Argmax</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>P₀</td>
<td>CityIn</td>
</tr>
<tr>
<td>P₁</td>
<td>BornIn</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Variables: <10

Functions: <10

Predicates: 23K

[Liung+ 2017]
Given definition of \( \text{Hop} \), need to output a predicate that is connected to \( R_2 \) (\( m.\text{USA} \)).

Decoder Vocab

- Variables: <10
- Functions: <10
- Predicates: 23K
- Valid Predicates: <100
REINFORCE Training

1. High variance
   Requires a lot of (expensive) samples

2. Cold start problem
   Without supervised pretraining, the gradients at the beginning

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

[Liang+ 2017]
Iterative Maximum Likelihood Training (Hard EM) [Liang+ 2017]

1. Spurious program mistake: PlaceOfBirth for PlaceOfDeath.
2. Lack of negative examples mistake: SiblingsOf for ParentsOf.

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

1. Reduce variance at the cost of bias
2. Mix in approximate gold programs to bootstrap and stabilize training

Beam search

Top \(k\) in beam

Policy gradient update

Updated Model

Approximate Gold Programs

(\(1 - \alpha\))

\(\alpha\)
Distributed Architecture

- 200 actors, 1 learner, 50 Knowledge Graph servers

[Qiang+ 2017]
Generated Programs

- **Question**: “what college did russell wilson go to?”
- **Generated program**:
  
  (hop v1 /people/person/education)
  (hop v2 /education/education/institution)
  (filter v3 v0 /common/topic/notable_types )
  <EOP>

  In which

  v0 = “College/University” (m.01y2hnl)
  v1 = “Russell Wilson” (m.05c10yf)

- **Distribution of the length of generated programs**

```
<table>
<thead>
<tr>
<th>#Expressions</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Percentage</strong></td>
<td>0.4%</td>
<td>62.9%</td>
<td>29.8%</td>
<td>6.9%</td>
</tr>
<tr>
<td><strong>F1</strong></td>
<td>0.0</td>
<td>73.5</td>
<td>59.9</td>
<td>70.3</td>
</tr>
</tbody>
</table>
```
New State-of-the-Art on WebQuestionsSP

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

<table>
<thead>
<tr>
<th>Model</th>
<th>Avg. Prec. @1</th>
<th>Avg. Rec. @1</th>
<th>Avg. F1 @1</th>
<th>Acc. @1</th>
</tr>
</thead>
<tbody>
<tr>
<td>STAGG</td>
<td>67.3</td>
<td>73.1</td>
<td>66.8</td>
<td>58.8</td>
</tr>
<tr>
<td>NSM – our model</td>
<td>70.8</td>
<td>76.0</td>
<td>69.0</td>
<td>59.5</td>
</tr>
<tr>
<td>STAGG (full supervision)</td>
<td>70.9</td>
<td>80.3</td>
<td>71.7</td>
<td>63.9</td>
</tr>
</tbody>
</table>
Augmented REINFORCE

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

<table>
<thead>
<tr>
<th>Settings</th>
<th>Train Avg. F1@1</th>
<th>Valid Avg. F1@1</th>
</tr>
</thead>
<tbody>
<tr>
<td>iterative ML only</td>
<td>68.6</td>
<td>60.1</td>
</tr>
<tr>
<td>REINFORCE only</td>
<td>55.1</td>
<td>47.8</td>
</tr>
<tr>
<td>Augmented REINFORCE</td>
<td>83.0</td>
<td>67.2</td>
</tr>
</tbody>
</table>

[Liang+ 2017]
From open IE to matching problems

[Berant & Liang 2014]

![Diagram]

- From open IE to matching problems
- "what is Italy money"
- "latest meeting"
- "Italy currency"
- "meeting with the largest end time"
- Italy.Currency
- argmax(type.meeting, end_time)
From open IE to matching problems

The beauty of the proposed approach

"what is Italy money"  "latest meeting"
"Italy currency"  "meeting with the largest end time"
large number of training examples

Itally.Currency
argmax(type.meeting, end_time)

[Berant & Liang 2014]
The Web as a KB

Depth (compositionality)

- Early Systems
- Scale to KBs

Semantic Parsing on Semi-Structured Data

- Web Search

[Berant+ 2013]
[Liang 2013]

[Liang+ 2011]

[Pasupat & Liang 2015]
The Web as a KB

**Early systems:** Parse very compositional questions into database queries

How many rivers are in the state with the largest population?

\[
\text{answer}(A, \text{count}(B, \text{river}(B), \text{loc}(B, C), \text{largest}(D, \text{state}(C), \text{population}(C, D))), A))
\]

**Compositionality:** High

**Knowledge source:** Database

- few entities / relations
- fixed schema

The Web as a KB

Scaling to large knowledge bases (KBs): Answer open-domain questions using curated KBs

In which comic book issue did Kitty Pryde first appear?

Compositionality: Lower

Knowledge source: Large KBs
  - lots of entities / relations
  - fixed schema

[Cai + Yates, 2013 / Berant et al., 2013 + 2014 / Fader et al., 2014 / Reddy et al., 2014 / ...]
The Web as a KB

QA on semi-structured data

**Input**: utterance $x$ and HTML table $t$

**Output**: answer $y$

**Training data**: list of $(x, t, y)$ — no logical form

**WikiTableQuestions dataset**:

- Tables $t$ are from Wikipedia
- Questions $x$ and answers $y$ are from Mechanical Turk
- Prompts are given to encourage compositionality
  
  e.g. Prompt: The question must contains "last" (or a synonym)

In what city did Piotr's last 1st place finish occur?
The Web as a KB

Greece held its last Summer Olympics in which year?

(1) Generation

(2) Ranking

R[λx[Year, Date, x]]. argmax(..., Index)

(3) Execution

Tables are represented as graphs

2004
Learning a Neural Semantic Parser from User Feedback

- neural sequence models to map utterances directly to SQL, bypassing any intermediate meaning representations

- These models are immediately deployed online to solicit feedback from real users to flag incorrect queries.

```sql
SELECT paper.paperId, paper.year
FROM paper, writes, author
WHERE paper.paperId = writes.paperId
  AND writes.authorId = author.authorId
  AND author.authorName = "michael i. jordan"
  AND paper.year =
    (SELECT max(paper.year)
     FROM paper, writes, author
     WHERE paper.paperId = writes.paperId
     AND writes.authorId = author.authorId
     AND author.authorName = "michael i. jordan");

I'd like to book a flight from San Diego to Toronto

SELECT DISTINCT f1.flight_id
FROM flight f1, airport_service a1, city c1,
     airport_service a2, city c2
WHERE f1.from_airport = a1.airport_code
  AND a1.city_code = c1.city_code
  AND c1.city_name = 'san diego'
  AND f1.to_airport = a2.airport_code
  AND a2.city_code = c2.city_code
  AND c2.city_name = 'toronto';
```
Learning Structured Natural Language Representations for Semantic Parsing

**Sentence**: which states do not border texas

**Non-terminal symbols in buffer**: which, states, do, not, border

**Terminal symbols in buffer**: texas

<table>
<thead>
<tr>
<th>Stack</th>
<th>Action</th>
<th>NT choice</th>
<th>TER choice</th>
</tr>
</thead>
<tbody>
<tr>
<td>answer ()</td>
<td>NT</td>
<td>answer</td>
<td></td>
</tr>
<tr>
<td>answer () exclude ()</td>
<td>NT</td>
<td>exclude</td>
<td></td>
</tr>
<tr>
<td>answer () exclude () states ()</td>
<td>TER</td>
<td>states</td>
<td>all</td>
</tr>
<tr>
<td>answer () exclude () states ( all )</td>
<td>RED</td>
<td>border</td>
<td></td>
</tr>
<tr>
<td>answer () exclude () states ( all )</td>
<td>NT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>answer () exclude () states ( all ), border ( )</td>
<td>TER</td>
<td></td>
<td>texas</td>
</tr>
<tr>
<td>answer () exclude () states ( all ), border ( texas )</td>
<td>RED</td>
<td></td>
<td></td>
</tr>
<tr>
<td>answer () exclude () states ( all ), border ( texas )</td>
<td>RED</td>
<td></td>
<td></td>
</tr>
<tr>
<td>answer () exclude () states ( all ), border ( texas )</td>
<td>RED</td>
<td></td>
<td></td>
</tr>
<tr>
<td>answer () exclude () states ( all ), border ( texas )</td>
<td>RED</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Actions taken by the transition system for generating the ungrounded meaning representation of the example utterance. Symbols in red indicate domain-general predicates.
Plan

- Information extraction
- Semantic parsing
- Semantic representation
The AI Elephant

- Each subfield of AI holds certain truth, but not all of it.
Putting things together

function approximation

correct training

structural bias

AlphaGo
Language & reasoning

- Language was primarily invented for reasoning
- Communication comes later
Lapata’s scream

I DON'T HAVE $$ FOR GPUS!

WHAT NOW? IS NLP DEAD?

AND LANGUAGE?

I REALLY LIKE MY FEATURES!
Noah’s Bias

(structural) bias

data

• Parsing sentences into predicate-argument structures
  • Fillmore frames
  • Semantic dependency graphs
• Language models that dynamically track entities
Why Relation Extraction Worked

- In very restricted domains
  - Closed domain queries
  - Semantic parser
  - KB relations
  - Text or html patterns
  - Web docs
Why Open Domain Relation Extraction Is Hard

- Open domain schemas are not compact enough

Open domain queries: q q q q q q q q q
Lexicon and matching
Open domain relations: R R R R R R R R
Reverb extraction rules
Web docs: d d d d d d d d d

how to define a schema for OIE?
Figure 2: Preprocessing methods applied to sentence (top row) - AMR graph (left column) pairs. Sentence-graph pairs after (a) graph simplification, (b) named entity anonymization, (c) named entity clustering, and (d) insertion of scope markers.
Question answering as a simple test bed

- A good semantic representation should support reasoning (computation)

Text → Knowledge Store → Answer → Expected Answer → Reward

Question → Program → Execute (i.e. no learning) → Generate (i.e. learning)
Thanks
Theory of cognitive development

- Piaget identified several important milestones in the mental development of children

"It is with children that we have the best chance of studying the development of logical knowledge, mathematical knowledge, physical knowledge, and so forth." -- Jean Piaget
Combine KB completion models with relation extractions [Dong+ 2014]