Machine Reasoning in NLP

Duyu Tang
Microsoft Research Asia
Past Tutorials on Machine Reasoning

- NLPCC–2020 Tutorial (3 hours)

Machine Reasoning in NLP
Duyu Tang
Microsoft Research Asia
NLPCC-2020 Tutorial

- EMNLP–2020 Tutorial (3 hours)

Machine Reasoning: Technology and Dilemma
Nan DUAN, Duyu TANG, Ming ZHOU
Microsoft Research Asia
{nanduan, dutang, mingzhou}@microsoft.com
EMNLP-2020 Tutorial

Both tutorials are available at https://tangduyu.github.io/
Microsoft AI Breakthroughs

- Gradually approaching human parity

<table>
<thead>
<tr>
<th>Year</th>
<th>Category</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>2016</td>
<td>Object recognition</td>
<td>96.0% RESNET 152 layers ImageNet Contest</td>
</tr>
<tr>
<td>2017</td>
<td>Speech recognition</td>
<td>94.9% Switchboard speech recognition test</td>
</tr>
<tr>
<td>January 2018</td>
<td>Machine reading comprehension</td>
<td>88.5% SQuAD reading comprehension test</td>
</tr>
<tr>
<td>March 2018</td>
<td>Machine translation</td>
<td>69.9% MT research system</td>
</tr>
<tr>
<td>March 2019</td>
<td>Conversational AI</td>
<td>89.9% CoQA conversational AI test</td>
</tr>
</tbody>
</table>

Image Courtesy: Harry Shum, 2020
Limitations:
1. Lack of transparency of the decision-making process
2. Highly rely on annotated data, ignore human/expert knowledge
Input → Model → Output

Explanation

Knowledge
Example #1: Simple Question Answering

Question: Which city hosted the Summer Olympics in 2008?

SQL: 

```
SEET Chate WEB Year = 208
```

Semantic Parsing

Execution

Answer: Beijing

<table>
<thead>
<tr>
<th>Year</th>
<th>City</th>
<th>Country</th>
<th>Nations</th>
</tr>
</thead>
<tbody>
<tr>
<td>1896</td>
<td>Athens</td>
<td>Greece</td>
<td>14</td>
</tr>
<tr>
<td>1900</td>
<td>Paris</td>
<td>France</td>
<td>24</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>2004</td>
<td>Athens</td>
<td>Greece</td>
<td>201</td>
</tr>
<tr>
<td>2008</td>
<td>Beijing</td>
<td>China</td>
<td>204</td>
</tr>
<tr>
<td>2012</td>
<td>London</td>
<td>UK</td>
<td>204</td>
</tr>
</tbody>
</table>

Required Knowledge: Symbolic
Example #2: Multi–Turn Question Answering

Q1: Which city hosted the Summer Olympics in 2008?
   SELECT Character WHERE Year = 2008
   A1: Beijing

Q2: How many nations participated that year?
   SELECT Nations WHERE Year = 2008
   A2: 204

Q3: How about 2004?
   SELECT Nations WHERE Year = 2004
   A3: 201

<table>
<thead>
<tr>
<th>Year</th>
<th>City</th>
<th>Country</th>
<th>Nations</th>
</tr>
</thead>
<tbody>
<tr>
<td>1896</td>
<td>Athens</td>
<td>Greece</td>
<td>14</td>
</tr>
<tr>
<td>1900</td>
<td>Pairs</td>
<td>France</td>
<td>24</td>
</tr>
<tr>
<td>2004</td>
<td>Athens</td>
<td>Greece</td>
<td>201</td>
</tr>
<tr>
<td>2008</td>
<td>Beijing</td>
<td>China</td>
<td>204</td>
</tr>
<tr>
<td>2012</td>
<td>London</td>
<td>UK</td>
<td>204</td>
</tr>
</tbody>
</table>

Required Knowledge: Symbolic + Conversation
Example #3: Medical Diagnosis

Multi-Turn Question Answering & Clarification

I am feeling cold today. I had a cough and a shortness of breath.

38.4º

PI

Mood: passive

Reasoning

Knowledge

• Facts
• Rules

Conclusions

You are most at risk. Have a PCR diagnosis first. Stay under quarantine.

You have symptoms of COVID-19, including cough, fever, and shortness of breath. You also have close contact with an infected person, which makes you most at risk.

Required Knowledge: Symbolic + Conversation + Domain Knowledge
### Example #3: Fact

This table illustrates the Situation by Country, Territory & Area as reported on the WHO Coronavirus Disease (COVID-19) Dashboard. The data is updated as of a specific date, shown as 30 May 2021, 13:47 (CEST).

#### Situation by Country, Territory & Area

<table>
<thead>
<tr>
<th>Name</th>
<th>Cases - cumulative total</th>
<th>Cases - newly reported in last 24 hours</th>
<th>Deaths - cumulative total</th>
<th>Deaths - newly reported in last 24 hours</th>
<th>Transmission Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global</td>
<td>35,109,317</td>
<td>294,763</td>
<td>1,035,341</td>
<td>4,526</td>
<td></td>
</tr>
<tr>
<td>United States</td>
<td>7,305,270</td>
<td>49,036</td>
<td>208,064</td>
<td>698</td>
<td>Community transmission</td>
</tr>
<tr>
<td>India</td>
<td>6,623,615</td>
<td>74,442</td>
<td>102,685</td>
<td>903</td>
<td>Clusters of cases</td>
</tr>
<tr>
<td>Brazil</td>
<td>4,906,833</td>
<td>26,310</td>
<td>145,987</td>
<td>599</td>
<td>Community transmission</td>
</tr>
<tr>
<td>Russian Fed…</td>
<td>1,225,889</td>
<td>10,888</td>
<td>21,475</td>
<td>117</td>
<td>Clusters of cases</td>
</tr>
<tr>
<td>Colombia</td>
<td>848,147</td>
<td>6,616</td>
<td>26,556</td>
<td>159</td>
<td>Community transmission</td>
</tr>
<tr>
<td>Peru</td>
<td>824,985</td>
<td>3,421</td>
<td>32,665</td>
<td>56</td>
<td>Community transmission</td>
</tr>
<tr>
<td>Argentina</td>
<td>790,818</td>
<td>11,129</td>
<td>20,795</td>
<td>196</td>
<td>Community transmission</td>
</tr>
<tr>
<td>Spain</td>
<td>789,932</td>
<td>0</td>
<td>32,086</td>
<td>0</td>
<td>Clusters of cases</td>
</tr>
<tr>
<td>Mexico</td>
<td>757,953</td>
<td>4,863</td>
<td>78,880</td>
<td>388</td>
<td>Community transmission</td>
</tr>
<tr>
<td>South Africa</td>
<td>681,289</td>
<td>1,573</td>
<td>16,976</td>
<td>38</td>
<td>Community transmission</td>
</tr>
</tbody>
</table>

[https://covid19.who.int/table](https://covid19.who.int/table)
Example #3: Rule

**SYMPTOM**
- IF <a person is infected> THEN <he/she may have fever>
- IF <a person is infected> THEN <he/she may have cough and shortness of breath>

**GROUPS AT RISK**
- IF <a person has close contact with animals> THEN <the person is at risk>
- IF <a person is a live animal market worker> THEN <he/she has close contact with animals>
- IF <a person has close contact with an infected person> THEN <the person is most at risk>
- IF <a person is a healthcare worker or a family member of infected person> THEN <the person has close contact with infected person>

**DIAGNOSIS & TREATMENT**
- IF <disease == COVID-19> THEN <diagnosis(disease)=PCR (Polymerase Chain)>
- IF <disease == COVID-19> THEN <treatment(disease)=none AND vaccine(disease)=none> (by Jan 31, 2020)
I am feeling cold today. I had a cough and a shortness of breath.

38.4º

Mood: passive

One of my family member is affected.

Have you contact with an infected person?

You are most at risk. Have a PCR diagnosis first. Stay under quarantine.

You have symptoms of COVID-19, including cough, fever, and shortness of breath. You also have close contact with an infected person, which makes you most at risk.
Features of Machine Reasoning

Input
- Image
- Physical signals
- Speech
- Text

Processor
- Knowledge
- Reasoning

Conclusion
- Decision
- Explanation
+ Propositional/First-Order Logic
+ Neuro-symbolic

Pre-trained models (e.g. ELMo, BERT, GPT)
Evidence (e.g. retrieved docs from Wikipedia/web, retrieved facts from Wikidata/ConceptNet)

First-Order Logic
Symbolic Operations

Input → Model → Output

\[(\alpha \land \beta) \equiv (\beta \land \alpha) \quad \text{commutativity of } \land\]
\[(\alpha \lor \beta) \equiv (\beta \lor \alpha) \quad \text{commutativity of } \lor\]
\[(\alpha \land (\beta \land \gamma)) \equiv (\alpha \land \beta) \land \gamma) \quad \text{associativity of } \land\]
\[(\alpha \lor (\beta \lor \gamma)) \equiv (\alpha \lor \beta) \lor \gamma) \quad \text{associativity of } \lor\]
\[\neg(\neg\alpha) \equiv \alpha \quad \text{double-negation elimination}\]
\[(\alpha \Rightarrow \beta) \equiv (\neg\beta \Rightarrow \neg\alpha) \quad \text{contraposition}\]
\[(\alpha \Rightarrow \beta) \equiv (\neg\alpha \lor \beta) \quad \text{implication elimination}\]

SEET  CON  CE  Tam  WEB
Obj  = "Yok"

lambda p(x) . p(x) . p(x) . of_bit h(
Dadd  Tam  x)
Agenda

Opening
Logic-based Models in NLP
Neural-Symbolic Models in NLP
Evidence-based Models in NLP
Summary
Logic–based Models in NLP
Outline

- Propositional Logic and First-Order Logic
- Inference/Theorem Proving: Forward and Backward Chaining
- Application in NLP
Propositional Logic

• **Logical constants**: true, false
• **Propositional symbols**: P, Q, S, ... (atomic sentences)
• Wrapping **parentheses**: ( ... )
• Sentences are combined by **connectives**:
  - ∧ ...and [conjunction]
  - ∨ ...or [disjunction]
  - → ...implies [implication / conditional]
  - ↔ ..is equivalent [biconditional]
  - ¬ ...not [negation]
• **Literal**: atomic sentence or negated atomic sentence

Attribution: Marie desJardins, Fall 2005
Propositional Logic Examples

- P means “It is hot”
- Q means “It is humid”
- R means “It is raining”

- \((P \land Q) \rightarrow R\)
  “If it is hot and humid, then it is raining”
- Q \rightarrow P
  “If it is humid, then it is hot”
- Q
  “It is humid.”

Attribution: Marie desJardins, Fall 2005
Propositional Logic Syntax

• Given: a set of proposition symbols \( \{X_1, X_2, \ldots, X_n\} \)
  • (we often add True and False for convenience)
• \( X_i \) is a sentence
• If \( \alpha \) is a sentence then \( \neg \alpha \) is a sentence
• If \( \alpha \) and \( \beta \) are sentences then \( \alpha \land \beta \) is a sentence
• If \( \alpha \) and \( \beta \) are sentences then \( \alpha \lor \beta \) is a sentence
• If \( \alpha \) and \( \beta \) are sentences then \( \alpha \Rightarrow \beta \) is a sentence
• If \( \alpha \) and \( \beta \) are sentences then \( \alpha \Leftrightarrow \beta \) is a sentence
• And p.s. there are no other sentences!

Attribution: Stuart Russell, 2019, ai.berkeley.edu
Logical Equivalence

\[
\begin{align*}
(\alpha \land \beta) & \equiv (\beta \land \alpha) \quad \text{commutativity of } \land \\
(\alpha \lor \beta) & \equiv (\beta \lor \alpha) \quad \text{commutativity of } \lor \\
((\alpha \land \beta) \land \gamma) & \equiv (\alpha \land (\beta \land \gamma)) \quad \text{associativity of } \land \\
((\alpha \lor \beta) \lor \gamma) & \equiv (\alpha \lor (\beta \lor \gamma)) \quad \text{associativity of } \lor \\
\neg(\neg\alpha) & \equiv \alpha \quad \text{double-negation elimination} \\
(\alpha \Rightarrow \beta) & \equiv (\neg\beta \Rightarrow \neg\alpha) \quad \text{contraposition} \\
(\alpha \Rightarrow \beta) & \equiv (\neg\alpha \lor \beta) \quad \text{implication elimination} \\
(\alpha \iff \beta) & \equiv ((\alpha \Rightarrow \beta) \land (\beta \Rightarrow \alpha)) \quad \text{biconditional elimination} \\
\neg(\alpha \land \beta) & \equiv (\neg\alpha \lor \neg\beta) \quad \text{De Morgan} \\
\neg(\alpha \lor \beta) & \equiv (\neg\alpha \land \neg\beta) \quad \text{De Morgan} \\
(\alpha \land (\beta \lor \gamma)) & \equiv ((\alpha \land \beta) \lor (\alpha \land \gamma)) \quad \text{distributivity of } \land \text{ over } \lor \\
(\alpha \lor (\beta \land \gamma)) & \equiv ((\alpha \lor \beta) \land (\alpha \lor \gamma)) \quad \text{distributivity of } \lor \text{ over } \land
\end{align*}
\]
Problems with Propositional Logic

- Hard to identify “individuals” (e.g., Mary, 3)
- Can’t directly talk about properties of individuals or relations between individuals (e.g., “Bill is tall”)
- Generalizations, patterns, regularities can’t easily be represented (e.g., “all triangles have 3 sides”)
- Lack of variables prevents stating more general rules
  - We need a set of similar rules for each cell
- First-Order Logic is expressive enough to concisely represent this kind of information
  - FOL adds relations, variables, and quantifiers, e.g.,
    - “Every elephant is gray”: ∀ x (elephant(x) → gray(x))
    - “There is a white alligator”: ∃ x (alligator(X) ^ white(X))
First–Order Logic

• First–order logic models the world in terms of
  • **Objects**, which are things with individual identities
  • **Properties** of objects that distinguish them from other objects
  • **Relations** that hold among sets of objects
  • **Functions**, which are a subset of relations where there is only one “value” for any given “input”

• Examples:
  • Objects: Students, lectures, companies, cars ...
  • Relations: Brother–of, bigger–than, outside, part–of, has–color, occurs–after, owns, visits, precedes, ...
  • Properties: blue, oval, even, large, ...
  • Functions: father–of, best–friend, second–half, one–more–than ...
First–Order Logic Examples

- Universal quantification
  - $(\forall x)P(x)$ means that $P$ holds for all values of $x$ in the domain associated with that variable
  - E.g., $(\forall x)\text{ dolphin}(x) \rightarrow \text{mammal}(x)$

- Existential quantification
  - $(\exists x)P(x)$ means that $P$ holds for some value of $x$ in the domain associated with that variable
  - E.g., $(\exists x)\text{mammal}(x) \land \text{lays–eggs}(x)$
  - Permits one to make a statement about some object without naming it

<table>
<thead>
<tr>
<th>Language</th>
<th>Propositional logic</th>
<th>First–order logic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Syntax</td>
<td>The world contains facts</td>
<td>The world contains objects, relations, and functions</td>
</tr>
<tr>
<td>Semantics</td>
<td>$\alpha \land \beta$ is true in a world iff $\alpha$ is true and $\beta$ is true (etc.)</td>
<td>$\phi(\sigma)$ is true in a world if $\sigma=\sigma_j$ and $\phi$ holds for $\sigma_j$; etc.</td>
</tr>
</tbody>
</table>
Outline

• Propositional Logic and First-Order Logic

• Inference/Theorem Proving: Forward and Backward Chains

• Application in NLP
Forward Chaining

- Start with given proposition symbols (atomic sentence)
  - e.g., A and B

- Iteratively try to infer truth of additional proposition symbols
  - e.g., $A \land B \Rightarrow C$, therefore we establish C is true

- Continue until
  - no more inference can be carried out, or
  - goal is reached

Attribution: Philipp Koehn, 2020
Forward Chaining Example: Proving Q

<table>
<thead>
<tr>
<th>Clauses</th>
<th>Count</th>
<th>Inferrred</th>
</tr>
</thead>
<tbody>
<tr>
<td>( P \implies Q )</td>
<td>1 N 0</td>
<td>A false true</td>
</tr>
<tr>
<td>( L \land M \implies P )</td>
<td>2 N 0</td>
<td>B false true</td>
</tr>
<tr>
<td>( B \land L \implies M )</td>
<td>2 N 0</td>
<td>M false true</td>
</tr>
<tr>
<td>( A \land P \implies L )</td>
<td>2 N 0</td>
<td>P false true</td>
</tr>
<tr>
<td>( A \land B \implies L )</td>
<td>0</td>
<td>Q false true</td>
</tr>
</tbody>
</table>

\( A \land B \implies L \) | 0 |

**Agenda**

- A
- B
Forward Chaining Algorithm

```plaintext
function PL-FC-ENTAILS?(KB, q) returns true or false

count ← a table, where count[c] is the number of symbols in c’s premise

inferred ← a table, where inferred[s] is initially false for all s

agenda ← a queue of symbols, initially symbols known to be true in KB

while agenda is not empty do
    p ← Pop(agenda)
    if p = q then return true
    if inferred[p] = false then
        inferred[p] ← true
        for each clause c in KB where p is in c.premise do
            decrement count[c]
            if count[c] = 0 then add c.conclusion to agenda

return false
```

Attribution: Sergey Levine and Stuart Russell, 2019, ai.berkeley.edu
Backward Chaining

- Idea: work backwards from the query $Q$:
  - to prove $Q$ by BC,
    - check if $Q$ is known already, or
    - prove by BC all premises of some rule concluding $q$

- Avoid loops: check if new subgoal is already on the goal stack

- Avoid repeated work: check if new subgoal
  - 1. has already been proved true, or
  - 2. has already failed

Attribution: Philipp Koehn, 2020
Outline

- Propositional Logic and First-Order Logic
- Inference/Theorem Proving: Forward and Backward Chaining
- Application in Knowledge Base Completion
  - Neural Backward Chaining
  - Logic as constraints
Neural Backward Chaining

Example Knowledge Base:
1. fatherOf(ABE, HOMER).
2. parentOf(HOMER, BART).
3. grandfatherOf(X, Y) :-
   fatherOf(X, Z),
   parentOf(Z, Y).

Logic as Constraints

\[
[F] = \begin{cases} 
\sigma(v_s \cdot v_{ij}) & \text{if } F = r_s(e_i, e_j), \text{ i.e., facts} \\
1 - [A] & \text{if } F = \neg A \\
[A] \star [B] & \text{if } F = A \land B 
\end{cases}
\]

\[r_1(e_i, e_j) \land \neg r_2(e_i, e_j) \Rightarrow r_3(e_i, e_j)\]
Performance V.S. Interpretability

• Neural Backward Chaining
  • Good interpretability, limited scope of application (e.g., completion on structured KB)

• Regularizing Neural Models
  • Good performance with neural models as backbone, limited interpretability
Neural–Symbolic Models in NLP
First-Order Logic: Every prime greater than two is odd.

\[ \forall x. \text{prime}(x) \land \text{more}(x, 2) \rightarrow \text{odd}(x) \]

Lambda Calculus: How many primes are less than 10?

\[ \text{count}(\lambda x. \text{prime}(x) \land \text{less}(x, 10)) \]

Lambda DCS: How many primes are less than 10?

\[ \text{count}(\text{prime} \cap (\text{less.10})) \]

- First-order logic fails to construct a set and manipulating it.
- The \( \lambda \) operator can be thought of as constructing a set of all \( x \) that satisfy the condition; in symbols, \( [\lambda x. f(x)]_c = \{ x : [f(x)]_c = \text{true} \} \).

Semantic Parsing

map an utterance $x$ in a context $c$ to an action $y$

$x$: *What is the largest prime less than 10?*

$c$: primes: $\{2, 3, 5, 7, 11, \ldots\}$

$y$: 7
A Derivation

What is the largest prime less than 10 ?

(R1)  \[ \text{NP}[\text{primes}] \]

(R2)  \[ \text{NP}[10] \]

(R3)  \[ QP[(-\infty, 10)] \]

(R4)  \[ \text{NP}[\text{primes} \cap (-\infty, 10)] \]

(R5)  \[ \text{NP}[\text{max}(\text{primes} \cap (-\infty, 10))] \]

(R7)  \[ \text{ROOT}[\text{max}(\text{primes} \cap (-\infty, 10))] \]

(R1)  \[ \text{prime} \quad \Rightarrow \quad \text{NP}[\text{primes}] \]

(R2)  \[ 10 \quad \Rightarrow \quad \text{NP}[10] \]

(R3)  \[ \text{less than NP}[z] \quad \Rightarrow \quad QP[(-\infty, z)] \]

(R4)  \[ \text{NP}[z_1] \quad QP[z_2] \quad \Rightarrow \quad \text{NP}[z_1 \cap z_2] \]

(R5)  \[ \text{largest NP}[z] \quad \Rightarrow \quad \text{NP}[\text{max}(z)] \]

(R6)  \[ \text{largest NP}[z] \quad \Rightarrow \quad \text{NP}[\text{min}(z)] \]

(R7)  \[ \text{What is the NP}[z] ? \quad \Rightarrow \quad \text{ROOT}[z] \]
Semantic Parsing

- Map **Natural Language** into **machine executable logical forms**

**Question**: How many CFL teams are from York College?

**Answer**: 2

<table>
<thead>
<tr>
<th>CFL Team</th>
<th>College</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hamilton Tiger-Cats</td>
<td>Wilfrid Laurier</td>
</tr>
<tr>
<td>Calgary Stampeders</td>
<td>York</td>
</tr>
<tr>
<td>Toronto Argonauts</td>
<td>York</td>
</tr>
</tbody>
</table>

**Question**: Where was Donald Trump given birth?

**Answer**: Queens, New York City
Outline

Part 1: Semantic Parsing

Part 2: Module Network

We are Here
LSNLIS (Lunar Sciences Natural Language Information System)

A question–answering system to enable a lunar geologist to conveniently access, compare, and evaluate the chemical analysis data on lunar rock and soil composition that is accumulating as a result of the Apollo moon missions.

Two DB files. One is a 13,000 line table of chemical and age analysis of the Apollo 11 samples extracted from the reports of a the First Annual Lunar Science Conference, and the second is a keyphrase index to those reports.

A dictionary about 3500 words.
Floating Parser

- Build formulas bottom-up according to a set of deduction rules
- Allow formulas to be created from nothing ("floating")

Greece held its last Summer Olympics in which year?


R[\lambda x[Year.Date.x]].argmax(Country.Greece, Index)

Greece held its last Summer Olympics in which year?

Greece
Country
Index

Ø \rightarrow Relation

Entities are anchored to token spans
Relations and Operations are not
Seq2Seq and Seq2Tree

Seq2Seq with 2-layer RNN

Seq2Tree with a hierarchical tree decode

Li Dong and Mirella Lapata. "Language to Logical Form with Neural Attention." ACL-2016.
Seq-to-Seq with Pointer Network

Decoder

Encoder

Question

Column Names

SQL

A distribution over the words from **question**, **column names** and limited vocabulary of the **SQL language**
Seq-to-Seq with Structural Decoding

Decoder

Encoder

Question

Pick # | CFL Team                    | Player       | Position | College   |
-------|-----------------------------|--------------|----------|-----------|
       | Hamilton Tiger-Cats         | Connor Healy | DB       | Wilfrid Laurier |
       | Calgary Stampeders         | Anthony Forgone | OL       | York     |
       | Toronto Argonauts          | Frank Hoffman | DL       | York     |
Seq-to-Seq with Typed Constrained Decoding

Action and Module

- Which super heroes came from Earth and first appeared after 2009?

```
SELECT Character
WHERE {Home World = Earth} \∧
{First Appeared > 2009}
```

Q = “Which super heroes came from Earth?”, A* = \{Dragonwing, Harmonia\}

Action and Module

- The goodness of a state: $V(s_t) = V(s_{t-1}) + \pi(s_{t-1}, a_t), V(s_0) = 0$
- Value of $\pi(s, a)$ is determined by a neural-network model
- Actions of the same type (e.g., select-column) share the same neural-network module

$\Rightarrow$ Which super heroes came from Earth?

Coarse-to-Fine Decoding

KBQA with Semantic Parsing (single-turn)

Where was the president of the United States born?

A1: \( S \rightarrow s \)
A4: \( s \rightarrow find(s \text{, } r1) \)
A4: \( s \rightarrow find(s \text{, } r2) \)
A15: \( s \rightarrow \{e\} \)
A16: \( e \rightarrow United \text{ States} \)
A17: \( r2 \rightarrow isPresidentOf \)
A17: \( r1 \rightarrow placeOfBirth \)

United States

KBQA with Semantic Parsing (multi-turn)

Where was president of the United States born?
Previous Question

New York City
Previous Answer

Where did he graduate from?
Current Question

Conversational Question Answering

Who is the wife of George Bush?

When you say George Bush, which one do you mean, George W. Bush or George H.W. Bush?

I mean the first, George W. Bush

Laura Bush

Part 1: Semantic Parsing

Part 2: Module Network
Reusable neural modules with different architectures

Attenton

Re-Attention

Combination

Classification

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein.
Performance V.S. Interpretability

- **Semantic Parser**
  - Good interpretability, good performance on limited applications
  - The extension of grammar to open domain is challenging

- **Module Network**
  - Good performance with neural models as backbone, limited interpretability
  - Moderate interpretability, compared with semantic parser.
  - The definition of grammar is typically task-specific, and manually designed by experts
Evidence–based Models in NLP
General Framework

· Consider evidence as an additional input of the model
Open Question Answering

- First large-scale evaluation of domain-independent QA systems.
- Participants were given 200 fact-based, short-answer questions.
- Each question was guaranteed to have at least one document in the collection that explicitly answered the question.
- Participants returned a ranked list of [document-id, answer-string] pairs per question such that each answer string was believed to contain an answer to the question.

Sparse Retrieval Model (DrQA)

**Document Retriever + Document Reader**
- Document retriever: finding relevant articles from 5 million Wikipedia articles
- Document reader (reading comprehension system): identifying the answer spans from those articles

Q: How many of Warsaw's inhabitants spoke Polish in 1933?

**Document Retriever**
- TF-IDF bag-of-words vectors + efficient bigram hashing (Weinberger et al., 2009)

**Document Reader**
- **Task**: given paragraph P and question Q, the goal is to find a span A in the paragraph which answers the question.
- **Model**: similar to AttentiveReader (Hermann et al, 2015; Chen et al, 2016). We aim to keep it simple!

[Diagram showing the flow of the model with Bidirectional LSTMs and equations for predicting start and end tokens]

**Datasets:**
- SQuAD (Rajpurkar et al, 2016)
- TREC (Baudiš and Šedivý, 2005)
- WebQuestions Freebase (Berant et al, 2013)
- WikiMovies (Miller et al, 2016)

Dense Retrieval Model (DPR, Dense Passage Retrieval)

Who is the bad guy in lord of the rings? Sala Baker is an actor and stuntman from New Zealand. He is best known for portraying the villain Sauron in the Lord of the Rings trilogy.

## DPR Results

<table>
<thead>
<tr>
<th>Training</th>
<th>Model</th>
<th>NQ</th>
<th>TriviaQA</th>
<th>WQ</th>
<th>TREC</th>
<th>SQuAD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single</td>
<td>BM25+BERT (Lee et al., 2019)</td>
<td>26.5</td>
<td>47.1</td>
<td>17.7</td>
<td>21.3</td>
<td>33.2</td>
</tr>
<tr>
<td>Single</td>
<td>ORQA (Lee et al., 2019)</td>
<td>33.3</td>
<td>45.0</td>
<td>36.4</td>
<td>30.1</td>
<td>20.2</td>
</tr>
<tr>
<td>Single</td>
<td>HardEM (Min et al., 2019a)</td>
<td>28.1</td>
<td>50.9</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Single</td>
<td>GraphRetriever (Min et al., 2019b)</td>
<td>34.5</td>
<td>56.0</td>
<td>36.4</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Single</td>
<td>PathRetriever (Asai et al., 2020)</td>
<td>32.6</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>56.5</td>
</tr>
<tr>
<td>Single</td>
<td>REALM\textsubscript{Wiki} (Guu et al., 2020)</td>
<td>39.2</td>
<td>-</td>
<td>40.2</td>
<td>46.8</td>
<td>-</td>
</tr>
<tr>
<td>Single</td>
<td>REALM\textsubscript{News} (Guu et al., 2020)</td>
<td>40.4</td>
<td>-</td>
<td>40.7</td>
<td>42.9</td>
<td>-</td>
</tr>
<tr>
<td>Single</td>
<td>BM25</td>
<td>32.6</td>
<td>52.4</td>
<td>29.9</td>
<td>24.9</td>
<td>38.1</td>
</tr>
<tr>
<td>Single</td>
<td>DPR</td>
<td><strong>41.5</strong></td>
<td>56.8</td>
<td>34.6</td>
<td>25.9</td>
<td>29.8</td>
</tr>
<tr>
<td></td>
<td>BM25+DPR</td>
<td>39.0</td>
<td>57.0</td>
<td>35.2</td>
<td>28.0</td>
<td>36.7</td>
</tr>
<tr>
<td>Multi</td>
<td>DPR</td>
<td><strong>41.5</strong></td>
<td>56.8</td>
<td><strong>42.4</strong></td>
<td>49.4</td>
<td>24.1</td>
</tr>
<tr>
<td></td>
<td>BM25+DPR</td>
<td>38.8</td>
<td><strong>57.9</strong></td>
<td>41.1</td>
<td><strong>50.6</strong></td>
<td>35.8</td>
</tr>
</tbody>
</table>

Table 4: End-to-end QA (Exact Match) Accuracy. The first block of results are copied from their cited papers. REALM\textsubscript{Wiki} and REALM\textsubscript{News} are the same model but pretrained on Wikipedia and CC-News, respectively. Single and Multi denote that our Dense Passage Retriever (DPR) is trained using individual or combined training datasets (all except SQuAD). For WQ and TREC in the Multi setting, we fine-tune the reader trained on NQ.
Joint Retrieval and Reader

- ORQA: Open–Retriever Question Answering
  - jointly learn the retriever and reader from question–answer string pairs
  - pre–train the retriever with an Inverse Cloze Task.

Pre-train Retrieval Model with Inverse Cloze Task

- In ICT, a sentence is treated as a pseudo-question, and its context is treated as pseudo-evidence.
- Given a pseudo-question, ICT requires selecting the corresponding pseudo-evidence out of the candidates in a batch.

Pre-train Retrieval Model with REALM

Knowledge Retriever: dense inner product model

\[
p(z \mid x) = \frac{\exp f(x, z)}{\sum_{z'} \exp f(x, z')},
\]

\[
f(x, z) = \text{Embed}_{\text{input}}(x)^\top \text{Embed}_{\text{doc}}(z),
\]

\[
\text{Embed}_{\text{input}}(x) = W_{\text{input}} \text{BERT}_{\text{CLS}}(\text{join}_{\text{BERT}}(x))
\]

\[
\text{Embed}_{\text{doc}}(z) = W_{\text{doc}} \text{BERT}_{\text{CLS}}(\text{join}_{\text{BERT}}(z_{\text{title}}, z_{\text{body}}))
\]

\[
\text{join}_{\text{BERT}}(x) = [\text{CLS}]x[\text{SEP}]
\]

\[
\text{join}_{\text{BERT}}(x_1, x_2) = [\text{CLS}]x_1[\text{SEP}]x_2[\text{SEP}]
\]
Pre-train Retrieval Model with REALM

Unsupervised pre-training

Unlabeled text

The [MASK] at the top of the pyramid (x)

Neural Knowledge Retriever (θ)

retrieve

Knowledge-Augmented Encoder (ϕ)

Answer

[MASK] = pyramidion (y)

Sample

Pre-training corpus (X)

Unsupervised pre-training

Supervised fine-tuning

Input query

what's the angle of an equilateral triangle? (x)

Neural Knowledge Retriever (θ)

retrieve

Knowledge-Augmented Encoder (ϕ)

Answer

60 degrees (y)

Sample

Supervised data

Realm: Retrieval-augmented language model pre-training
Summary

- Topics covered by this talk
  - Logic-based Models in NLP
  - Neural-Symbolic Models in NLP
  - Evidence-based Models in NLP

- Directions worth pursuing
  - Interpretable models and methods
  - Deep understanding with reasoning ability
Challenge: Performance vs. Explainability

Learning Techniques (today)
- Neural Nets
- Deep Learning
- Statistical Models
- AOGs
- SVMs
- Graphical Models
- Bayesian Belief Nets
- Ensemble Methods
- Random Forests
- Decision Trees
- Neural Nets

Explainability (notional)
- Prediction Accuracy
- Explainability

- Deep Explanation
  - Modified deep learning techniques to learn explainable features
- Interpretable Models
  - Techniques to learn more structured, interpretable, causal models
- Model Induction
  - Techniques to infer an explainable model from any model as a black box

Explainable Artificial Intelligence (XAI), David Gunning, DARPA/I2O
We are hiring!

- Both interns and employees.

- Topics:
  - Semantic Parsing
  - Code Intelligence
  - Machine Reasoning

- Send email to dutang@microsoft.com