Algorithms that learn to think on their feet
What is NLP?

➢ Fundamental goal: deep understanding of text
  ➢ Not just string processing or keyword matching
➢ End systems that we want to build
  ➢ Simple: Spelling correction, text categorization, etc.
  ➢ Complex: Speech recognition, machine translation, information extraction, dialog interfaces, question answering
➢ Unknown: human-level comprehension (more than just NLP?)
Why is language hard?

- Ambiguity abounds (some headlines)
  - Iraqi Head Seeks Arms
  - Teacher Strikes Idle Kids
  - Kids Make Nutritious Snacks
  - Stolen Painting Found by Tree
  - Local HS Dropouts Cut in Half
  - Enraged Cow Injures Farmer with Ax
  - Hospitals are Sued by 7 Foot Doctors
  - Ban on Nude Dancing on Governor's Desk
  - Scientists study whales from space

- Why are these funny?
- What does ambiguity imply about the role of learning?
Despite ambiguity, language is predictable

I like my coffee with cream and asparagus.

This is crummy weather for San ta Claus.

➢ The brain uses this information!

➢ Can we use predictability to make decisions before all of the input is observed?

YES!!!
Outline

Quizbowl (Incremental Question Answering)

Alvin Grissom II

He He

Simultaneous Machine Interpretation

Ich bin mit dem Zug nach Ulm traveled by train to Ulm

(. . . . . . waiting. . . . . . )

Hal Daum III (me@hal3.name)
Simultaneous (machine) interpretation

➢ Dozens of defendants
➢ Judges from four nations (three languages)
➢ Status quo: speak, then translate
➢ After Nuremberg, simultaneous translations became the norm
➢ Long wait → bad conversation

Nuremberg Trials
Why simultaneous interpretation is hard

- Human languages have vastly different word orders
  - About half are OV, the other half are VO
  - This comes with a lot more baggage than just verb-final

Running (German/English) Example:

Ich bin mit dem Zug nach Ulm gefahren
I am with the train to Ulm traveled
I (……. waiting.……) traveled by train to Ulm
Model for interpretation decisions

- We have a set of actions (predict / translate)
  - Wait
  - Predict clause-verb
  - Predict next word
  - Commit (“speak”)

- In a changing environment (state)
  - The words we've seen so far
  - Our models' internal predictions

- With a well defined notion of “optimal action” at training time
Example of interpretation trajectory

Ich bin mit dem Zug nach Ulm gefahren
I am with the train to Ulm traveled
I (....... waiting....... ) traveled by train to Ulm
DAgger: Dataset Aggregation

- Collect trajectories from expert $\pi^\ast$
- Dataset $D_0 = \{ (s, \pi^\ast(s)) \mid s \sim \pi^\ast \}$
- Train $\pi_1$ on $D_0$
- Collect new trajectories from $\pi_1$
  - But let the expert steer!
- Dataset $D_1 = \{ (s, \pi^\ast(s)) \mid s \sim \pi_1 \}$
- Train $\pi_2$ on $D_0 \cup D_1$
- In general:
  - $D_n = \{ (s, \pi^\ast(s)) \mid s \sim \pi_n \}$
  - Train $\pi_n$ on $\bigcup_{i<n} D_i$

If $N = T \log T$, $L(\pi_n) < T\epsilon_N + O(1)$ for some $n$
Evaluating performance and baselines

Grissom II et al., EMNLP 2014

Source Sentence

Psychic

Monotone

Batch

Policy Prediction

Good Translation

Bad Translation

Good Translation

Bad Translation

Good Translation

Bad Translation

Good Translation

Bad Translation
Evaluating performance and baselines

Source Sentence

Psychic

He went to the store

Monotone

He

Batch

Policy Prediction

He

Good Translation

Bad Translation

Good Translation

Bad Translation

Good Translation

Bad Translation

Good Translation

Bad Translation

Grissom II et al., EMNLP 2014

Hal Daumé III (me@hal3.name)
Evaluating performance and baselines

Source Sentence

- Psychic: He went to the store

Monotone

- He

Batch

Policy Prediction

- He

Grissom II et al., EMNLP 2014
Evaluating performance and baselines

Source Sentence

Psychic

Monotone

Batch

Policy Prediction

He went to the store

He to the

He went

Good Translation

Bad Translation

Good Translation

Bad Translation

Good Translation

Bad Translation

Good Translation

Bad Translation
Evaluating performance and baselines
Evaluating performance and baselines

(Grisom II et al., EMNLP 2014)
Evaluating performance and baselines

Grissom II et al., EMNLP 2014

Hal Daumé III (me@hal3.name) Algorithms that think on their feet
Training the policy

➢ Actions:
  ➢ Commit: translate(revealed words)
  ➢ Predict (verb/next): translate(revealed + predicted)
  ➢ Wait: get_next_words()

➢ Delayed feedback: latency BLEU

➢ Features:
  ➢ Output & confidence of predictors
  ➢ Internal translation / language model scores
  ➢ Previous decisions made by policy
Evaluating performance

![Graph showing performance evaluation with different algorithms: Batch, Monotone, Optimal, and Learned. The Y-axis represents performance scores ranging from 0 to 1.4. The graph illustrates the comparative performance of these algorithms in terms of their efficiency and accuracy.](image-url)
Quizbowl
(Incremental Question Answering)

Mohit Iyyer

He He

double-e with
fridge

train to Ulm
Simultaneous Machine Interpretation
Humans doing incremental prediction

- Game called “quiz bowl”
- Two teams play each other
  - Moderator reads a question
  - When a team knows the answer, they buzz in
  - If right, they get points; otherwise, rest of the question is read to the other team
- Hundreds of teams in the US alone
- Example . . .
Quizbowl example

With Leo Szilard, he invented a doubly-eponymous
Quizbowl example

With Leo Szilard, he invented a doubly-eponymous refrigerator with no moving parts. He did not take interaction with neighbors into account when formulating his theory.
Quizbowl example

With Leo Szilard, he invented a doubly-eponymous refrigerator with no moving parts. He did not take interaction with neighbors into account when formulating his theory of heat capacity, so
With Leo Szilard, he invented a doubly-eponymous refrigerator with no moving parts. He did not take interaction with neighbors into account when formulating his theory of heat capacity, so Debye adjusted the theory for low temperatures. His summation convention automatically sums repeated indices in tensor products. His name is attached to the A and B coefficients.
Quizbowl example

With Leo Szilard, he invented a doubly-eponymous refrigerator with no moving parts. He did not take interaction with neighbors into account when formulating his theory of heat capacity, so Debye adjusted the theory for low temperatures. His summation convention automatically sums repeated indices in tensor products. His name is attached to the A and B coefficients for spontaneous and stimulated emission, the subject of one of his multiple groundbreaking 1905 papers. He further developed the model of statistics sent to him by
Quizbowl example

With Leo Szilard, he invented a doubly-eponymous refrigerator with no moving parts. He did not take interaction with neighbors into account when formulating his theory of heat capacity, so Debye adjusted the theory for low temperatures. His summation convention automatically sums repeated indices in tensor products. His name is attached to the A and B coefficients for spontaneous and stimulated emission, the subject of one of his multiple groundbreaking 1905 papers. He further developed the model of statistics sent to him by Bose to describe particles with integer spin. For 10 points, who is this German physicist best known for formulating
With Leo Szilard, he invented a doubly-eponymous refrigerator with no moving parts. He did not take interaction with neighbors into account when formulating his theory of heat capacity, so Debye adjusted the theory for low temperatures. His summation convention automatically sums repeated indices in tensor products. His name is attached to the A and B coefficients for spontaneous and stimulated emission, the subject of one of his multiple groundbreaking 1905 papers. He further developed the model of statistics sent to him by Bose to describe particles with integer spin. For 10 points, who is this German physicist best known for formulating the special and general theories of relativity?
Solving incrementally

- **Action:** buzz now or wait
  - Content Model is constantly generating guesses
  - Oracle provides examples where it is correct
  - The Policy generalizes to test data
  - Features represent our state

---

**Qatar**

From Wikipedia, the free encyclopedia

For other places with the same name, see Qatar (disambiguation).

Qatar ( /kɑːˈtɑːr/, /kɑːˈtɑɹ/ or /kɑːˈtær/; Arabic: قطر, romanized: Qatar [qatˤar]; local the State of Qatar (Arabic: دولة قطر, Dawlat Qatar), is a sovereign Arab the small Qatar Peninsula on the northeastern coast of the Arabian Penin to the south, with the rest of its territory surrounded by the Persian Gulf. , from the nearby island kingdom of Bahrain. In 2013, Qatar's total populat and 1.5 million expatriates.[8]
Evaluation methodology

- Mechanical Turk to collect human data
- 7000 questions were answered in the first day
- Over 43,000 questions were answered in the space of two weeks
- Total of 461 unique users
- Leaderboard to encourage users

Big problem:

“this man shot at Aaron Burr”

is very different from

“Aaron Burr shot at this man”
Challenge: modeling compositionality
Challenge: modeling compositionality
Challenge: modeling compositionality

- **invented**
  - **he**
    - **a**
  - **fridge**
    - **double-e**
      - **with**
        - **parts**
          - **no**
          - **moving**
Challenge: modeling compositionality

\[
e(\text{fridge}) = f( w_{\text{fridge}} + \forall \mathbb{V} [ e(a) e(\text{double-e}) e(\text{with}) ] )
\]
Results on question-answering task

History Questions

- BOW (QB)
- RNN (QB)
- BOW (wiki)
- Combined

(Iyer et al., ACL 2014)
Results on question-answering task
Moving to more general frameworks

- Lots of NLP (+al) problems can be cast at test time as integer linear programs
- ILPs are usually solved using branch and bound

Branch and bound involves a complex heuristic search. Can we learn to perform this search efficiently?
Some intuition

- A good search strategy should:
  - find a good incumbent solution early
  - identify non-promising nodes before expansion

- “Good” varies depending on your position in the tree
  - DFS should only be used at nodes that promise to lead to a good feasible solution that may replace the incumbent
  - Best-bound-first search can quickly discard unpromising nodes, but should not be used frequently at the top

- We will learn a heuristic based on features that can capture this intuition
Training and experiments

- Same algorithm as before
- Four (standard) ILP datasets (non-NLP-based)
- Comparison to:
  - DFS (baseline)
  - Gurobi (thousands of person-hours of effort)

Measures:
- Optimality Gap, Integrality Gap, and improvement from initial heuristic solution

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Ours (Dagger training)</th>
<th>DFS</th>
<th>Gurobi</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OGap</td>
<td>IGap</td>
<td>Impr</td>
</tr>
<tr>
<td>MIK</td>
<td>0.23</td>
<td>16.63</td>
<td>4.39</td>
</tr>
<tr>
<td>Regions1</td>
<td>0.54</td>
<td>4.53</td>
<td>10.57</td>
</tr>
<tr>
<td>Regions2</td>
<td>1.22</td>
<td>6.76</td>
<td>19.36</td>
</tr>
<tr>
<td>Hybrid</td>
<td>0.87</td>
<td>20.28</td>
<td>24.46</td>
</tr>
</tbody>
</table>

We achieve less than 1.2 optimality gap while exploring 0.05%, 1.5%, 5.1% and 47% of the nodes explored by Gurobi!
Reasoning with incomplete information is useful for *speed* and *modeling*.

*Imitation learning* can help us build such systems.

- Plug: even when you can't construct a perfect oracle (see *LOLS*, ICML 2015).

Wide range of new, interesting problems to work on:
- How to learn from human interpreters?
- How to learn to compete?
- How to *not need* BOW in deepNN models?

Thanks! Questions?