learning language through interaction

Hal Daumé III | University of Maryland | me@hal3.name | @haldaume3
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?)
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

Nuremburg Trials
Skype translator demo
Skype translator demo
“Sorry Melanie, say that again”
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

*dinosaur*-TOP *store*-LOC *go*-PAST
the dinosaur went to the store

*food*-OBJ *buy*-DESIRE *dinosaur*-TOP *store*-LOC *go*-PAST
the dinosaur who wanted to buy food went to the store
One slide on statistical machine translation

- Collect a bunch of text that's been translated by humans ("parallel corpus")
- Break it into "aligned" sentence pairs
- Learn statistical model to map from the "source" to "target"
- Apply that model to novel "unseen" sentences that you want to translate
つまり例えばこの表現一は認識できますが二から四は認識できない

**Batch**
They might recognize expression one but not expression two to four.

**Interp**
The phrase number one only is accepted and phrases two, three, four were not accepted.
General diffs of Interp vs Batch

- Inversion
  - Segmentation into multiple sentences
  - Passivization of single sentence

- Word generalization
  - (lower retrieval time)

- Summarization and omission
  - (to catch up)
Example (gen + segment)

(S) この日本語の待遇表現の特徴ですが英語から日本語へ直訳しただけでは表現できないといった特徴があります

(Batch) One of the characteristics of honorific Japanese is that it can not be adequately expressed when using a direct translation from English to Japanese.

(Interp) Now let me talk about the characteristic of the Japanese polite expressions.  <segment/> And such expressions can not be expressed enough just by translating directly.
Example (gen + summarize)

(S) で三番目の特徴としてはですねえ出来る限り自然な日本語の話言葉とてその出力をするといったような特徴があります。

(Batch) Its third characteristic is that its output is, as much as possible, in the natural language of spoken (( Japanese )).

(Interp) And the third feature is that the translation could be produced in a very natural spoken language.
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

- Have example data from human interpreters that demonstrates what to do!
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 $\pi^*$
- Dataset $D_0 = \{(s, \pi^*(s)) | s \sim \pi^*\}$
- Train $\pi_1$ on $D_0$
- Collect new trajectories from $\pi_1$
  - But let the expert steer!
- Dataset $D_1 = \{(s, \pi^*(s)) | s \sim \pi_1\}$
- Train $\pi_2$ on $D_0 \cup D_1$

In general:

- $D_n = \{(s, \pi^*(s)) | s \sim \pi_n\}$
- Train $\pi_n$ on $\bigcup_{i<n} D_i$

If $N = T \log T$, $L(\pi_n) < T \epsilon_N + O(1)$ for some $n$
Training the policy

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

➢ Features:
  ➢ Output & confidence of predictors
  ➢ Internal translation / language model scores
  ➢ Previous decisions made by policy
Evaluating performance
But wait, how good are humans....???

- Everyone makes errors:
  - Best human interpreters make some
  - Most human interpreters make many

- Mimicking human behavior is suboptimal!

- But then, what can we do?
Learning to search: AggraVaTe

1. Let learned policy $\pi$ drive for $t$ timesteps to obs. $o$

2. For each possible action $a$:
   - Take action $a$, and let expert $\pi^{\text{ref}}$ drive the rest
   - Record the overall loss, $c_a$

3. Update $\pi$ based on example:
   \[(o, \langle c_1, c_2, \ldots, c_K \rangle)\]

4. Goto (1)
Learning to search: AggraVaTe

On the bright side:
Doesn't need immediate feedback on every decision

On the not-so-bright side:
Have to evaluate a ton of trajectories
Need to assume driver is optimal
Motivating setting

Daume 3 Hal
iPhone
November 11, 2016 at 9:12 PM

Transcription Beta
“________ me I are on PCH and have fun be there in about hopefully less than a half hour funny minutes or so anyway see you soon love you bye.”

Was this transcription useful or not useful?
1. Let learned policy $\pi$ drive for $t$ timesteps to obs. $o$

2. Choose a single “deviation” action $a$:
   - Take action $a$, and let policy $\pi^{out}$ drive the rest
   - Record the overall loss, $c_a$

3. Estimate the cost of all $a' \neq a$
   by regression

4. Update $\pi$ based on example:
   $(o, \langle c_1, c_2, \ldots, c_K \rangle)$

5. Goto (1)
BanditLOLS: learning from bandit feedback

- Human feedback (thumbs up/thumbs down) only required on one prediction per input

- Evaluation balances exploration & exploitation

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Exploration</th>
<th>POS Accuracy</th>
<th>Dependency UAS</th>
<th>Chunking F-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference</td>
<td>-</td>
<td>47.24</td>
<td>44.15</td>
<td>74.73</td>
</tr>
<tr>
<td>LOLS</td>
<td>$\varepsilon$-greedy</td>
<td>2.29</td>
<td>18.55</td>
<td>31.76</td>
</tr>
</tbody>
</table>

All code available, integrated into Vowpal Wabbit (github.com/hal3/)
Humans are not optimal at *games* either!

- 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
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.
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?
Evaluation methodology

- Mechanical Turk to collect human data
- 7000 questions were answered in the first day
- Over 43000 questions were answered in the space of two weeks
- Total of 461 unique users
- Leaderboard to encourage users
Deep Q-learning

- **Q-learning**
  - Estimate $Q(s, a)$: expected future reward starting from $(s,a)$
  - Reward function: quiz bowl scoring

  After the death of this man, The Liberators’ Civil War was fought in
  
  $\begin{align*}
  0.11 & \text{ Mithridates} \\
  0.09 & \text{ Julius Caesar} \\
  0.07 & \text{ Sulia}
  \end{align*}$

  What is the **final** score I can get if I BUZZ / WAIT now

- **Deep Q-learning**
  - Use a neural network to approximate the Q-function

\[
\begin{align*}
\phi_t^s & \quad \text{State features} \\
\rightarrow & \\
\rightarrow & \\
\rightarrow & \\
Q(s_t, a_t) & \quad \text{Square loss}
\end{align*}
\]
Opponent as part of the world

DQN-self:

*imitate the expert: buzz whenever the current prediction is correct*

DQN-world:

*reinforcement learning from long term feedback*

Interacting with opponent helps learn!

Reward against human

<table>
<thead>
<tr>
<th>Reward</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>DQN-world</td>
<td>0.9</td>
</tr>
<tr>
<td>DQN-self</td>
<td>0.75</td>
</tr>
</tbody>
</table>

(He et al., ICML 2016)
Exploit the opponent?

- **Human performance**
  - Fast and good
  - Cautious
  - Aggressive

- **Machine performance**
  - More words seen, higher accuracy

(Hal Daumé III (me@hal3.name))
How to incorporate opponent?

**Concatenation**

\[
\phi_t^s \rightarrow h^s \rightarrow h^o; h^s \rightarrow h \rightarrow Q \cdot \pi^o (s_t, a_t)
\]

**Mixture-of-Experts**

\[
\begin{align*}
\text{experts} & \quad \phi_t^s \rightarrow h^s \\
& \quad \cdot \cdot \cdot \quad Q_k \\
\text{gating} & \quad \phi_t^o \rightarrow h^o \\
& \quad \cdot \cdot \cdot \quad w_i \rightarrow Q \cdot \pi^o (s_t, a_t)
\end{align*}
\]
Overall reward against human players

![Bar chart showing overall reward against human players for DRON-concat, DRON-MoE, DQN-world, and DQN-self. The chart indicates that DRON-MoE has the highest reward, followed by DRON-concat, DQN-world, and DQN-self.](image)
Students

Leonardo Claudino
Alvin Grissom II
He He
Mohit Iyyer
Sudha Rao
Amr Sharaf

Other Collaborators

Alekh Agarwal
Jordan Boyd-Graber
Alina Beygelzimer
Kai-Wei Chang
Akshay Krishnamurthy
John Langford
Paul Mineiro
Stéphane Ross
Learning from interaction requires:

- Build a system and let (simulated) people use it...
- They provide some feedback (implicit/explicit)...
- System improves performance over time...

Open questions:

- How can we best use offline-acquired data?
- Can we (substantially) reduce the amount of interaction required to learn something interesting?
- How can we convert implicit feedback into an implicit reward signal?

THANKS! Questions?