Interpretation

Advanced Machine Learning for NLP
Jordan Boyd-Graber
HUMAN IN THE LOOP EVALUATION
Evaluation

- User study
- 40 minutes
- Sort documents into categories
- What information / interface helps best
Evaluation

- User study
- 40 minutes
- Sort documents into categories
- What information / interface helps best
Evaluation

- User study
- 40 minutes
- Sort documents into categories
- What information / interface helps best
  - Train a classifier on human examples
  - Compare classifier labels to expert judgements
Evaluation

- User study
- 40 minutes
- Sort documents into categories
- What information / interface helps best
  - Train a classifier on human examples *(don’t tell them how many labels)*
  - Compare classifier labels to expert judgements
Evaluation

- User study
- 40 minutes
- Sort documents into categories
- What information / interface helps best
  - Train a classifier on human examples (don’t tell them how many labels)
  - Compare classifier labels to expert judgements (purity)

\[
purity(U, G) = \frac{1}{N} \sum_l \max_j |U_l \cap G_j|,
\]  

(1)
Which is more Useful?

Who should drive?
Which is more Useful?

Active Learning

Topic Models

TOPIC 1: computer, technology, system, service, site, phone, internet, machine

TOPIC 2: sell, sale, store, product, business, advertising, market, consumer

TOPIC 3: play, film, movie, theater, production, star, director, stage
To provide for payments to certain natural resource trustees to assist in re... A bill to authorize the Secretary of the Army to carry out activities to man... A bill to prevent forfeited fishing vessels from being transferred to private ... To reauthorize various Acts relating to Atlantic Ocean marine fisheries. To amend the Magnuson-Stevens Fishery Conservation and Management... To prevent forfeited fishing vessels from being transferred to private parti... To amend the Magnuson-Stevens Fishery Conservation and Management... A bill to require the Secretary of the Army to study the feasibility of the hy... Making appropriations for disaster relief requirements for the fiscal year e... To rescind any unobligated discretionary appropriations returned to the F... To amend the Robert T. Stafford Disaster Relief and Emergency Assistanc... Making appropriations for energy and water development and related age...
To rescind any unobligated discretionary appropriations returned to the F...
Active learning if time is short
Active learning can help topic models
Topic models help users understand the collection
Moral: machines and humans together (if you let them)
This city’s mass transit system’s director, Paul Wiedefeld, undertook SafeTrack to restore reliability of the system. For ten points each ...

[10] Name this American city with stations designed by Harry Weese.

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Washington Metro</td>
<td>stations designed by Harry Weese</td>
</tr>
<tr>
<td>Baltimore</td>
<td>Paul Wiedefeld</td>
</tr>
<tr>
<td>Washington, D.C.</td>
<td>this American city</td>
</tr>
</tbody>
</table>

- Have a task humans are doing
- Show them predictions of good system
- Measure how much human performance improves (degrades)
Improving Visualizations

• If you have feedback on performance, can use RL
• Change parameters of visualizations
  ○ How many/which hypotheses are shown
  ○ How much/which evidence is shown
  ○ Is anything shown at all?
• Can optimize for team performance
Improving Visualizations

- If you have feedback on performance, can use RL
- Change parameters of visualizations
  - How many/which hypotheses are shown
  - How much/which evidence is shown
  - Is anything shown at all?
  - Which questions are shown
- Can optimize for team performance or just human performance
Evaluation Takeaway

• Measure what you care about
• If you care about prediction, likelihood is good
• If you care about a particular task, measure that