End-to-End Reinforcement Learning for Automatic Taxonomy Induction

Yuning Mao, Xiang Ren, Jiajing Shen, Xiaotao Gu, Jiawei Han
Department of Computer Science, University of Illinois at Urbana-Champaign, USA
Department of Computer Science, University of Southern California, CA, USA
{yuningm2, js2, xiaotao2, hanj}@iillinois.edu  xiangren@usc.edu

# Task

- **Goal:** Automatic Taxonomy Induction
  - Input: 1) a set of training taxonomies
  - 2) related resources (e.g., background text corpora).
- **Output:** given vocabulary \( V_T \) construct a taxonomy \( T \) by adding terms from \( V_T \).

# Hypernymy Detection:

- Hypernymy pairs (is-a relations) are extracted: (banana, fruit), (panda, mammal), ...
- A noisy hypernym graph is generated

# Hypernymy Organization:

- Organize is-a term pairs into a tree-structured hierarchy -> graph pruning
- maximum spanning tree (MST) (Bansal et al., 2014 [1])
- minimum cost flow (MCF) (Gupta et al., 2017 [2])
- other pruning heuristics (Panchenko et al., 2016 [3])

# RL Component - States:

- The state at time \( t \) comprises:
  - the current taxonomy \( T_t \) (terms & structure)
  - the remaining vocabulary \( V_T \)
  - Update deterministically

# RL Component - Actions:

1. select a term \( x \) from the remaining vocabulary \( V_T \)
2. remove \( x \) from \( V_T \)
3. attach \( x \) as a hyponym of one term \( y \) that is already on the current taxonomy \( T_t \)
- Action Space: \( |V_T| \times |T_t| \)
- Episode Length: \( |V_T| \)

# RL Component - Rewards:

- Evaluation Metrics:
  - Ancestor-F1
  - Edge-F1
- Reward Shaping: \( R_t = \text{Edge-F1}(t) - \text{Edge-F1}(t-1) \)

# Action (term-pair) Representation

- Dependency Paths between \( x \) and \( y \)
- \( W_x \): Word Embedding of \( x \)
- \( W_y \): Word Embedding of \( y \)
- \( f(x, y) \): Surface (Ends with, Contains, etc.), Frequency (pattern-based co-occurs info), and Generality (edge not too general or narrow) Features

# Methodology

# Experimental Results

- **Compared methods:**
  - TAXI [3]: pattern-based method that ranked 1st in the SemEval-2016 Task 13 competition
  - HyperNET [4]: state-of-the-art hypernymy detection method
  - HypeNET + MST (maximum spanning tree): post-processing of HyperNET to prune the hypernym graph into a tree
  - Bansal et al. (2014) [1]: state-of-the-art taxonomy induction method
  - SubSeq [2]: state-of-the-art results on the SemEval-2016 Task 13
  - Taxo-RL (RE, with virtual root embedding), Taxo-RL (NR, with new root addition), Taxo-RL (partial, allows partial taxonomy), Taxo-RL (full, has to use all terms in the vocabulary)

# Performance Study on End-to-End Taxonomy Induction:

- WordNet (532/144/144 taxonomies for training, validation, and test set, size [10, 50, depth=4, animals, daily necessities, etc.])

<table>
<thead>
<tr>
<th>Model</th>
<th>( P_1 )</th>
<th>( R_1 )</th>
<th>( F_1 )</th>
<th>( P_L )</th>
<th>( R_L )</th>
<th>( F_L )</th>
</tr>
</thead>
<tbody>
<tr>
<td>TAXI (TAG)</td>
<td>50.1</td>
<td>32.7</td>
<td>39.6</td>
<td>33.8</td>
<td>26.8</td>
<td>29.9</td>
</tr>
<tr>
<td>TAXI (tree)</td>
<td>67.5</td>
<td>30.8</td>
<td>42.3</td>
<td>41.1</td>
<td>23.1</td>
<td>29.6</td>
</tr>
<tr>
<td>SubSeq</td>
<td>51.6</td>
<td>36.4</td>
<td>42.7</td>
<td>37.5</td>
<td>24.2</td>
<td>29.4</td>
</tr>
<tr>
<td>Taxo-RL (Partial)</td>
<td>47.2</td>
<td>54.6</td>
<td>50.6</td>
<td>52.3</td>
<td>32.3</td>
<td>32.3</td>
</tr>
<tr>
<td>Taxo-RL (RL)</td>
<td>61.6</td>
<td>41.7</td>
<td>47.9</td>
<td>38.8</td>
<td>34.8</td>
<td>36.7</td>
</tr>
<tr>
<td>Taxo-RL (NR)</td>
<td>26.3</td>
<td>83.8</td>
<td>43.5</td>
<td>44.8</td>
<td>28.8</td>
<td>35.1</td>
</tr>
<tr>
<td>Taxo-RL (Full)</td>
<td>48.6</td>
<td>34.4</td>
<td>48.9</td>
<td>45.9</td>
<td>33.0</td>
<td>41.8</td>
</tr>
<tr>
<td>Taxo-RL (RL)</td>
<td>68.3</td>
<td>52.9</td>
<td>59.6</td>
<td>57.9</td>
<td>37.9</td>
<td>37.9</td>
</tr>
</tbody>
</table>

# Testing on Hypernymy Organization:

- SemEval-2016 Task 13 (test set only, hundreds of terms, environment, science domain)

<table>
<thead>
<tr>
<th>Model</th>
<th>( P_1 )</th>
<th>( R_1 )</th>
<th>( F_1 )</th>
<th>( P_L )</th>
<th>( R_L )</th>
<th>( F_L )</th>
</tr>
</thead>
<tbody>
<tr>
<td>TAXI (TAG)</td>
<td>74.1</td>
<td>39.8</td>
<td>42.9</td>
<td>33.8</td>
<td>25.6</td>
<td>28.9</td>
</tr>
<tr>
<td>TAXI (tree)</td>
<td>55.9</td>
<td>30.8</td>
<td>39.8</td>
<td>31.8</td>
<td>20.6</td>
<td>23.1</td>
</tr>
<tr>
<td>SubSeq</td>
<td>51.6</td>
<td>36.4</td>
<td>42.7</td>
<td>37.5</td>
<td>24.2</td>
<td>29.4</td>
</tr>
<tr>
<td>Taxo-RL (Partial)</td>
<td>47.2</td>
<td>54.6</td>
<td>50.6</td>
<td>52.3</td>
<td>32.3</td>
<td>32.3</td>
</tr>
<tr>
<td>Taxo-RL (RL)</td>
<td>61.6</td>
<td>41.7</td>
<td>47.9</td>
<td>38.8</td>
<td>34.8</td>
<td>36.7</td>
</tr>
<tr>
<td>Taxo-RL (NR)</td>
<td>26.3</td>
<td>83.8</td>
<td>43.5</td>
<td>44.8</td>
<td>28.8</td>
<td>35.1</td>
</tr>
<tr>
<td>Taxo-RL (Full)</td>
<td>48.6</td>
<td>34.4</td>
<td>48.9</td>
<td>45.9</td>
<td>33.0</td>
<td>41.8</td>
</tr>
<tr>
<td>Taxo-RL (RL)</td>
<td>68.3</td>
<td>52.9</td>
<td>59.6</td>
<td>57.9</td>
<td>37.9</td>
<td>37.9</td>
</tr>
</tbody>
</table>

# Ablation Study:

- Multiple sources of information are complementary to each other

<table>
<thead>
<tr>
<th>Model</th>
<th>( P_1 )</th>
<th>( R_1 )</th>
<th>( F_1 )</th>
<th>( P_L )</th>
<th>( R_L )</th>
<th>( F_L )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distributional Info</td>
<td>72.3</td>
<td>24.3</td>
<td>25.6</td>
<td>13.8</td>
<td>27.4</td>
<td>27.4</td>
</tr>
<tr>
<td>Pooled-based Info</td>
<td>72.8</td>
<td>48.5</td>
<td>33.7</td>
<td>27.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D + P</td>
<td>36.7</td>
<td>39.4</td>
<td>37.9</td>
<td>38.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D + P + Surface Features</td>
<td>41.3</td>
<td>49.2</td>
<td>44.9</td>
<td>35.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D + P + S + FG</td>
<td>52.9</td>
<td>58.6</td>
<td>55.6</td>
<td>43.8</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

# Case Studies:

- Numbers indicate the orders of term pair selections
  - (air filter, filter, 2) -> correct root
  - (fuel filter, filter, 3), (coffee filter, filter, 4) -> substring inclusion
  - (colander, strainer, 13), (glass wool, filter, 16) -> path and distributional info

# Conclusion and References

- **Conclusion:**
  - Learns the representations of term pairs by optimizing a holistic tree metric
  - Reduces error propagation between two phases
  - Achieves new state-of-the-art results

- **References:**