FUSION OF KNOWLEDGE-BASED AND DATA-DRIVEN APPROACHES TO GRAMMAR INDUCTION

Introduction


  I WANT TO TRAVEL TO NEW YORK
  I WANT TO TRAVEL TO <TOCITY>
  <TOCITY> = Miami | Hawaii | Alaska

- Two grammar induction approaches: data-driven (bottom-up) and knowledge-based (top-down).

  Rule: Merge of bottom-up and top-down grammar.

  Example: Merge of bottom-up and top-down grammar.

  Corpus creation: Generate sentences via top-down grammar. Use top-down corpus as input for bottom-up induction.

  Two terminal rule induction:
  - Terminal rules: Leaf lexicon and terminal rules.
  - Non-terminal rules: Conceptualized as fragments composed of words and terminal rules.

Bottom-Up Induction

Seed examples guide corpus extraction of enhance candidates. Rules are enhanced with most common frequent candidates.

1. Terminal rule induction: Based on distributional hypothesis of meaning. Semantic Distance: Context-based Manhattan Norm. <CITY> = Boston | Denver <STATE> = Miami | Hawaii | Alaska


Late Fusion

Induction approaches run independently and resulting grammars are fused. Grammars vary in structure: two techniques to deal with overgeneralization and overlap.

1. Simple Union: Merge of bottom-up and top-down grammar.
2. Augmentation: Use one grammar as input to enhance the other.

   (a) Concept Matching: (Optional) Many-to-one terminal rule matching to address terminal rule inconsistency.
   (b) Mapping Level:
      - Rule: Overall rule similarity considered for mapping to rules.
      - Fragment: Fragments directly map to rules, regardless of encapsulating rule.

Analysis of Experiments

- Domain: Air Travel (English).
- Bottom-up Induction: 2 seed fragments, 10 requests per rule.
- Top-down Induction: Air Travel ontology and hand-crafted lexicon comprising 67 ontology element lexicalizations.
- Top-down and web-harvested bottom-up corpora comprise of 6,017 and 17,564 sentences respectively.
- Early and Mid Fusion: Runs repeated x10, evaluations averaged.
- Evaluation: non-terminal rules wrt. F-measure (Fm), where Precision (Pr): percentage of correct enhancements per rule. Recall (Rc): correct enhancements’ coverage wrt. gold standard.

Table: Fusion of grammar approaches

<table>
<thead>
<tr>
<th>Fusion</th>
<th>Grammar</th>
<th>Pr</th>
<th>Rt</th>
<th>Fm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Early</td>
<td>-</td>
<td>0.65</td>
<td>0.44</td>
<td>0.52</td>
</tr>
<tr>
<td>Top-down (BU)</td>
<td>0.81</td>
<td>0.18</td>
<td>0.30</td>
<td>0.71</td>
</tr>
<tr>
<td>Mid Fusion</td>
<td>BU Induction</td>
<td>0.66</td>
<td>0.37</td>
<td>0.47</td>
</tr>
<tr>
<td>Early Fusion</td>
<td>BU Induction</td>
<td>0.63</td>
<td>0.54</td>
<td>0.52</td>
</tr>
<tr>
<td>Mid Fusion</td>
<td>BU Induction</td>
<td>0.64</td>
<td>0.52</td>
<td>0.58</td>
</tr>
<tr>
<td>Late Fusion</td>
<td>BU Induction</td>
<td>0.65</td>
<td>0.52</td>
<td>0.55</td>
</tr>
<tr>
<td>Late Fusion</td>
<td>BU Rule Augm.</td>
<td>0.27</td>
<td>0.46</td>
<td>0.34</td>
</tr>
<tr>
<td>Late Fusion</td>
<td>BU Frag. Augm.</td>
<td>0.58</td>
<td>0.58</td>
<td>0.58</td>
</tr>
<tr>
<td>Late Fusion</td>
<td>TD Rule Augm.</td>
<td>0.61</td>
<td>0.20</td>
<td>0.30</td>
</tr>
<tr>
<td>Late Fusion</td>
<td>TD Frag. Augm.</td>
<td>0.79</td>
<td>0.21</td>
<td>0.34</td>
</tr>
</tbody>
</table>

1. Early Fusion: Quality and size of top-down corpus affect performance, enhancement provides better results.
2. Mid Fusion: Outperforms bottom-up baseline, 0.58 on induction and 12.5% coverage improvement on enhancement.
3. Late Fusion: 20% relative improvement on simple union. Fragment mapping addresses overgeneralization, applies best on bottom-up augmentation. Concept matching reaches 0.55.

Aknowledgements

Partially funded by the FPT ELI-IST PortDial and SpeDial projects, grant number 611396 and 296170 respectively.