Using Lexical Knowledge to Evaluate the Novelty of Rules Mined from Text

Sugato Basu, Raymond J. Mooney, Krupakar V. Pasupuleti, Joydeep Ghosh

Presented by Joseph Schlecht
Problem Description

• Modern data-mining techniques discover large number of relationships (rules)
  – Antecedent $\rightarrow$ Consequent
• Few may actually be of interest
  – CS job hunting: SQL $\rightarrow$ database
• How do we find rules that are interesting and novel?
• Notice this is subjective
Problem Formalization

• Authors consider text mining
  – Rules consist of words in natural language
• Use WordNet and define semantic distance between two words
• Novelty is defined w.r.t the semantic distance between words in the antecedent and consequent of a rule
Semantic Distance

Given words $w_i$ and $w_j$,

$$d(w_i, w_j) = Dist(P(w_i, w_j)) + K \times \text{Dir}(P(w_i, w_j))$$

- $Dist(p)$ is the distance along path $p$
  - Weighted by relation type (15 in WordNet)
- $Dir(p)$ is the number of directional changes on $p$
  - Defined 3 directions according to relation type
- $K$ is a chosen constant
# Weight and Direction Info

<table>
<thead>
<tr>
<th>Relation</th>
<th>Weight</th>
<th>Direction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synonym, Attribute, Pertainym, Similar</td>
<td>0.5</td>
<td>Horizontal</td>
</tr>
<tr>
<td>Antonym</td>
<td>2.5</td>
<td>Horizontal</td>
</tr>
<tr>
<td>Hypernym, (Member</td>
<td>Part</td>
<td>Substance), Meronym</td>
</tr>
<tr>
<td>Hyponym, (Member</td>
<td>Part</td>
<td>Substance) Holonym, Cause, Entailment</td>
</tr>
</tbody>
</table>
Novelty

• For each rule, a *score* of novelty is generated

• Let $A = \{\text{set of antecedent words}\}$ and $C = \{\text{set of consequent words}\}$ in a given rule

• For each word $w_i$ in $A$ and $w_j$ in $C$
  – $\text{Score}(w_i, w_j) \leftarrow d(w_i, w_j)$

• Score of rule = average of all $(w_i, w_j)$ scores
Experiment

- Measure success by comparing the heuristic’s results of novelty scoring to humans’
- Used rules generated by DiscoTEX from 9000 Amazon.com book descriptions
- Four random samples of 25 rules were made
- Four groups of humans scored each sample
  - 0.0 (least interesting) to 10.0 (most interesting)
- One set was used as training for the heuristic (to find $K$), the other three were used for experiments
## Results

<table>
<thead>
<tr>
<th></th>
<th>Human-Human Correlation</th>
<th>Heuristic-Human Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Raw</td>
<td>Rank</td>
</tr>
<tr>
<td>Group1</td>
<td>0.350</td>
<td>0.338</td>
</tr>
<tr>
<td>Group2</td>
<td>0.412</td>
<td>0.393</td>
</tr>
<tr>
<td>Group3</td>
<td>0.337</td>
<td>0.339</td>
</tr>
</tbody>
</table>

Raw = Pearson’s Raw Score

Rank = Spearman’s Ranks Score
Results (cont)

Example of rules scored by the heuristic

• High Score (9.5)
  romance love heart $\rightarrow$ midnight

• Medium Score (5.8)
  author romance $\rightarrow$ characters love

• Low Score (1.9)
  Astronomy science $\rightarrow$ space
Discussion

• Humans rarely agreed with each other
• Correlation between heuristic and human was similar to human-human correlation
  – Success, but not too meaningful
• Provided statistical evidence that correlation is unlikely due to random chance
• Future tests would use dataset that had higher human-human correlation