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 → Consequent
- Few may actually be of interest
  - CS job hunting: SQL → 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 ,

$$d(w_i, w_j) = Dist(P(w_i, w_j)) + K * Dir(P(w_i, w_j))$$

- $Dist(p)$ 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

Relation	Weight	Direction
Synonym, Attribute, Pertainym, Similar	0.5	Horizontal
Antonym	2.5	Horizontal
Hypernym, (Member|Part|Substance), Meronym	1.5	Up
Hyponym, (Member|Part|Substance) Holonym, Cause, Entailment	1.5	Down

Novelty

- For each rule, a *score* of novelty is generated
- Let A = {set of antecedent words} and C = {set of consequent words} in a given rule
- For each word w_i in A and w_j in C
  - Score(w_i, w_j) ← $d(w_i, w_j)$
- 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

	Human-Human Correlation		Heuristic-Human Correlation	
	Raw	Rank	Raw	Rank
Group1	0.350	0.338	0.187	0.137
Group2	0.412	0.393	0.386	0.363
Group3	0.337	0.339	0.339	0.338

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 → *midnight*

- Medium Score (5.8)

  author romance → *characters love*

- Low Score (1.9)

  Astronomy science → *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