Extracting semantic relations from unlabeled text

Chandra Prakash Vishal Kumar Gupta

Mentor: Dr. Amitabha Mukerjee

March 21, 2013

Motivation

A 14yo bxr owned by a reputable breeder is being treated for IBD with pred.

Motivation

[A 14yo bxr]$_{ANIMAL}$ owned by [a reputable breeder]$_{HUMAN}$ is being treated for [IBD]$_{DISEASE}$ with [pred]$_{DRUG}$. [4]

Why the problem is hard ?

- Huge amount of data available on the web
- No manual tags or labels available
- Don't know exactly how many types of entities are present
- Presence of many irrelevant relations as noise

Problem Definition

Given a corpus of data of extracted relational tuples of the form $r(a, b)$, clusters the data using their relationship and also determine the best match for a given relation.

Related Work

- TextRunner: identifies relational tuples in one pass of the web [3]
- Semantic Network Extractor: Jointly cluster relation and object string [5]
- Infinite Relational Model [1]

Algorithm Specification

- $P(z^1, ...z^n \mid R^1, R^2, ...R^m)$
- Generative Model

$$P(R^1, R^2, ...R^m, z^1, ...z^n) = \prod_{i=1}^{m} P(R^i \mid z^1, ...z^n) \prod_{j=1}^{n} P(z^j)$$

- $P(z^j)$ is calculated using Chinese Restaurant Process
- $R(i,j) \mid z, \eta(a,b)$ is calculated using Bernoulli Distribution
- Chinese Restaurant Process (CRP) also determines the number of clusters

Output Matrics

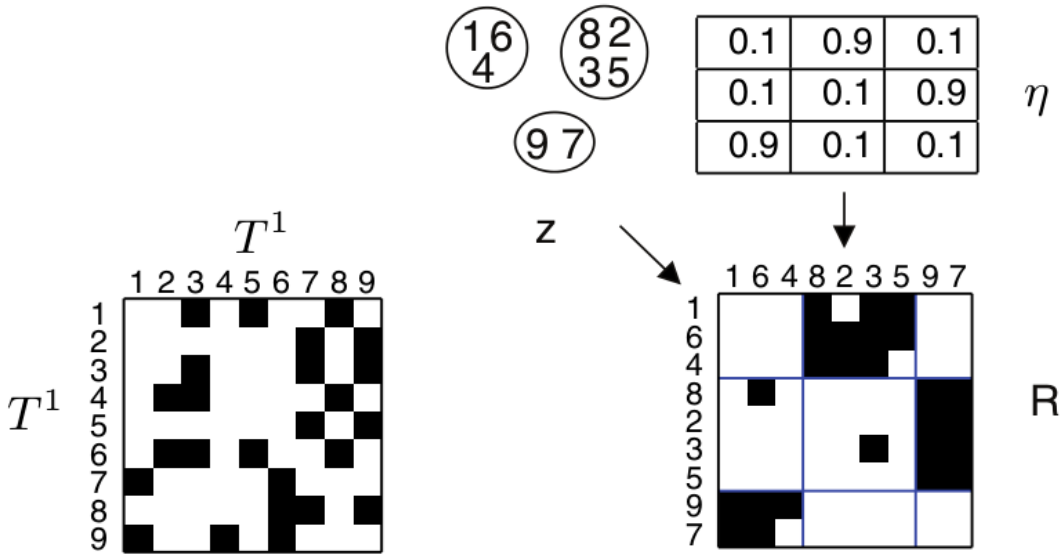

Figure: Output Matrics

References

1. Kemp Charles, Tenenbaum Joshua B, Griffiths Thomas L, Yamada Takeshi, and Ueda Naonori. Learning systems of concepts with an infinite relational model. 21(1):381, 2006.
2. Turney Peter D, Pantel Patrick, et al. From frequency to meaning: Vector space models of semantics. Journal of Artificial Intelligence Research, 37(1):141–188, 2010.
3. Banko Michele. Open information extraction for the web. PhD thesis, University of Washington, 2009.
4. Huang Ruihong and Riloff Ellen. Inducing domain specific semantic class taggers from (almost) nothing. Proceedings of the Association for Computational Linguistics (ACL), 2010.
5. Kok Stanley and Domingos Pedro. Extracting semantic networks from text via relational clustering. Proceedings of ECML, 2008.

Source code and Dataset

- The Source code for IRM is publicly available at *http://www.psy.cmu.edu/~ckemp/code/irm.html*
- The dataset is available at *http://knight.cis.temple.edu/~yates/data/resolver_data.tar.gz*

Questions !!

Formulae Specifications

- Generative Model

$$P(R^1, R^2, ...R^m, z^1, ...z^n) = \prod_{i=1}^{m} P(R^i \mid z^1, ...z^n) \prod_{j=1}^{n} P(z^j)$$

- Generating Clusters (CRP)

$P(z_i = a \mid z_1, ..., z_{i-1}) = \frac{n_a}{i-1+\gamma}$ if $n_a > 0$

$P(z_i = a \mid z_1, ..., z_{i-1}) = \frac{\gamma}{i-1+\gamma}$ if a is a new cluster

- Generating Relations from clusters

$z \mid \gamma \sim CRP(\gamma)$

$\eta(a, b) \mid \beta \sim Beta(\beta, \beta)$

$R(i, j) \mid z, \eta \sim Bernoulli(\eta(z_i, z_j))$

- Inference

$$P(R \mid z) = \prod_{a,b \epsilon N} \frac{Beta(m(a, b) + \beta), Beta(\bar{m}(a, b) + \beta)}{Beta(\beta, \beta)}$$

Semantic Network Extractor

---

function $SNE(S_r, S_x, S_y, R)$
  inputs: S_r, set of relation symbols
    S_x, set of object symbols that appear as first arguments
    S_y, set of object symbols that appear as second arguments
    R, ground $r(x, y)$ atoms formed from the symbols in S_r, S_x, and S_y
  output: a semantic network, $\{(\gamma_r, \gamma_x, \gamma_y) \in \Gamma_r \times \Gamma_x \times \Gamma_y : (\gamma_r, \gamma_x, \gamma_y)$ contains at least one true ground atom$\}$
for each $i \in \{r, x, y\}$
  $\Gamma_i \leftarrow unitClusters(S_i)$
$mergeOccurred \leftarrow true$
while $mergeOccurred$
  $mergeOccurred \leftarrow false$
  for each $i \in \{r, x, y\}$
    $CandidateMerges \leftarrow \emptyset$
    for each $(\gamma, \gamma') \in \Gamma_i \times \Gamma_i$
      $\Delta P \leftarrow$ change in $P(\{\Gamma_r, \Gamma_x, \Gamma_y\}|R)$ if γ, γ' are merged
      if $\Delta P > 0$, $CandidateMerges \leftarrow CandidateMerges \cup \{(\gamma, \gamma')\}$
    sort $CandidateMerges$ in descending order of ΔP
    $MergedClusters \leftarrow \emptyset$
    for each $(\gamma, \gamma') \in CandidateMerges$
      if $\gamma \notin MergedClusters$ and $\gamma' \notin MergedClusters$
        $\Gamma_i \leftarrow (\Gamma_i \setminus \{\gamma, \gamma'\}) \cup \{\gamma \cup \gamma'\}$
        $MergedClusters \leftarrow MergedClusters \cup \{\gamma\} \cup \{\gamma'\}$
        $mergedOccurred \leftarrow true$
return $\{(\gamma_r, \gamma_x, \gamma_y) \in \Gamma_r \times \Gamma_x \times \Gamma_y : (\gamma_r, \gamma_x, \gamma_y)$ contains at least one true ground atom$\}$

Thank You