Predicting Cognitively Salient Modifiers of the Constitutive Parts of Concepts

Gerhard Kremer  Marco Baroni

CIMeC, University of Trento, Italy

July 15, 2010
Describing a Concept...

“Dog”

- has 4 paws
- has a tail
- barks
Describing a Concept...

“Dog”

- has 4 paws
- has a tail
- barks

- vs. -

- has a heart
- can see
Topic and Focus

Feature Norms (e.g. McRae et al.’s)

Concept representations – used in simulations of cognitive tasks
**Topic and Focus**

**Feature Norms (e.g. McRae et al.’s)**

Concept representations – used in simulations of cognitive tasks

**Efforts on extracting such descriptions**

... using text corpora

(getting norms without experiments; better models based on more data)
**Topic and Focus**

**Feature Norms (e.g. McRae et al.’s)**

Concept representations – used in simulations of cognitive tasks

**Efforts on extracting such descriptions**

... using text corpora

(getting norms without experiments;
更好的 models based on more data)

**(New) focus here**

**Composite** part properties (adj modifier + noun) of concepts,

e.g. **rabbit**: **long** ears
Approach

- **Aim:**
  Extract cognitively salient modifiers for given concept–part pairs

- **Idea:**
  Create ranked list based on corpus frequencies and select 5 highest ranked modifiers

- **Resource:**
  WaCky web corpus

- **Evaluation** against feature production norms
Rank List Methods

1. Modifier–Part pair frequencies ("contextless")
2. Log-Likelihood ratios of frequencies
3. Frequencies of modifier–part pairs in concept context
   \[ part ? (20 \text{ sent.}) \text{concept} (20 \text{ sent.}) \text{part} ? \]
4. Summed log-rescaled frequencies
5. Productwise combination of frequencies
Example:

Concept “Bear” With Part “Fur”

<table>
<thead>
<tr>
<th>rank</th>
<th>freq</th>
<th>modifier</th>
<th>rank</th>
<th>freq</th>
<th>modifier</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>507</td>
<td>thick</td>
<td>1</td>
<td>16</td>
<td>thick</td>
</tr>
<tr>
<td>2</td>
<td>209</td>
<td>dense</td>
<td>2</td>
<td>14</td>
<td>white</td>
</tr>
<tr>
<td>3</td>
<td>204</td>
<td>soft</td>
<td>3</td>
<td>11</td>
<td>small</td>
</tr>
<tr>
<td>4</td>
<td>185</td>
<td>black</td>
<td>4</td>
<td>11</td>
<td>soft</td>
</tr>
<tr>
<td>5</td>
<td>175</td>
<td>long</td>
<td>5</td>
<td>9</td>
<td>dense</td>
</tr>
</tbody>
</table>
Performance (ITA)

- parts in concept context (freq)
- contextless (freq)
- combination productwise (freqs)
- baseline: averaged random guessing
Plausibility Judgements (GER)

Setting

- Top 5 candidates of best method (productwise combination)
- “The part of a concept is modifier.”
- Plausible/unlikely to be used in concept explanation?

Evaluation

... for those concept–modifier–part triples with acceptance ≥ 75%
Performance Based on Plausibility Ratings (GER)

- **Recall** vs **Precision** plot
- Red line: combination productwise (freqs)
- Purple line: baseline: averaged random guessing

Axes:
- X-axis: Recall
- Y-axis: Precision

Values range from 0.0 to 1.0 on both axes.
Discussion

Automatic corpus-based extraction

... works best when combining in-context and contextless list

... performs similarly well across languages

... works comparably well based on both production and perception
Further Work

Extension

- Include numerals
- Decide if modifier necessary for specific part

Evaluation

Filter unlikely modifiers (more production data, judgements)

Next

- Salient parts (as preceding step)
- Extract other relation types
... thank you.