Speaker-Generated and Corpus-Generated Features

Poesio, Baroni, Murphy, Barbu, Lombardi (Trento), Almuhareb (KAST, Riyadh), Vinson, Vigliocco (UC London)

Overview

- Objectives
- Speaker-Generated Features
  - published norms
  - cognitive plausibility
  - feature typologies
- Corpus-Generated Features
  - flat word spaces
  - cognitive plausibility
  - semi-supervised feature extraction
  - unsupervised feature extraction and classification

Objectives

- Searching for parallels among views on concepts and their features:
  - Psychology: Rosch, Miller
  - Linguistics: pretty much everyone, including Fillmore, Pustejovsky
  - Neuroscience:
    - Warrington and others: sensory/functional dichotomy
    - Caramazza and others: living/non-living dichotomy
    - Martin and others: finer grained (living/non-living neighbours)
    - Pulvermüller: embodied experience

Why not just ask people?
Speaker-Generated Features

- Recent published concept/feature norms:
  - McRae et al:
    - 541 concepts, 30 informants each, 7000 pairs
    - asked for physical (perceptual), functional properties (used how, where, when, what for), what category and other facts (e.g. where from)
  - Garrard et al:
    - 50 concepts, 20 informants each, 1600 pairs
    - asked what a concept is ..., has ... and can ...
  - Vinson and Vigliocco
    - 169 (nouny) concepts, 20 informants each, 1000 pairs
    - same task as McRae et al
    - asked not to give associative features
  - Compared to Latent Semantic Analysis:
    - Superior category clustering
    - Superior predictor of speeded error production during picture naming
    - Superior predictor of word interference to picture naming
    - Superior predictor of priming in word/no-word decision task

SGF Overlaps

- **motorcycle**
  - Vinson & Vigliocco:
    - wheel, motor, loud, vehicle, wheel, fast, handle, ride, transport, bike, human, danger, noise, seat, brake, drive, fun, gas, machine, object, open, small, travel, wind
  - Garrard et al:
    - vehicle, wheel, fast, handlebar, light, seat, make a noise, tank, metal, unstable, tyre, coloured, sidecar, indicator, pannier, pedal, speedometer, manoeuvrable, race, brakes, stop, move, engine, petrol, economical, gears
  - McRae et al:
    - wheels, 2_wheels, dangerous, engine, fast, helmets, Harley_Davidson, loud, 1_or_2_people, vehicle, leather, transportation, 2_people, fun, Hell's_Angels, gasoline
  - Mutual correlation of ranks ranges from 0.4 to 0.7

Typologies of SGFs

<table>
<thead>
<tr>
<th>Vinson Vigliocco</th>
<th>McRae et al</th>
<th>Garrard et al</th>
<th>for example</th>
</tr>
</thead>
<tbody>
<tr>
<td>visual colour</td>
<td>visual-motion</td>
<td>visual-surface</td>
<td>sensory</td>
</tr>
<tr>
<td>visual motion</td>
<td>visual-parts</td>
<td></td>
<td>brown</td>
</tr>
<tr>
<td>visual surface</td>
<td>smell</td>
<td></td>
<td>gallop</td>
</tr>
<tr>
<td>visual parts</td>
<td>sound</td>
<td></td>
<td>tony</td>
</tr>
<tr>
<td>other perceptual</td>
<td>touch</td>
<td></td>
<td>bump</td>
</tr>
<tr>
<td></td>
<td>taste</td>
<td></td>
<td>roar</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>bristly</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>rather not know</td>
</tr>
<tr>
<td>functional</td>
<td></td>
<td>functional</td>
<td>ride</td>
</tr>
<tr>
<td>motoric</td>
<td></td>
<td></td>
<td>mount</td>
</tr>
<tr>
<td>other</td>
<td>encyclopaedic</td>
<td>encyclopaedic</td>
<td>from Middle-East</td>
</tr>
<tr>
<td></td>
<td>categorising</td>
<td></td>
<td>animal</td>
</tr>
</tbody>
</table>
Corpus-Generated Features

- **Word Space model of meaning**
  - Meaning of words defined by *set of contexts* in which word occurs
  - Similarity of words represented as *geometric distance* among *context vectors*
  - “You should tell a word by the company it keeps” (Firth 1957)

<table>
<thead>
<tr>
<th></th>
<th>leash</th>
<th>walk</th>
<th>run</th>
<th>owner</th>
<th>pet</th>
</tr>
</thead>
<tbody>
<tr>
<td>dog</td>
<td>3</td>
<td>5</td>
<td>2</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>cat</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>lion</td>
<td>0</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>light</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>bark</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>car</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>0</td>
</tr>
</tbody>
</table>

Cognitive Plausibility of CGFs?

- **Word Space model on TOEFL task** (Rapp 2004)
  - Target: *levied*
  - Candidates: *believed, requested, imposed, correlated*
  - Average foreign test taker: 64.5%
  - Macquarie University staff (Rapp 2004):
    - Average of 5 non-natives: 86.75%
    - Average of 5 natives: 97.75%
  - Best reported WSM results (Rapp 2003): *92.5%*

The Word Space World is Flat

- some nearest neighbours of *motorcycle*
  - motor: *component*
  - car: *co-hyponym*
  - diesel: *component?*
  - to race: *proper function*
  - van: *co-hyponym*
  - bmw: *hyponym*
  - to park: *proper function*
  - vehicle: *hyponym*
  - engine: *component*
  - to steal: *qualia?*
- Impossible to distinguish in this word space model
CGFs from Hand Crafted Patterns

- Hearst (1992) automatic ontology extraction
  - vehicles *such as* motorcycles

- Abdulrahmen & Poesio: extract attributes and values
  - e.g. *the seat of the* motorcycle *is* broken; *the* motorcycle *has* two wheels
  - massive corpora: Google fragments; enWac 2 billion words of UK English
  - some rubbish
    - bicycle: safety, electric, recumbent, modern, exercise, motorised, carrera, folding, mountain, tandem, stationary, motor, road, racing, brand, name, old, american, first, utility, new, touring, giant, antique, ...
  - feature rank correlation of 0.4 with Vinson and Vigliocco data
  - concept distance correlation of 0.7 with Vinson and Vigliocco data

CGFs by Pattern Discovery

- Procedure
  - Extract list of potential concept + pattern + feature tuples
  - Rank concept + feature pairs on the basis of number of distinct tuples in which they occur
  - Assign type to concept + feature pair based on analysis of shared parts in patterns that connect them

- Intuition:
  - frequent patterns could simply be (part of) fixed phrases
  - true semantic relations are likely to be expressed by a variety of different superficial patterns
  - year: *year of the tiger, *year of some tigers, *tigers have years ...
  - tail: tail of the tiger, tail of some tigers, tigers have JJ tails, tiger with its tail ...

Typed features: motorcycle

- property
  - ride
  - rider
  - vehicle
  - moped
  - road
  - park
  - scooter
  - car
  - insurance
  - bike
- type
  - verb _ concept
  - noun on concept
  - noun as concept
  - noun for concept
  - concept on noun
  - verb _ concept
  - noun up concept
  - noun as concept
  - noun for concept
  - noun out concept

Typed features: tiger

- property
  - jungle
  - cat
  - species
  - stripe
  - animal
  - to maul
  - habitat
  - lion
  - tame
  - zoo
- type
  - concept in noun
  - noun as concept
  - noun as concept
  - noun as concept
  - verb by concept
  - concept in noun
  - noun as concept
  - verb _ concept
  - concept in noun
Typed features: book

property  type
 to read verb _ concept
 author concept by noun
 to write verb _ concept
 reader concept for noun
 chapter noun in concept
 library concept in noun
 publish verb _ concept
 reading noun from concept
 publisher concept from noun
 review noun on concept

Cluster of Features = Concept

- **fruit**: it is a fruit, it is eaten, it is tasted, it is sliced, it is a flavour, it is used for juice, it is in bowls, it ripens, it is peeled, it is picked
- **animal**: it is an animal, it is killed, it is fed, it is bred, it is a mammal, it is in cages, it is a species, it eats stuff, it is in zoos, it is rescued
- **illness**: it is a disease, treatments have a function for it, it causes stuff, it is pain, it is cured, it is a condition, it is common, it is an infection, it has something to do with dying, it is an ailment
- **creator**: they are employed, they create stuff, they are asked, they are artists, they are in studios, they build stuff, they are commissioned stuff, cameras have a function for them, they are hired, they sell stuff

The End