Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures

Written by
Alexander Budanitsky
Graeme Hirst
Retold by
Keith Alcock

Definitions

- Semantic relatedness
  - General term involving many relationships
    - car-wheel (meronymy)
    - hot-cold (antonymy)
    - pencil-paper (functional)
    - penguin-Antarctica (association)
- Semantic similarity
  - More specific term involving likeness
    - bank-trust company (synonymy)
- Distance
  - Inverse of either one
    - reldist(x)=semantic relatedness^{-1}(x)
    - simdist(x)=semantic similarity^{-1}(x)

Evaluation

- Theoretical examination
  - Coarse filter
- Comparison with human judgment
  - Lack of data
- Performance in NLP applications
  - Many different applications (with potentially conflicting results)
    - Word sense disambiguation
    - Discourse structure
    - Text summarization and annotation
    - Information extraction and retrieval
    - Automatic indexing
    - Automatic correction of word errors in text

Equation: Hirst— St-Onge

$$\mathrm{rel}_{\mathrm{HS}}(c_1, c_2) = C - \text{path length} - k \times d$$

c_1, c_2 : synsets

d : number of changes of direction in the path

C : constant

k : constant

$$\mathrm{rel}_{\mathrm{HS}}(c_1, c_2) = k_1 - \mathrm{len}(c_1, c_2) - k_2 \, \mathrm{dirChanges}(c_1, c_2)$$

Equation: Leacock— Chodorow

$$\mathrm{sim}_{\mathrm{LC}}(c_1, c_2) = -\log\left(\frac{\mathrm{len}(c_1, c_2)}{2D}\right)$$

c_1, c_2 : synsets

D : overall depth of the taxonomy

$$\mathrm{sim}_{\mathrm{LC}}(c_1, c_2) = \log(2) - \log(\mathrm{len}(c_1, c_2)) + \log(D)$$

Equation: Resnik

$$\mathrm{sim}_{\mathrm{R}}(c_1, c_2) = -\log(\mathrm{p}(\mathrm{lso}(c_1, c_2)))$$

c_1, c_2 : synsets

$\mathrm{p}(x)$: probability of encountering x

in a specific corpus

$\mathrm{lso}(x, y)$: lowest super - ordinate

Equation: Jiang— Conrath

$$\text{dist}_{\text{JC}}(c_1, c_2) = 2\log(\text{p}(\text{lso}(c_1, c_2))) - (\log(\text{p}(c_1)) + \log(\text{p}(c_2)))$$

c_1, c_2 : synsets

$\text{p}(x)$: probability of encountering x

in a specific corpus

$\text{lso}(x, y)$: lowest super – ordinate

$$\text{simdist}_{\text{JC}}(c_1, c_2) = \log\left(\frac{\text{p}^2(\text{lso}(c_1, c_2))}{\text{p}(c_1)\,\text{p}(c_2)}\right)$$

Equation: Lin

$$\mathrm{sim}_{\mathrm{L}}(c_1, c_2) = \frac{2 \times \log(\mathrm{p}(\mathrm{lso}(c_1, c_2)))}{\log(\mathrm{p}(c_1)) + \log(\mathrm{p}(c_2))}$$

c_1, c_2 : synsets

$\mathrm{p}(x)$: probability of encountering x in a specific corpus

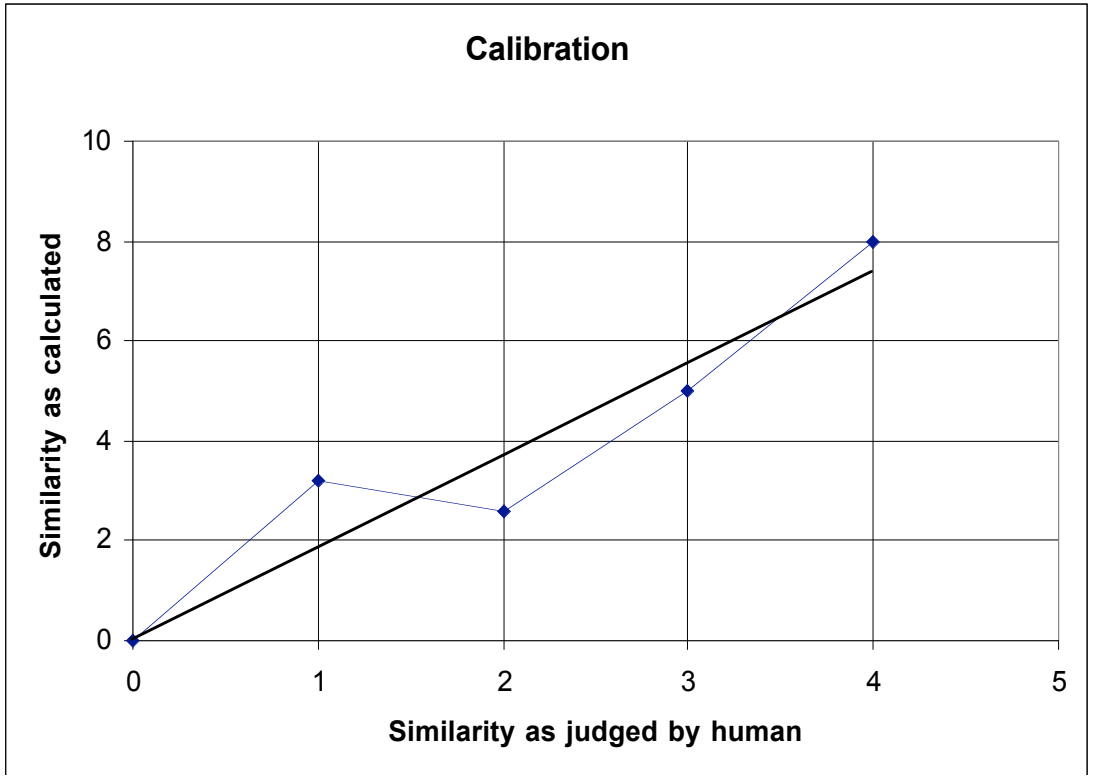
$\mathrm{lso}(x)$: lowest super-ordinate

$$\mathrm{sim}_{\mathrm{L}}(c_1, c_2) = \frac{\log(\mathrm{p}^2(\mathrm{lso}(c_1, c_2)))}{\log(\mathrm{p}(c_1)\,\mathrm{p}(c_2))}$$

Calibration: Step 1

- Rubenstein & Goodenough (1965)
  - Humans judged semantic synonymy
    - 51 subjects
    - 65 pairs of words
    - 0 to 4 scale
- Miller & Charles (1991)
  - Different humans, subset of words
    - 38 subjects
    - 30 pairs of words
    - 10 low (0-1), 10 medium (1-3), 10 high (3-4)

Calibration: Step 2



Testing: Simulation

- Malapropism
  - Real-word spelling error
  - *He lived on a diary farm.
  - When after insertion, deletion, or transposition of intended letters, a real word results
- Material
  - 500 articles from *Wall Street Journal* corpus
  - 1 in 200 words replaced with spelling variation
  - 1408 malapropisms

Testing: Assumptions

- The writer's intended word will be semantically related to **nearby** words
- A malapropism is unlikely to be semantically related to **nearby** words
- An intended word that is not related is unlikely to have a spelling variation that is related to **nearby** words

Testing: Suspicion

- Suspect is unrelated to other nearby words
- True suspect is a malapropism

$$P_S = \text{Precision}_S = \frac{\text{number of true suspects}}{\text{number of suspects}}$$

$$R_S = \text{Recall}_S = \frac{\text{number of true suspects}}{\text{number of malapropisms in text}}$$

$$\text{F-measure}_S\big|_{\beta=1} = \left.\frac{(\beta^2+1)P_S R_S}{\beta^2 P_S + R_S}\right|_{\beta=1} = \frac{2P_S R_S}{P_S + R_S}$$

Testing: Detection

- Alarm is a spelling variation related to nearby words
- True alarm is a malapropism that has been detected

$$P_D = \text{Precision}_D = \frac{\text{number of true alarms}}{\text{number of alarms}}$$

$$R_D = \text{Recall}_D = \frac{\text{number of true alarms}}{\text{number of malapropisms in text}}$$

$$\text{F-measure}_D\big|_{\beta=1} = \frac{(\beta^2+1)P_D R_D}{\beta^2 P_D + R_D}\bigg|_{\beta=1} = \frac{2P_D R_D}{P_D + R_D}$$

Results: Suspicion

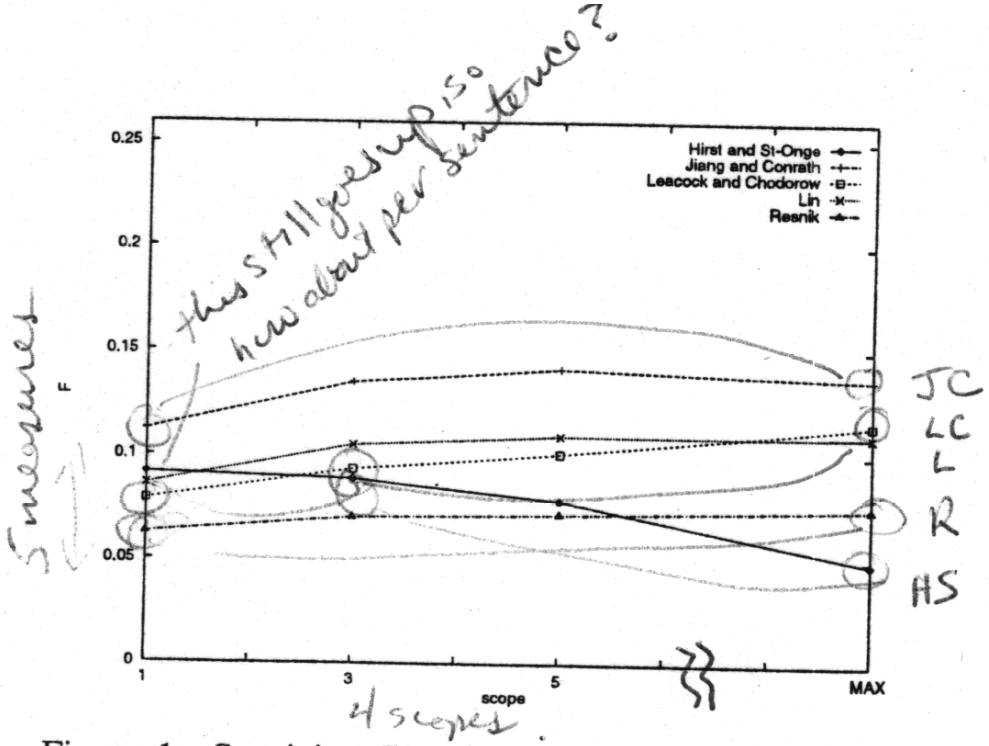


Figure 1: Suspicion F-measure (F_S), by measure and scope.

Results: Detection

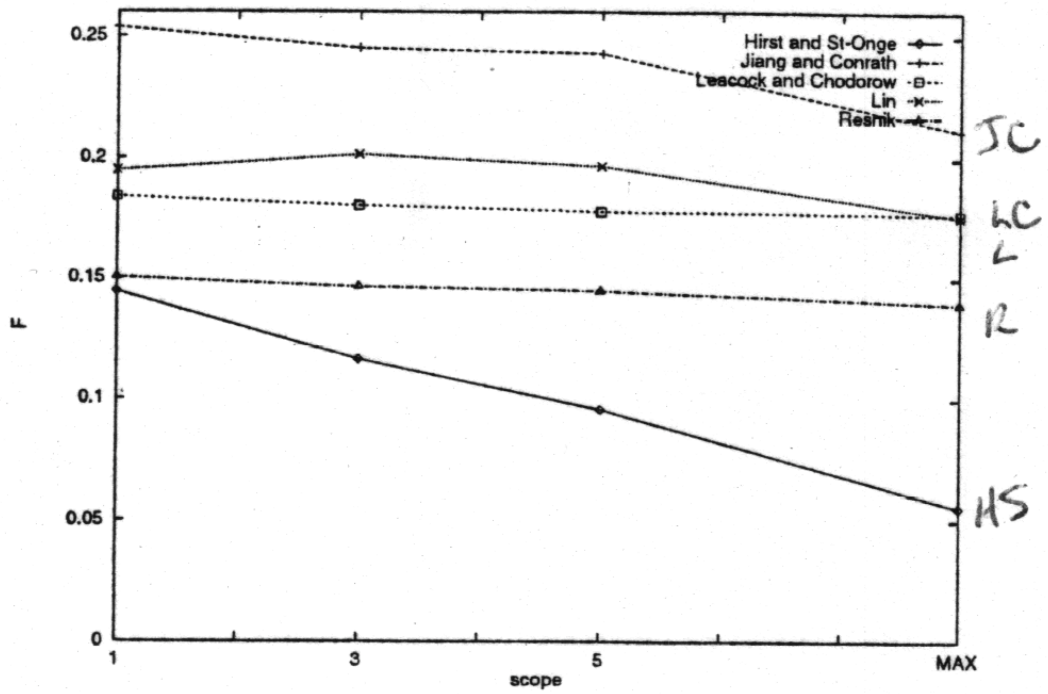


Figure 2: Detection F-measure (F_D), by measure and scope.

Conclusion

- Measures are significantly different
  - $simdist_{JC}$ on single paragraph is best
    - 18% precision
    - 50% recall
  - rel_{HS} is worst
- Relatedness doesn't outperform similarity
  - WordNet gives obscure senses the same prominence as more frequent senses

Discussion

- Calibration of relatedness with similarity data
- Calibration point inaccurate
- Substitution errors untested
- Semantic bias in human typing errors not addressed
- Binary threshold not best choice
- Frequency on synset, word, or word sense