
NLP

Text Similarity

Thesaurus-based Word Similarity Methods

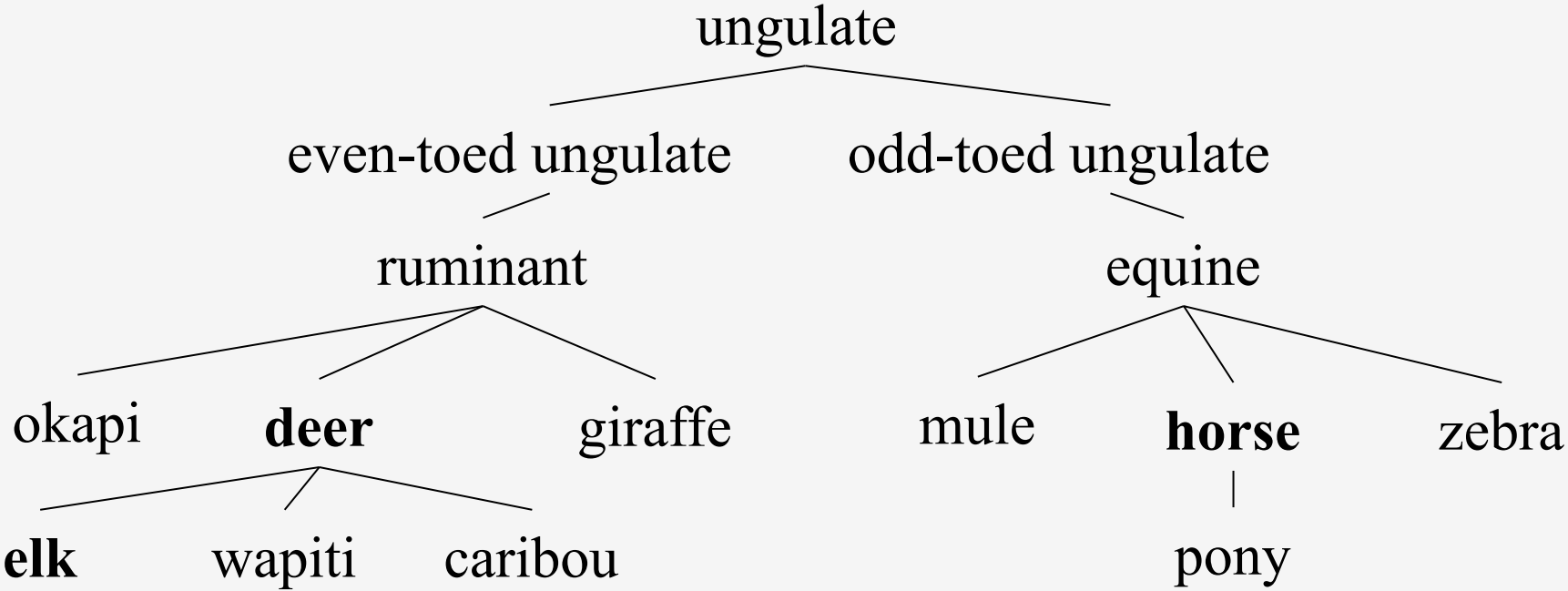
Quiz

- Which pair of words exhibits the greatest similarity?
  - 1. Deer-elk
  - 2. Deer-horse
  - 3. Deer-mouse
  - 4. Deer-roof

Quiz Answer

- Which pair of words exhibits the greatest similarity?
  - **1. Deer-elk**
  - 2. Deer-horse
  - 3. Deer-mouse
  - 4. Deer-roof
- Why?
- Remember the Wordnet tree:

Remember Wordnet



Path Similarity

- Version 1
  - Sim (v,w) = - pathlength (v,w)
- Version 2
  - Sim (v,w) = - log pathlength (v,w)

Problems With This Approach

- There may be no tree for the specific domain or language
- A specific word (e.g., a term or a proper noun) may not be in any tree
- IS-A (hypernym) edges are not all equally apart in similarity space

Path Similarity Between Two Words

- Version 3 (Philip Resnik)

  Sim (v,w) = – log *P*(LCS(v,w))

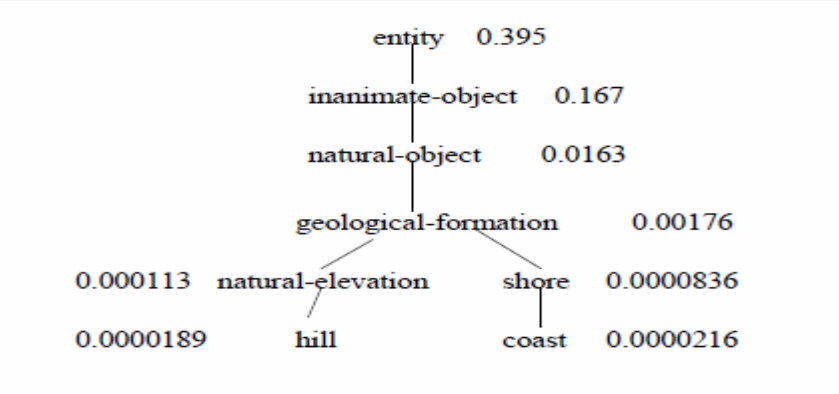
  where LCS = lowest common subsumer, e.g.

  ungulate for deer and horse

  deer for deer and elk

Information Content

- Version 4 (Dekang Lin)
  - Wordnet augmented with probabilities (Lin 1998)
  - IC(c) = -log P(c)
  - Sim (v,w) = 2 x log *P*(LCS(v,w)) / (log *P*(v) + log *P*(w))



$$\text{sim}(\text{Hill, Coast}) = \frac{2 \times \log P(\text{Geological-Formation})}{\log P(\text{Hill}) + \log P(\text{Coast})}$$

$= 0.59$

Wordnet Similarity Software

- WordNet::Similarity (Perl)
  - http://www.d.umn.edu/~tpederse/similarity.html
- NLTK (Python)
  - http://www.nltk.org

```
>>> dog.lin_similarity(cat, brown_ic)
0.879
>>> dog.lin_similarity(elephant, brown_ic)
0.531
>>> dog.lin_similarity(elk, brown_ic)
0.475
```


NLP