WordNet

LING83600
Kyle Gorman
Graduate Center, City University of New York

Outline

- Definitions of key lexico-semantic relations
- Introduction to WordNet
- WordNet-based word similarity

Definitions

Word senses

Word senses refer to distinct meanings of a word.

- *homonyms*: unrelated senses that have the same spelling; e.g., *bank* (a financial institution) vs. *bank* (of a river)
- *polysemes*: related-but-distinct senses; e.g., *Bank* of America vs. blood *bank*
- *metonyms*: metaphorical stand-ins; e.g., *Washington* used to mean "the US federal government" or "the executive branch"
- *homographs*: distinct or related senses that have the same spelling but different pronunciations; e.g., *bass* (fish) vs. *bass* guitar

Sense relationships

The *is-a* relationship, as in *a dog is-a mammal*, has two directional forms:

- *hypernyms*: more general, less specific; e.g., *mammal* is a hypernym of *dog*
- *hyponyms*: less general, more specific; e.g., *dog* is a hyponym of *mammal*

Other commonly-used relations include *meronyms* (part-to-whole) and *holonyms* (whole-to-part).

WordNet

Introducing WordNet

The best-known resource for computational lexical semantics is WordNet (Fellbaum, 1998), developed by George A. Miller and Christiane Fellbaum at Princeton. The original WordNet was for English, but there are now hundreds of "WordNets" for different languages (though size and quality vary enormously). NLTK (Bird et al., 2009) includes an easy-to-use Python interface for (English) WordNet; similar interfaces are available for other languages (e.g., see Pedersen et al. 2004 for Perl).

WordNet contents

WordNet consists of

- A simple rule-based lemmatizer,

```
>>> from nltk.stem.wordnet import WordNetLemmatizer
>>> lemmatizer = WordNetLemmatizer()
>>> lemmatizer.lemmatize("wolves")
'wolf'
```

- a one-to-many mapping from lemmas (lexical nouns, verbs, adjectives, and adverbs) to *synsets*,
- a tree-structured ontology in which each (non-)terminal node is a synset, and
- methods to compute the similarity between two synsets.

Synsets

A *synset* (“synonym set”) is a group of one or more lemmata (and their associated senses) assumed to be synonymous.

```
>>> from nltk.corpus import wordnet
>>> wordnet.synsets("dog")
[Synset('dog.n.01'), Synset('frump.n.01'),
 Synset('dog.n.03'), Synset('cad.n.01'),
 Synset('frank.n.02'), Synset('pawl.n.01'),
 Synset('andiron.n.01'), Synset('chase.v.01')]
```

Synset methods

The synsets are returned in order of frequency (in some corpus). Let us select the first synset, which gives the common-sense sense of *dog* (a domesticated animal).

```
>>> dog = wordnet.synsets("dog")[0]
>>> dog.definition()
'a member of the genus Canis...'
>>> dog.examples()
['the dog barked all night']
```

Hypernyms

Synset objects can enumerate their (immediate) hypernyms and hyponyms.

```
>>> dog.hyponyms()
[Synset('basenji.n.01'), Synset('corgi.n.01'),
 Synset('cur.n.01'), Synset('dalmatian.n.02'),
 Synset('great_pyrenees.n.01'), Synset('griffon.n.02'),
 Synset('hunting_dog.n.01'), Synset('lapdog.n.01'),
 Synset('leonberg.n.01'), Synset('mexican_hairless.n.01'),
 Synset('newfoundland.n.01'), Synset('pooch.n.01'), ...]
>>> dog.hypernyms()
[Synset('canine.n.02'), Synset('domestic_animal.n.01')]
```

Closure

Or one can enumerate all their hypernyms and hyponyms.

```
>>> dog.closure(lambda s: s.hypernyms())
[Synset('canine.n.02'), Synset('domestic_animal.n.01'),
 Synset('carnivore.n.01'), Synset('animal.n.01'),
 Synset('placental.n.01'), Synset('organism.n.01'),
 Synset('mammal.n.01'), Synset('living_thing.n.01'),
 Synset('vertebrate.n.01'), Synset('whole.n.02'),
 Synset('chordate.n.01'), Synset('object.n.01'),
 Synset('physical_entity.n.01'), Synset('entity.n.01')]
```

WordNet word similarity

The WordNet ontology can also be used as a way to measure/quantify the similarity between pairs of words (or synsets).

Path similarity

The simplest form of this is known as *path similarity*, defined as $PS(c_1, c_2) : C \times C \longrightarrow \mathbb{N}$ where C is the set of senses and $\mathbb{N}$ is a counting number. Let A be the minimum number of "arcs" one must traverse between the two synsets. Then,

$$PS(c_1, c_2) = \frac{1}{1 + A}.$$

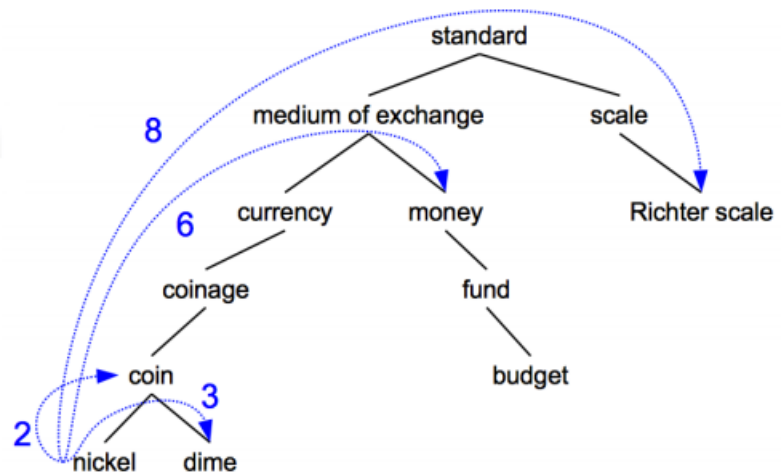

Figure: Example of WordNet path similarity (from slides by D. Jurafsky).

Path similarity examples

```
>>> dog = wordnet.synsets("dog")[0]
>>> cat = wordnet.synsets("cat")[0]
>>> spider = wordnet.synsets("spider")[0]
>>> dime = wordnet.synsets("dime")[0]
>>> round(dog.path_similarity(cat), 2)
0.2
>>> round(dog.path_similarity(spider), 2)
0.14
>>> round(dog.path_similarity(dime), 2)
0.06
>>> round(spider.path_similarity(dime), 2)
0.05
```

Alternatives to path similarity

Path similarity is somewhat limited however, because it assumes a uniform semantic distance between parent/child nodes in the ontology. This motivates alternative definitions based on the notion of *information content* of subtrees of the ontology.

Information content

Let $P(c) : C \times [0, 1]$ be a random variable denoting the probability that a randomly-sampled word in a corpus is an instance of, or hypernym of, sense c; this can be computed using maximum likelihood estimation, given a corpus and an ontology. Then,

$$IC(c) = -\log P(c).$$

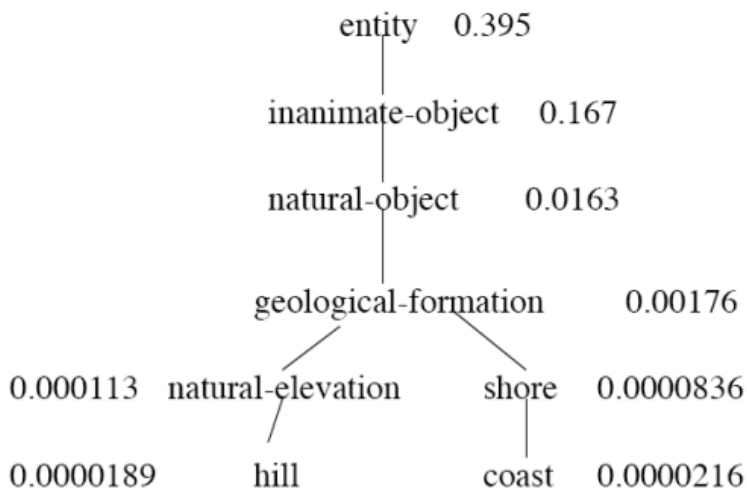

Figure: Example of WordNet information content (from slides by M. Capuat).

Information content in NLTK

NLTK's WordNet wrapper provides several pre-computed tables of information content.

```
>>> from nltk.corpus import wordnet_ic
>>> brown_ic = wordnet_ic.ic("ic-brown.dat")
```

Resnik similarity

Let $LCS(c_1, c_2)$ be the *least common subsumer* of synsets c_1, c_2, i.e., the lowest node that dominates both. Then, Resnik (1995) similarity $RP(c_1, c_2) : C \times C \rightarrow \mathbb{R}$ is given by

$$RP(c_2, c_2) = ICS(LCS(c_1, c_2))$$

Resnik similarity examples

```
>>> round(dog.res_similarity(cat, brown_ic), 2)
7.91
>>> round(dog.res_similarity(spider, brown_ic), 2)
4.95
>>> round(dog.res_similarity(dime, brown_ic), 2)
-0.0
>>> round(spider.res_similarity(dime, brown_ic), 2)
-0.0
```

Alternatives to Resnik similarity

Alternatives to Resnik's method—but still based on information content—include the methods of:

- Wu and Palmer (1994),
- Jiang and Conrath (1997),
- Leacock and Chodorow (1998), and
- Lin (1998).

All of these are supported by NLTK. Budanitsky and Hirst (2006) is just one of many studies that attempts to compare and evaluate these methods.

Alternatives to WordNet

WordNet is not the only lexical ontology out there. For instance, researchers in medical informatics often use MeSH ("Medical Subject Headings"), an ontology developed by the National Institute of Health's National Library of Medicine (NLM). Tens of thousands of medical articles have been tagged using the MeSH system. Domain-specific ontology development is a common task for linguists working in tech.

References I

S. Bird, E. Klein, and E. Loper. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, 2009.

A. Budanitsky and G. Hirst. Evaluating WordNet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47, 2006.

C. Fellbaum. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

J. J. Jiang and D. W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings on International Conference on Research in Computational Linguistics*, pages 19–33, 1997.

C. Leacock and M. S. Chodorow. Combining local context and WordNet similarity for word sense identification. In C. Fellbaum, editor, *WordNet: An Electronic Lexical Database*, pages 265–283. MIT Press, 1998.

D. Lin. An information-theoretic definition of similarity. In *International Conference on Machine Learning*, pages 296–304, 1998.

References II

T. Pedersen, S. Patwardhan, and J. Michelizzi. WordNet::Similarity: measuring the relatedness of concepts. In *Demonstration Papers at HLT-NAACL 2004*, pages 38–41, 2004.

P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pages 448–453, 1995.

Z. Wu and M. Palmer. Verb semantics and lexical selection. In *32nd Annual Meeting of the Association for Computational Linguistics*, pages 133–138, 1994.