

WordNet: Design, Contents, Limitations

Christiane Fellbaum
Cognitive Science Laboratory
Princeton University

What is WordNet?

- A large lexical database, or “electronic dictionary”
- Covers most English nouns, verbs, adjectives, adverbs
- Electronic format makes it amenable to automatic manipulation
- Used in many applications (document retrieval and sorting, machine translation,...)

What's so special about WordNet?

- Traditional paper dictionaries are organized alphabetically, so words that are grouped together (on the same page) are unrelated
- WordNet is organized by meaning, so words in close proximity are related
- Users can browse WordNet and find words that are meaningfully related to their queries (like in a thesaurus)

Motivation and Inspiration

- AI models for Knowledge Representation (e.g., Collins and Quillian 1986)
- People have knowledge about tens of thousands of concepts
- Knowledge encompasses lexical (word) knowledge and encyclopedic (world) knowledge
- Effortless storage and retrieval
- Concepts are represented in human semantic memory efficiently and economically

Knowledge about concepts

Minimal knowledge about concepts:

- X is a kind of Y
- X has part Y
- an X Ys
- X is Y/has property Y

Knowledge about concepts

These types of knowledge encode (much of) the meanings of nouns (entities), verbs (events), and adjectives (properties)

The statements can be expressed as relations among nouns, verbs, and adjectives

Empirical question:

Can most (all?) concepts referring to entities/events/properties be linked by means of a few relations?

Basic design of WordNet

Semantic relations interlink lexicalized
concepts

Result: large semantic network (graph)

Concepts, Words, Relations

- Lexicon: labeling of concepts \Rightarrow words
- Humans label salient concepts
- Concepts differ in systematic ways: contrasts and similarities
- Consistent differences = relations
- If a few relations suffice to interlink most labeled concepts, then labeling is systematic (lexicon is regular)

Two kinds of semantic relations

- Lexical (word-word) relations
- Conceptual (concept-concept) relations

Most important lexical relation: synonymy

WordNet groups (roughly) synonymous,
denotationally equivalent, words into unordered
sets of synonyms (synsets)

{hit, beat}

{big, large}

{queue, line}

Each synset expresses a distinct concept.

Currently, WordNet contains appr. 117,000 synsets

Synonymy

Note: Natural Languages map concepts and words
many-to many

Synonymy: One concept can be expressed by many
words

{car, auto, automobile}

{close, shut}

Most WordNet synsets contain >1 member

Another lexical relation: Antonymy

Psycholinguistic experiments show strong bi-directional “clang” association between pairs of antonymous adjectives:

big-little

large-small

wet-dry

hot-cold

etc.

Antonymy

This strong association is often word-specific

big-little (not: *big-small*)

large-small (not: *large-little*)

hard-easy (not: *hard-simple*)

Note: pairs share selectional restrictions

Antonymy

Of course, semantic opposition/contrast holds among **all** members of antonymous synsets

(but it's not as salient as the bond between the lexical pairs)

{big, large}-{little, small}

{hard, difficult}-{easy, simple}

This synset-synset relation is conceptual

Conceptual-semantic relations

Synsets, the nodes of the network, are interrelated via conceptual-semantic relations

Whence the relations?

Inspired by

--Aristotle's *Metaphysics*

--AI models and some experiments

--traditional lexicographers' definitions

--psycholinguistic evidence

Traditional dictionary definitions reflect relations

an X is a kind/type of (a) Y

X and Z are Ys

(super/subordinate relation)

an X is a part of (a) Y

a Y has an X (part/whole relation)

X: not Y (antonymy/contrast)

Psycholinguistic Evidence for Relations

- Word association norms (robust!)
- Patterns of loss and sparing in aphasia

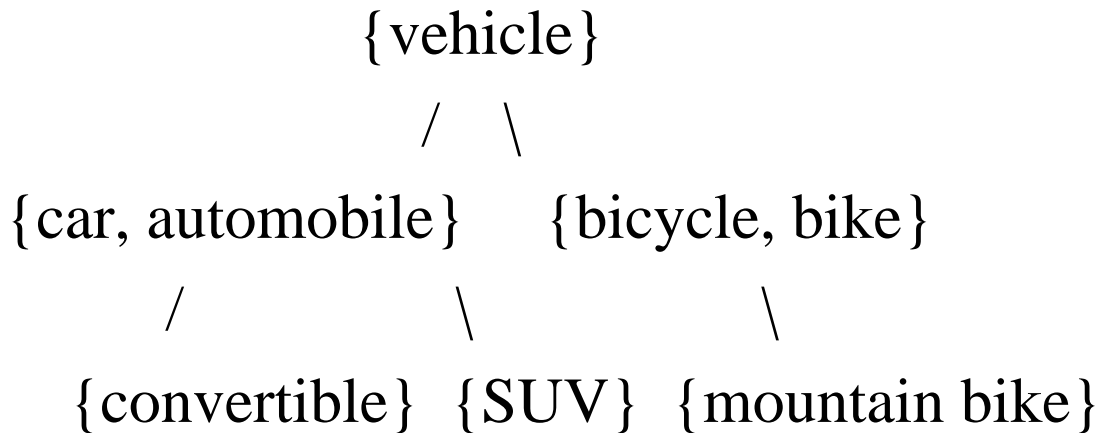
Knowledge about concepts

Reaction time experiments indicated that knowledge about concepts is stored at superordinate nodes and inherited downward

- X is a kind of Y
- X has part Y

Hyponymy relates noun synsets

Denotes more/less general concepts:



“A car is a kind of vehicle”

“The class of vehicles contains cars, SUVs, mountain bikes”

Hyponymy

Transitivity:

A car is a kind of vehicle

An SUV is a kind of car

=> An SUV is a kind of vehicle

Instances vs. types

Instances are hyponyms but not types/kinds:

A lake is a kind of body of water

**Lake Ontario is a kind of a lake*

Lake Ontario is an instance of a lake

Instances are expressed by proper nouns

Instances are leaves (terminal nodes) without hyponyms

Types vs. Roles

WordNet does not (yet) distinguish types from roles

{poodle}-*{dog}* (a “type” relation)

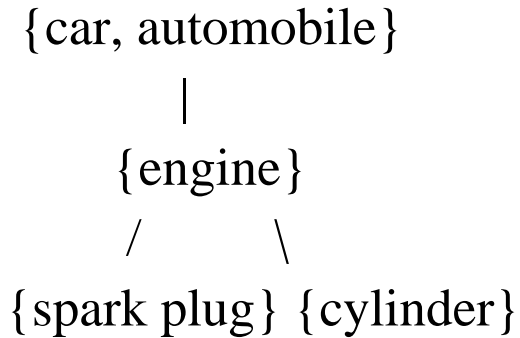
{poodle}-*{pet}* (a “role” relation)

poodle is a type of *dog*, but not a type of *pet*

poodle can (but must not) play the “role” of *pet*--the relation between *poodle* and *pet* is defeasible

Meronymy among noun synsets

Meronymy/holonymy (part/whole)



“An engine has spark plugs”

“Spark plus and cylinders are parts of an engine”

Meronymy

Inheritance:

A car has an engine

An engine has spark plugs

=> A car has spark plugs

Three Kinds of Meronymy

- Meronymy is a polysemous relations (Chaffin et al. 1991)
- WordNet's meronymy is underspecified
- Distinguish 3 kinds only

Three Kinds of Meronymy

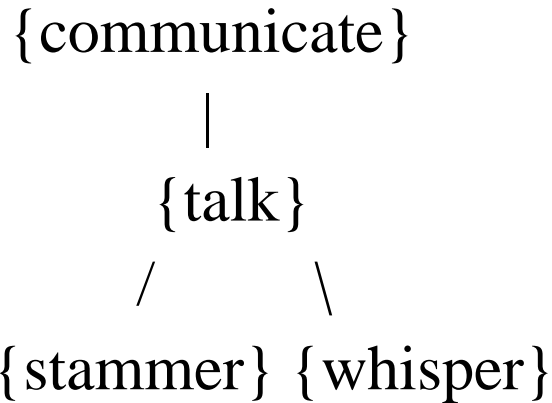
Proper parts (separable, in principle):

table-leg, finger-hand, petal-blossom

Substances: *oxygen-water*

Groups/Members: *forest-tree, student-class*

“Manner” relation links verb synsets



“To talk is to communicate in some manner”

“To whisper is to talk in some manner”

“Manner” relation

Similar to noun hyponymy:

- most frequently encoded (manner relation structures the verb lexicon)
- builds tree structures (more shallow than noun hierarchies)
- inheritance

Other relations among verb synset

Semantics of events (verbs) are very different
from semantics of entities (nouns)

Relation refer to temporal properties of events

--partial and complete overlap of two events

--prior or posterior events

Entailment relations among verb synsets

- Temporal inclusion:
 - co-extensiveness (manner): *whisper-talk*
 - proper inclusion: *walk-step*
- No temporal inclusion:
 - backward presupposition: *forget-know*
 - cause: *show-see, raise-rise*

Limitations of WordNet

WordNet encodes mostly paradigmatic relations:
related terms are substitutable for each other (same
POS)

Few syntagmatic relations, where related words co-
occur in a small window of context (different
POS)

Limitations of WordNet

As a result, WordNet consists of four largely unconnected WordNets:

--nouns

--verbs

--adjectives

--adverbs

Exceptions: “topical/domain” links (*car-traffic*)
and derivational relations (*direct-director-directive*)

Limitations of WordNet

This is inconsistent with early AI models of KR,
psychological evidence of human semantic
organization (association norms)

It also limits WordNet's potential for NLP

Current work to increase density (WNPlus)

Network is (too) sparsely connected

- Why not encode/find links among ALL synsets?
- Recent work with Boyd-Graber, Osherson, Schapire (2006)

Overcome WN's shortcomings:

- overcome sparseness of connections
- both intra- and intercategory
- attach weights to arcs
- direct arcs

Add to WordNet

- Cross-POS links (*traffic, congested, stop*)
- More relations: *Holland-tulip, sweater-wool, axe-tree, buy-shop, red-flame,...*
- Relations need not be labeled
- Arcs are directed: *dollar-green/*green-dollar*
- Arcs are weighted

Evocation

“How strongly does concept A evoke concept B in people’s minds?”

NOT: similarity (*pear-apple*)

association (*dress-button*)

Procedure

Identify 1K “core” synsets

- highly frequent (in British National Corpus)
- highly salient
- 500 N, 250 V, 250 Adj

Experiment

- Collected 120K judgments for randomly chosen synsets (subset of 1K)
- Designed interface for ratings
- Wrote rating manual
- Strength of evocation ranged from 0-100
- Five anchor points

Human Ratings

- Raters were warned not to use personal, idiosyncratic evocations (*dog-grandmother*)
- Avoid evocation of word form (rhyme, same initial letter, etc.)
- Raters were tested for consistency with themselves and agreement with others

Results

- Median correlation on test set for the 24 annotators was .72
- lowest correlation was .64
- Average correlation with themselves: .70

Results

- 67% of evocations were rated “zero” (expected)
- High consistency for zero ratings

Comparison with other similarity measures

- Lesk (overlap of words in glosses)
- Paths in WN (verbs, nouns)
- Latent semantic indexing (strings not necessarily senses)

- Only weak correlation of our results with each measure!
- Evocation captures something similarity doesn't!

Next task

- Completely fill in net of 1K synsets
- Too much for human ratings
- Machine learning!

Machine Learning

Input features:

- major similarity measures
- context vectors from British National Corpus (tagged for POS; eliminates some polysemy)

Learn evocations

- Apply boosting techniques (Schapire)
- Divide data into 5 categories of evocation strengths (0 is its own category)
- 80% training data, 20% testing
- Results: incorrect assignments ~25%
- More work is ongoing...

WordNet and Ontology

- WordNet is a **lexical** ontology
- WN has been mapped onto formal ontologies
- E.g., SUMO and MILO (Adam Pease)
- Formal ontology may be the interlingua for crosslinguistic wordnets

Where to find WordNet

Freely downloadable:

<http://wordnet.princeton.edu/>

Database, browser, documentation

Information on related work, FQA

Interfaces, browsers, APIs, ...

Thank you!

For questions, comments, and papers:

fellbaum@princeton.edu