UNIVERSITÉ DE LYON

ICOM INSTITUT DE LA COMMUNICATION

INFO STAT

Using knowledge in text mining

Master Data Mining

Julien Velcin

Using knowledge

- In this short introduction, using knowledge means mainly integrating an ontology, but there are **many other resources**: dictionaries, lexicons, but also websites such as wikipedia !
- Ontologies are the most general formalism for describing data objects
- Very popular for the Semantic Web (e.g. OWL)
- Ontologies can be of various complexity

Ontologies and the Semantic Web

- Origin in philosophy: *"study of the nature of being, existence and reality as such, as well as the basic categories of being and their relations"* [Wikipedia]
- In CS, represent a domain and use it to **reason** about the **objects** and the **relations** between these objects
- Today, the ontologies are central for the Semantic Web
- Most of the SW standards (XML, RDF, OWL) are concerned with some level of ontological representation of the knowledge

Which elements represent an ontology?

- An ontology typically consists of the following elements:

  Instances – the basic or "ground level" objects

  Classes – sets, collections, or types of objects

  Attributes – properties, features, characteristics, or parameters that objects can have and share

  Relations – ways that objects can be related to one another

WordNet

- Thesaurus has a main function to connect different surface word forms with the same meaning into one sense (synonyms)
- Using thesauri + the hypernym-hyponym relation leads to a more compact representation of the knowledge
- The most commonly used general thesaurus is **WordNet** which exists in many other languages (e.g. EuroWordNet, BalkanNet) http://www.illc.uva.nl/EuroWordNet
- WordNet group at Princeton http://wordnet.cs.princeton.edu
- Try it online: http://www.wordnet-online.com (v. 2.0)
- WordNet 2.1 (Windows), WordNet 3.0 (UNIX-based)

Different kinds of relations

- WN addresses different kinds of relations among word surface forms and their senses:
  - **Hypernym-Hyponym:** taxonymic "is a" relation
    e.g.: breakfast and meal
  - **HasPart-PartOf:** wholes and parts
    e.g.: table and leg
  - **Antonym:** opposites
    e.g.: leader and follower
  - **Synonym** (embedded in **synsets**): different form, same meaning
    e.g.: singer, vocalist

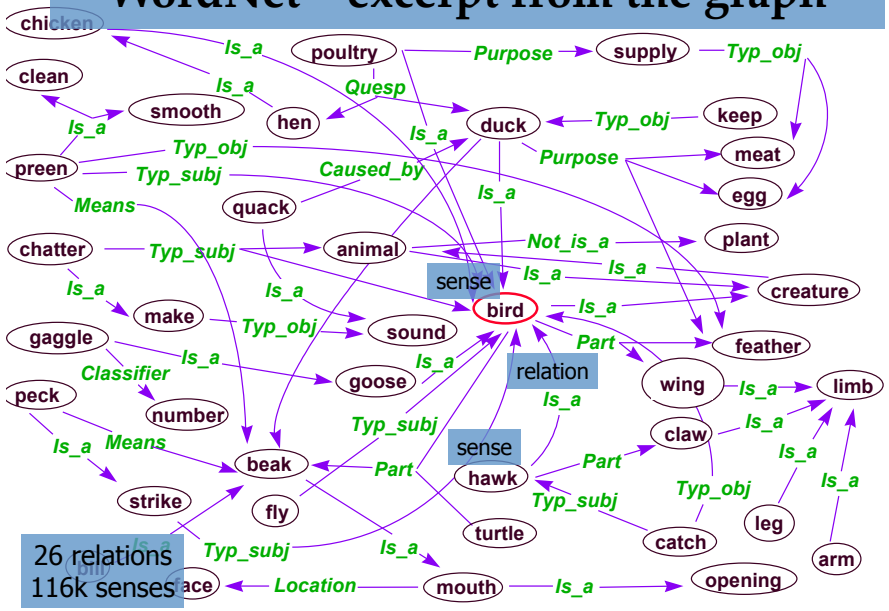
WordNet: a database of lexical relations

- WordNet is the most well developed and widely used lexical database for English (147,249 words in v2.1)
  - 4 databases: nouns, verbs, adjectives, and adverbs
  - Noun network = 80%
  - Maximum depth of the noun hierarchy is 16
- Each database consists from sense entries (**synsets**) – each sense consists from a set of synonyms

Category	Unique Forms	Number of Senses
Noun	94474	116317
Verb	10319	22066
Adjective	20170	29881
Adverb	4546	5677

E.g.:
- musician, instrumentalist, player
- person, individual, someone
- life form, organism, being

WordNet – excerpt from the graph



WordNet 2.0

http://www.wordnet-online.com

Noun bird has 5 senses

1. **bird** - warm-blooded egg-laying vertebrates characterized by feathers and forelimbs modified as wings
   - $\rightarrow_1$ is a kind of vertebrate, craniate
   - $\rightarrow_1$ is a member of Aves, class Aves; flock
   - $\rightarrow_1$ has parts:
     - beak, bill, neb, nib, pecker; furcula; feather, plume, plumage; wing; pennon, pinion; bird's foot; uropygium; air sac; uropygial gland, preen gland; syrinx; bird, fowl
   - $\rightarrow_1$ has particulars:
     - dickeybird, dickey-bird, dickybird, dicky-bird; cock; hen; nester; night bird; bird of passage; protoavis; archaeopteryx, archeopteryx, Archaeopteryx lithographica; Sinornis; Ibero-mesornis; archaeornis; ratite, ratite bird, flightless bird; carinate, carinate bird, flying bird; passerine, passeriform bird; nonpasserine bird; bird of prey, raptor, raptorial bird; gallinaceous bird, gallinacean; parrot; cuculiform bird; coraciiform bird; apodiform bird; caprimulgiform bird; piciform bird; trogon; aquatic bird; twitterer

   Derived form: *verb* bird$_1$
2. **bird, fowl** - the flesh of a bird or fowl (wild or domestic) used as food
   - $\rightarrow_2$ is a kind of meat
   - $\rightarrow_2$ is a part of bird
   - $\rightarrow_2$ has parts:
     - wishbone, wishing bone; drumstick; second joint, thigh; wing; giblet, giblets; oyster; parson's nose, pope's nose; dark meat
   - $\rightarrow_2$ has particulars: poultry; wildfowl
3. dame, doll, wench, skirt, chick, **bird** - informal terms for a (young) woman
   - $\rightarrow_3$ is a kind of girl, miss, missy, young lady, young woman, fille
4. boo, hoot, Bronx cheer, hiss, raspberry, razzing, snort, **bird** - a cry or noise made to express displeasure or contempt
   - $\rightarrow_4$ is a kind of cry, outcry, call, yell, shout, vociferation
5. shuttlecock, **bird**, birdie, shuttle - badminton equipment consisting of a ball of cork or rubber with a crown of feathers
   - $\rightarrow_5$ is a kind of badminton equipment

Verb bird has 1 sense

1. **bird**, birdwatch - watch and study birds in their natural habitat
   - $\rightarrow_1$ is one way to observe

   Derived forms: *noun* birder$_1$, *noun* bird$_1$

   Sample sentence:
   In the summer they like to go out and bird

Semantic measures

- Estimating the **semantic relatedness** between two words using the *various* relations of WN
- Relatedness ≠ similarity !
- Relatedness of concepts ≈ synsets
- Can be highly usefull, especially for WSD

Computing semantic relatedness

Some notations [Budanitsky and Hirst, 05]

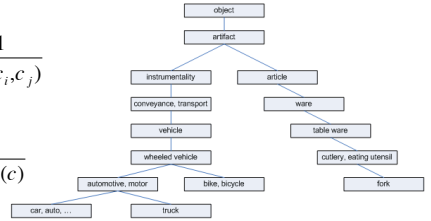
- len(c_i,c_j) = length of the shortest path
- depth(c_i) = len(root, c_i)
- lso(c_i,c_j) = lowest super-ordinate
- rel(c_i,c_j) = semantic relatedness between concepts
- rel(w_i,w_j) = semantic relatedness between words

$$rel(wi,wj) = \max_{c_p \in s(w_i), c_q \in s(w_j)} \left[rel(c_p, c_q) \right]$$

Some measures

- Path length: $sim_{PL}(c_i,c_j) = \frac{1}{len(c_i,c_j)}$
- Leacock & Chodorow:

$$sim_{LC}(c_i,c_j) = -\log \frac{len(c_i,c_j)}{2 \times \max_{c \in WordNet} depth(c)}$$



- Wu & Palmer:

$$sim_{WP}(c_i,c_j) = \frac{2 \times depth(lso(c_i,c_j))}{len(c_i,lso(c_i,c_j)) + len(c_j,lso(c_i,c_j)) + 2 \times depth(lso(c_i,c_j))}$$

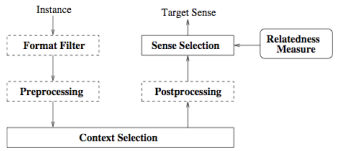
- Resnik:

$$sim_R(c_i,c_j) = -\log p(lso(c_i,c_j))$$

etc.

An application to WSD

- Using WordNet for WSD [Patwardhan et al., 2007]
- Evaluation in SemEval-2007 (task 1)
- System UMND1:



- Calculating a score for each sense t_i:

$$score(t_i) = \sum_{j=1}^{2n} \max_{k=1 to W_j} (rel(t_i, w_{jk}))$$

Material on the Internet

- Tutorial at EDBT'06
  rene-witte.net/system/files/IntroductionToTextMining.pdf
- WordNet: An Electronic Lexical Database
  http://mitpress.mit.edu/book-home.tcl?isbn=026206197X
- [Budanitsky and Hirst, 05]
  http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.106.7617&rep=rep1&type=pdf
- Ted Pedersen's website
  http://marimba.d.umn.edu/cgi-bin/similarity/similarity.cgi