Foundations of Language Science and Technology

Semantics 3

Manfred Pinkal
Saarland University

Overview

- Semantic Processing - Introduction
- Logic-based meaning representation and processing: Truth-conditional interpretation, entailment, deduction
- Word Meaning: Lexical-semantic resources, ontologies, similarity-based approaches
  - Informal overview
  - Semantic Relations, WordNet
  - Corpus-based semantic similarity measures
  - Comparison
- Semantic Composition: Composing sentence and text meaning from word meaning
- Textual Entailment and Inference

Representing Conceptual Content

- Semantic Decomposition: Representing conceptual content through:
  - Feature sets:
    bachelor --> [+male, +adult, - married]
  - Structured representations, made up of a small set of „primitive concepts“:
    kill(X,Y) := CAUSE(X, BECOME(NOT(ALIVE(Y))))
- A difficulty:
  - There is hardly a limited set of features or primitives that permit to semantically model the full lexicon.

Semantic Relations

Text: *... Volvo sells trucks ...*
Question: *Which companies sell motor vehicles?*

- To find out whether the text contains a piece of information which is relevant to answer the question we need to know something about the meaning of the words *company, truck, and motor vehicle*.
- We need no definition or full description of their meaning however: It is sufficient to know that trucks are motor vehicles (i.e., truck stands in a hyponymy or subconcept relation to motor vehicle), and that Volvo is a company.
- With a small inventory of semantic relations, we can capture a lot of lexical meaning structure which is needed to infer relevant information from natural-languag expressions.

Basic Semantic Relations

- Synonymy: car - auto - automobile - machine
- Hypernomy / hyponomy, the sub-/superconcept relation:
  - *car - truck, dog - animal, kill - murder*
- Meronymy, the part-of relation, and its inverse relation, holonymy, with three (well-motivated) sub-relations:

  Physical Part - Whole relation: *branch - tree*

  Member - Group relation: *tree - forest*

  Substance - Object relation: *wood - tree*
- Antonymy, a general super-concept for opposition/ contrast, comprising
  - Contrast (or antonymy in the narrower sense): *good - bad, expensive - cheap*
  - Complementarity: *man - woman, married - single*
  - Converse/ inverse relation: *buy - sell, ancestor - descendant*

  (according to Lyons 1979)

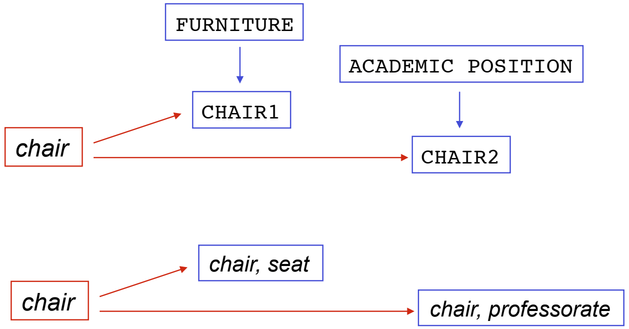
Lexical Ambiguity

- The mapping between phonological/ orthographic words and meanings is one-to-many: One (phonological or orthographic) word can be associated with several concepts or word senses.
- Ambiguity between unrelated senses: Homonymy
  - *bank: river bank - financial institution*
- Ambiguity between semantically related concepts: Polysemy
  - *bank: financial institution - blood bank*
  - *case: carton - case: suitcase - case: pillowcase*
  - *to serve a meal - to serve as president*
- Homonyms are typically represented as different lexical entries, cases of polysemy as single entries with multiple sense descriptions.
- There are systematic patterns of polysemy:
  - *rabbit, dear, chicken*: animal - meat - fur
  - *fast car - fast road - fast driver*

Thesaurus, Ontology, WordNet

- Large dictionaries/ lexical databases structured as a hierarchy by hypernymy/hyponymy relation are called thesaurus (e.g., Roget's thesaurus for English)
- Semantic relations are strictly speaking not relations between words, but rather between concepts or word senses.
- The task of representing meaning relations for the lexicon is a two-fold one:
  - Describe meaning relations between concepts: This is typically the done by **ontologies (see below)**.
  - Specify mappings from words to concepts, which associate words with their possible word senses.
- The currently most important lexical-semantic resource, WordNet, does both in one, using the concept of a "synset".

Thesaurus, Ontology, WordNet



FURNITURE

ACADEMIC POSITION

CHAIR1

chair

CHAIR2

chair, seat

chair

chair, professorate

WN-Senses/Synsets of *car*

- S: (n) **car**, auto, automobile, machine, motorcar
- S: (n) **car**, railcar, railway car, railroad car
- S: (n) **car**, gondola
- S: (n) **car**, elevator car
- S: (n) cable car, **car**

WN: Synsets, glosses, examples

- car
  - { car, auto, automobile, machine, motorcar }
  - a motor vehicle with four wheels; usually propelled by an internal combustion engine
  - *"he needs a car to get to work"*

WN: Hyponyms of *motor vehicle*

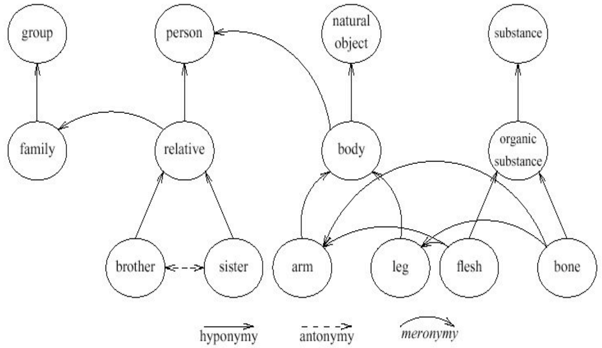
- S: (n) **motor vehicle**, automotive vehicle (a self-propelled wheeled vehicle that does not run on rails)
- ***direct hyponym*** / *full hyponym*
  - S: (n) amphibian, amphibious vehicle (a flat-bottomed motor vehicle that can travel on land or water)
  - S: (n) bloodmobile (a motor vehicle equipped to collect blood donations)
  - S: (n) car, auto, automobile, machine, motorcar (a motor vehicle with four wheels; usually propelled by an internal combustion engine) *"he needs a car to get to work"*
  - S: (n) doodlebug (a small motor vehicle)
  - S: (n) four-wheel drive, 4WD (a motor vehicle with a four-wheel drive transmission system)
  - S: (n) go-kart (a small low motor vehicle with four wheels and an open framework; used for racing)
  - S: (n) golfcart, golf cart (a small motor vehicle in which golfers can ride between shots)
  - S: (n) hearse (a vehicle for carrying a coffin to a church or a cemetery; formerly drawn by horses but now usually a motor vehicle)
  - S: (n) motorcycle, bike (a motor vehicle with two wheels and a strong frame)
  - S: (n) snowplow, snowplough (a vehicle used to push snow from roads)
  - S: (n) truck, motortruck (an automotive vehicle suitable for hauling)

WordNet

- WordNet is kind of an ontology with the additional feature that the word-concept mapping is implicit in the way word senses or concepts are represented:
- As sets of synonymous words („synsets“).
- Synsets directly provide synonymy information, and information about the word-concept mapping: A (orthographic) word has all those senses/ synsets as readings, of which it is a member.
- This holds in principle. In practice, no or too few synonyms are typically available for sense distinction. WordNet glosses and examples help to disambiguate.
- WordNet provides information about all semantic relations mentioned so far: Hyponymy, Meronymy, Antonymy, and their inverse relations.

A small fragment of the WN graph

Figure 2. Network representation of three semantic relations among an illustrative variety of lexical concepts



WordNet

- English WordNet is by far the largest currently available lexical-semantic resource:
  - 150.000 lexical items
  - 120.000 synsets
  - 200.000 word-sense pairs
- WordNet is extensively used in many Language technology applications.
- Versions of WordNet currently available for about 45 languages (with large differences in coverage, design, and availability).
- "GermaNet": a German WordNet version with about 100.000 lexical items.

Using WordNet for Information Access Tasks

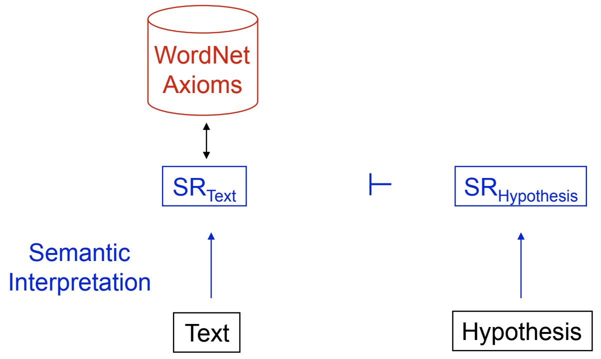
- WordNet is highly useful for both entailment and similarity-based approaches to information access.
- From the logical point of view, WordNet forms a large database of axioms which can be used to support deduction / inference.

WordNet Relations in FOL

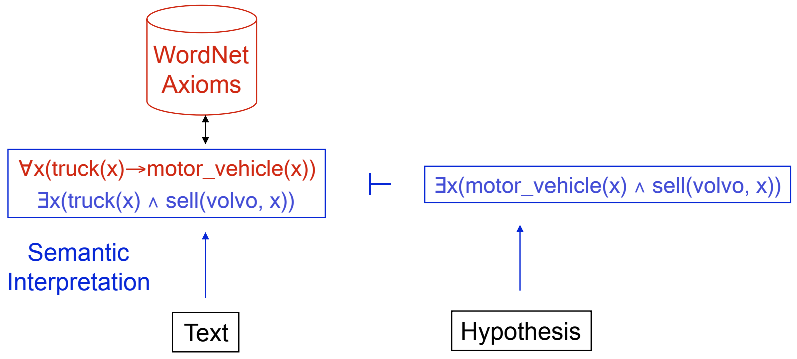
- WordNet relations can be re-written as FOL formulae:

$$\forall x(\text{family}(x)\rightarrow\text{group}(x))$$
$$\forall x(\text{person}(x) \rightarrow \exists y(\text{substance\_m}(y,x) \wedge \text{body}(y))$$
$$\forall x(\text{body}(x) \rightarrow \exists y(\text{part\_m}(y,x) \wedge \text{leg}(y))$$
$$\forall x(\text{body}(x) \rightarrow \exists y(\text{part\_m}(y,x) \wedge \text{arm}(y))$$

WordNet and Logical Entailment



WordNet and Logical Entailment


WordNet Axioms
∀x(truck(x)→motor_vehicle(x))
∃x(truck(x) ∧ sell(volvo, x))
⊢
∃x(motor_vehicle(x) ∧ sell(volvo, x))
Semantic Interpretation
Text
Hypothesis

Description Logic

- Description Logics are fragments of FOL which are tailored to representation with terminological information. Below, WN axioms are given in DL notation - just for illustration.

- ∀x(family(x)→group(x))
  family ⊑ group
  ∀x(relative(x)→person(x))
  relative ⊑ person
  ∀x(person(x) → ∃y(substance_m(y,x) ∧ body(y))
  person ⊑ ∃substance_m.body
  ∀x(body(x) → ∃y(part_m(y,x) ∧ leg(y))
  body ⊑ ∃part_m.leg

Description Logic

- DL reasoning is less expressive, but much more efficient than FOL deduction.
- Trade-off between expressive power and computational complexity
- DL reasoners: FaCT, Racer, Protégé, supporting different reasoning tasks for different DL versions.
- Description Logics form the core or backbone of Semantic Markup Languages (e.g., OWL) and various ontologies.

Ontologies

- An ontology is a shared conceptualization of a domain
- An ontology is a set of definitions in a formal language for terms describing the world (Definition taken from slides of Adam Pease)

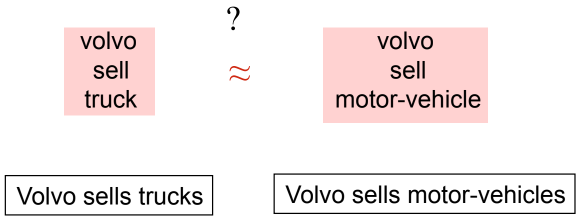
- Another definition: Ontologies are
  - Hierarchical data structures
  - Providing formally rigorous information about concepts and relation
  - Within a specific domain (domain ontologies)
  - Or concepts and relations of foundational, domain-independent relevance (upper ontologies)
- Upper Ontologies:
  - DOLCE, CYC, SUMO
- WordNet is a linguistically motivated and language related upper ontology, called a thesaurus as well as a „language ontology“.

Overview

- Semantic Processing - Introduction
- Logic-based meaning representation and processing: Truth-conditional interpretation, entailment, deduction
- Word Meaning: Lexical-semantic resources, ontologies, similarity-based approaches
  - Informal overview
  - Semantic Relations, WordNet
  - Semantic similarity measures
  - Comparison
- Semantic Composition: Composing sentence and text meaning from word meaning
- Textual Entailment and Inference

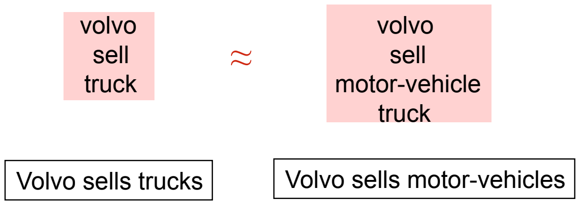
Query Expansion

- WordNet also feeds similarity-based approaches to infomation access:



Query expansion

- WordNet also feeds similarity-based approaches to infomation access:



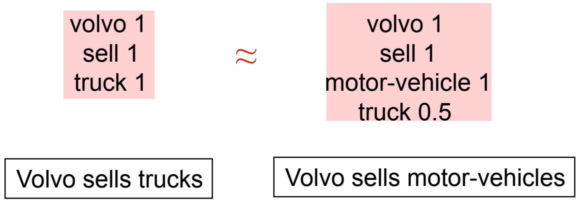
- Query expansion by WordNet synonyms/ hyponyms/ concepts that are „WordNet-related“ in some or the other way

WordNet Similarity

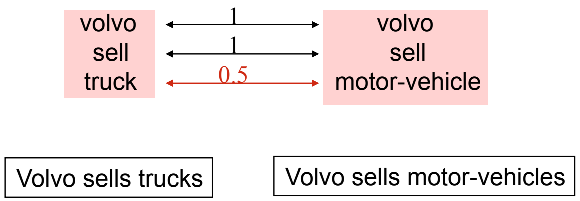
- Based on WordNet, quantitative measures of distance / similarity between two concepts or word senses can be defined:
- A simple distance measure: Path length $dist_{WN} = pathlength(s_1,s_2)$
- A simple similarity measure: Inverse of path length $sim_{WN} = \frac{1}{pathlength(s_1,s_2)}$
- Normalisation by medium total path length.
- More complex corpus-related WN measures are based on the ratio between the informativity of the compared senses and the informativity of their lowest common hypernym. - Informativity of s measured as the negative log value of the probability that a content word in the corpus is a hypomyn of s.

A different use of WordNet

- Use WN similarity for:
  - Thresholds for query expansion
  - Discounting of added concepts
  - Direct computation of T - H similarity



A different use of WordNet



Vector-space models of meaning

d1: About Dolphins

Dolphins are mammals, not fish. They are warm-blooded like humans, and give birth to one baby called a calf at a time. At birth a bottlenose dolphin calf is about 90-130 cms long and will grow to approx. 4 metres, living up to 40 years. They are highly sociable animals, living in pods which are fairly fluid, with dolphins from other pods interacting with each other from time to time.

d2: About Whales

Whales are marine mammals of order Cetacea which are neither dolphins - members, in other words, of the families delphinidae or platanistoidae - nor porpoises. They include the blue whale, the largest animal ever to have lived. Like all mammals, whales breathe air into lungs, are warm-blooded, feed their young milk from mammary glands, and have some (although very little) hair.

d3: About Language Technology

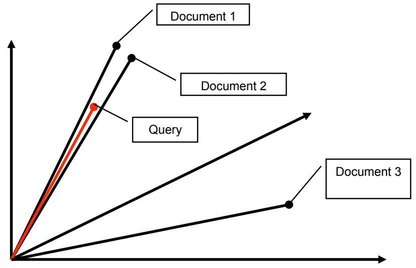
Applied CL focusses on the practical outcome of modelling human language use. The methods, techniques, tools and applications in this area are often subsumed under the term language engineering or (human) language technology. Although existing CL systems are far from achieving human ability, they have numerous possible applications. The goal is to create software products that have some knowledge of human language.

Vector-Space Models in IR

- In Information Retrieval, semantic or informational similarity between documents is (among other things) measured by similar occurrence patterns of words (similar words occuring with similar frequency).
- Document meaning can be approximately represented as vector in the word space. The relevant property is the direction, expressing the proportion of different words, rather than the length which corresponds to the length of (number of words in) the document.
- Similarity between documents can be calculated on the basis of the angle between the vectors d1 and d2. Typically, it is defined as the cosine of the angle between d1 and d2.

Word-Document Matrix

	d1	d2	d3	…
dolphin	3	2	0	…
human	1	0	3	…
language	0	0	5	…
like	1	1	0	…
mammal	1	1	0	
technology	0	0	1	…
warm-blooded	1	1	0	…
whale	0	3	0	…



Document Similarity by Text Frequency

- Standard measure for similarity between word vectors is cosine again:

$$sim_{\cos ine}(\vec{x},\vec{y}) = \frac{\vec{x}\cdot\vec{y}}{|\vec{x}||\vec{y}|} = \frac{\sum_{i=1}^{n} x_i \times y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}$$

- Cosine is 1, if vectors have indentical directions ($cos(0^0)=1$), it is 0, if vectors are orthogonal ($cos(90^0)=1$).
- Our example:

$$\cos(\vec{d}_1,\vec{d}_2) = 0.62$$
$$\cos(\vec{d}_1,\vec{d}_3) = 0.14$$
$$\cos(\vec{d}_2,\vec{d}_3) = 0$$

Are dolphins mammals?

	d1	d2	d3	q
dolphin	3	2	0	1
human	1	0	3	0
language	0	0	5	0
like	1	1	0	0
mammal	1	1	0	1
technology	0	0	1	0
warm-blooded	1	1	0	0
whale	0	3	0	0

$$\cos(\vec{q},\vec{d}_1) = 0.78$$

$$\cos(\vec{q},\vec{d}_2) = 0.53$$

$$\cos(\vec{q},\vec{d}_3) = 0$$

Vector-Space Models of Word Meaning

- The document-word matrix can be also looked at from a different viewpoint:
- Two words are semantically similar, if their distribution pattern over documents is similar.
- We can represent word meaning as a vector in the document space.

One further step:

- Two words are semantically similar, if their co-occurrence patterns with other words are similar.
- Thus, we can compute word meanings directly as vectors in the word space.
- Co-occurrence can be defined on the level of documents, paragraphs, sentences, or context windows of fixed length n (measured in text words, e.g., n = 2, 10, 50).
- Different similarity measures are available, one of the most prominent ones is cosine again.

Word-Word Matrix

	dolphin	human	language	like	technol.	warm-bl.	whale
dolphin	5	1	0	2	0	2	3
human	3	4	5	1	1	1	0
language	0	3	5	0	1	0	0
like	1	1	0	2	0	2	3
technology	0	3	5	0	1	0	0
warm -blooded	5	1	0	2	0	2	3
whale	2	0	0	1	0	1	3