Semantic Theory

Summer 2006

Lexical Semantics

M. Pinkal / A. Koller

Final Exam

- Thu 20.7. 11:00-13:00 (120 min.!)
- Registration for Final Exam: Deadline Thu 6.7. !!!
- Question time, Sample Exam discussion: Tue 18.7.

Structure of this course

- Sentence semantics
- Discourse semantics
- Lexical semantics

The dolphin text

Dolphins are mammals, not fish. They are warm blooded like man, 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.

Sentence Semantics (Predicate Logic)

Dolphins are mammals, not fish. They are warm blooded like man, 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.

Discourse semantics

Dolphins are mammals, not fish. They are warm blooded like man, 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.

Lexical semantics

Dolphins are mammals, not fish. They are warm blooded like man, 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.

Lexical semantics

Dolphins are mammals, not fish. They are warm blooded like man, 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.

Main semantic categories of words

- Function words:
  - Connectives and quantifiers
  - Modal verbs and particles
  - Anaphoric pronouns, articles
  - Degree modifiers, Copula, ...

- Content words
  - Standard one-place predicates: Common nouns, adjectives, (intrans. verbs)
  - Relational concepts with overt argument: In particular verbs, but also nouns, adjectives, prepositions
- Other
  - Named Entities (Persons, institutions, geographic entities, dates)
  - Numbers
  - ...

Challenges in lexical semantics [1]: Size of the lexicon

- Provision of lexical-semantic information is highly labor- and cost-intensive because the lexicon is
  - very large, actually
  - undelimitable
  - heterogeneuos
  - subject to extreme application-dependent variation
- The basic challenge is not efficient processing techniques, but methods for acquisition and organisation.

Challenges in lexical semantics [2]: Ambiguity

- Single words can be multiply ambiguous, in particular in central areas of the lexicon.
- There is no clear boundary for the set of readings of a lexical item, because of meaning extensions and figurative uses (metaphor, metonymy):
  - ➢ *to like Shakespeare, to eat rabbit, to wear rabbit, to grasp an idea*
- There is no clear criterion to distinguish between different uses of one reading and different readings:
  - ➢ *bank (river bank, financial institute)*
  - ➢ *onion (eating onions – growing onions)*

The dolphin again

Challenges in lexical semantics [3]

- The concepts corresponding to single readings of a word are typically multi-layered, consisting of heterogeneous kinds of information:
  - "Propositional" layer
  - Layer of visual (or other sensory) prototypes
  - Stereotypical information
- No sharp boundary between word meaning and other kinds of knowledge.
- No boundary between commonsense and expert knowledge constituting meaning.

Lexical-semantic resources

- Monolingual dictionaries, alpabetically ordered, provide meaning information about the readings of a word informally, in form of synonyms, glosses, typical examples, etc.
  - Oxford English Dictionary
  - Webster's
  - Wahrig /Duden
- A thesaurus presents the lexicon of a language in a hierarchical ordering:
  - Roget's Thesaurus (English, since 1805)
  - Dornseiff's "Deutscher Wortschatz nach Sachgruppen" (German, 1910)

Semantic Relations

- Thesauri provide implicit information about the basic semantic relation of
  - ➢ Hyponymy/Hypernymy (the "ISA relation", e.g., dolphin – mammal)
- There are a number of additional important semantic relations:
  - ➢ Synonymy : case – bag
  - ➢ Meronymy/Holonymy
    - Part – Whole : branch – tree
    - Member – Group: tree – forest
    - Matter – Object: wood – tree
  - ➢ Contrast:
    - Complementarity: boy – girl
    - Antonymy: long – short

Semantic Networks

- A network of semantic relations appears to be a natural way of representing the semantic lexicon of a language.
- However, it is not the words themselves that stand in semantic relations to each other, but rather the concepts corresponding to the different readings of a word.
- There is no 1:1 relation between words and concepts:
  - ➢ The same word can express different concepts (ambiguity)
  - ➢ The same concept can be expressed by different words (synonymy)

WordNet

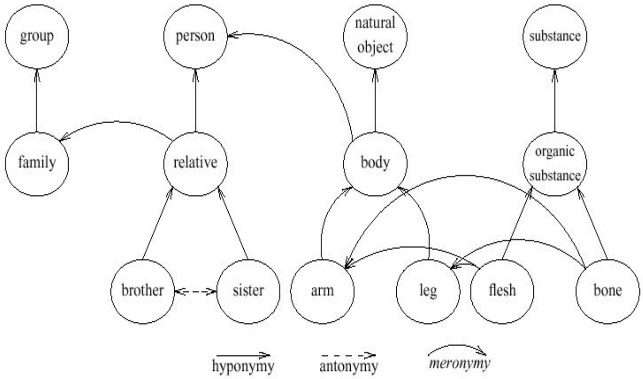
- WordNet is a large lexical-semantic resource, organised as a semantic network.
- Concepts/readings in WordNet are represented by „synsets“: Sets of synonymous words. „Synsets“ form the nodes of the semantic network.
- In cases where no synonyms are available to distinguish readings, WordNet uses glosses.

An example: *case*

- {*case, carton*}
- {*case, bag, suitcase*}
- {*case, pillowcase, slip*}
- {*case, cabinet, console*}
- {*case, casing* (the enclosing frame around a door or window opening)}
- {*case* (a small portable metal container)}

An example

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



WordNet

- English WordNet: about 150.000 lexical items
  - ➢ Web Interface: http://wordnet.princeton.edu/perl/webwn
  - ➢ General Info: http://wordnet.princeton.edu/
- "GermaNet": a German WordNet version with about 90.000 lexical items
- Versions of WordNet for available for about 30 languages
- WordNet consists of different, basically unrelated databases for common nouns, verbs, adjectives (and adverbs)
- ´The respective hierarchies have a number of "unique beginners" each.

Unique Beginners for WordNet Nouns

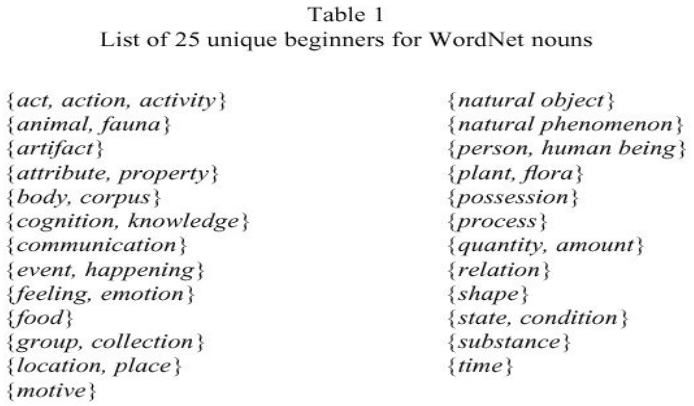


Table 1
List of 25 unique beginners for WordNet nouns

{*act, action, activity*}
{*animal, fauna*}
{*artifact*}
{*attribute, property*}
{*body, corpus*}
{*cognition, knowledge*}
{*communication*}
{*event, happening*}
{*feeling, emotion*}
{*food*}
{*group, collection*}
{*location, place*}
{*motive*}
{*natural object*}
{*natural phenomenon*}
{*person, human being*}
{*plant, flora*}
{*possession*}
{*process*}
{*quantity, amount*}
{*relation*}
{*shape*}
{*state, condition*}
{*substance*}
{*time*}

WordNet: Advantages

- WordNet is big and has very large coverage (concerning both words and word senses (compared to SUMO/MILO)
- WordNet is a highly valuable resource for different kinds of information management and access tasks, e.g.:
  - ➢ query expansion for Information Retrieval
  - ➢ basic inferences via semantic relations, in particular hyponymy / subsumption
- WordNet concepts (in a given language) are linked to natural language expressions in the most direct and natural way.

WordNet: Constraints

- Different parts of WordNet have different granularity for the description of word senses. In general, WordNet is too fine-granular for many purposes.
- There are WordNet versions for a large number of languages, but there is no real multi-lingual WordNet: The different WordNet differ in coverage, format, and availability.
- WordNet focusses on paratactic semantic relations between single words. It does not provide the core lexical information needed for composition:
  - ➢ Predicate-argument structure /Semantic roles.

Ontologies

- An **ontology** is the product of an attempt to formulate an exhaustive and rigorous conceptual scheme about a domain. An ontology is typically a hierarchical data structure containing all the relevant entities and their relationships and rules within that domain (eg. a **domain ontology**). The computer science usage of the term *ontology* is derived from the much older usage of the term ontology in philosophy.

- An ontology which is not tied to a particular problem domain but attempts to describe general entities is known as a **foundation ontology** or **upper ontology**.

  (Wikipedia, the whole article is worth reading)

Ontologies

- In philosophy, **ontology** (from the Greek *ov* = *being* and *λόγος* = *word/speech*) is the most fundamental branch of metaphysics. It studies being or existence as well as the basic categories thereof -- trying to find out what entities and what types of entities exist. Ontology has strong implications for the conceptions of reality.

- Basic Aristotelian categories:
  - ➢ Substance, Quantity, Quality, Relation, Place, Time, Posture, State, Action, and Passion

Ontologies, Overview

- Special Ontologies: Terminological information for certain subjects /areas of research and technology. Most wiede-spread are bio-medical ontologies.
- "Upper-model ontologies" provide common-sense, general terminological knowledge.
- Ontologies are typically formalised, using a logical representation formalism to encode conceptual knowledge:
  - ➢ Versions of Description Logic (→ OWL)
  - ➢ Predicate /modal logic

- Ontologies are intended to provide language-independent conceptual information. Interface to the natural-language lexicon must be provided. Typically throuhg WordNet.
- Available upper-model ontologies:
  - ➢ CYC: a huge ontology which is very expensive (and maybe not really useful). "Open CYC" is free, but has no coverage.
  - ➢ SUMO (The Suggested Upper Merged Ontology) – Size: 2.600 concepts, 6.000 relations, 2.000 rules

- Web interface for SUMO http://berkelium.teknowledge.com:8080/sigma/home.jsp

Fish in SUMO [1]

- Description of Concept:
- (documentation Fish "A cold-blooded aquatic Vertebrate characterized by fins and breathing by gills. Included here are Fish having either a bony skeleton, such as a perch, or a cartilaginous skeleton, such as a shark. Also included are those Fish lacking a jaw, such as a lamprey or hagfish.")
- Relationship to other concepts:

```
 (subclass Fish ColdBloodedVertebrate)
 (disjointDecomposition ColdBloodedVertebrate
    Amphibian Fish Reptile)
```

Fish in SUMO [2]

- A rule:

```
(=>
      (instance ?FISH Fish)
      (exists
         (?WATER)
         (and
            (inhabits ?FISH ?WATER)
            (instance ?WATER Water))))
```

- ... and its semi-colloquial paraphrase:

"if instance FISH Fish, then there exists WATER such that inhabits FISH WATER and instance WATER Water"

Words with overt arguments

Dolphins are mammals, not fish. They are warm blooded like man, 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.

Thematic roles: Some observations

- *Mary likes John*
- *John pleases Mary*

like(x,y) ↔ please (y,x)

- *Mary gave Peter the book*
- *Peter received the book from Mary*

give (x,y,z) ↔ receive_from (y,x,z)

Some observations [3]:

- *The window broke*
- *A rock broke the window*
- *John broke the window with a rock*

$$\text{break}_3(x,y,z) \models \text{break}_2(z,y) \models \text{break}_1(y)$$

Thematic Roles (Fillmore 1968)

- Frames are the units for the conceptual modelling of the world: structured schemata representing complex situations, events, and actions. The meaning of words in terms of the part which they play in frames.
- Thematic roles describe the conceptual participants in a situation in a generic way, independent from their grammatical realization.

Examples for Thematic Roles

- Agent
- Theme/ Patient/ Object
- Recipient
- Instrument
- Source
- Goal
- Beneficient
- Experiencer

Examples Annotated with Thematic Roles

- *[The window]$_{pat}$ broke*
- *[A rock]$_{inst}$ broke [the window]$_{pat}$*
- *[John]$_{ag}$ broke [the window]$_{pat}$ [with a rock]$_{inst}$*

- *[Peter]$_{ag}$ gave [Mary]$_{rec}$ [the book]$_{pat}$*
- *[Mary]$_{rec}$ received [the book]$_{pat}$ [from Peter]$_{ag}$*

Thematic Roles

- allow more abstract/ generic semantic representations
- support the encoding and application of general inference rules
- support the semantic interpretation process (→ role linking)

Role linking, example

give: SB → Agent
OA → Theme
OD → Recipient

get: SB → Recipient
OA → Theme
OP-from → Agent

The „Role Dilemma“

- In Fillmore's original theory and in early KR research a small, closed, and universally applicable inventory of roles is postulated.
- This assumption is untenable, given the semantic richness of natural languages.

Fillmores Frame-semantic Concept (1976)

- „...first identify the phenomena, experiences, or scenarios represented by the meanings of the *target words ...“*
- *„...*then identify labels to the parts or aspects of these which are associated with specific means of linguistic expression ...*frame elements ...“*

… implemented in the Berkeley FrameNet Database (since 1996)

- Frames: an inventory of conceptual structures modelling a prototypical situation like “COMMERCIAL_TRANSACTION”, “COMMUNICATION_REQUEST”, "SELF_MOTION"
- Semantic roles are locally valid only (and accordingly called “Frame Elements” (FE):
  - FEs of the COMMUNICATION_REQUEST frame: SPEAKER, ADDRESSEE, MESSAGE, ...
  - FEs of the COMMERCIAL_TRANSACTION frame: BUYER, SELLER, GOODS, PRICE, ...
- A set of "target words" associated with each frame: e.g., for COMMERCIAL_TRANSACTION:
  - buy, sell, pay, spend, cost, charge,
  - price, change, debt, credit, merchant, broker, shop
  - tip, fee, honorarium, tuition

An example [1]

- Airbus sells five A380 superjumbo planes to China Southern for 220 million Euro
- China Southern buys five A380 superjumbo planes from Airbus for 220 million Euro
- Airbus arranged with China Southern for the sale of five A380 superjumbo planes at a price of 220 million Euro
- Five A380 superjumbo planes will go for 220 million Euro to China Southern

An example [2]

- COMMERCIAL_TRANSACTION
  - SELLER: Airbus
  - BUYER: China Southern
  - GOODS: five A380 superjumbo planes
  - PRICE: 220 million Euro

The Berkeley FrameNet Database

The FrameNet database consists of:

- A data-base of frames with
  - ➢ Descriptions of frames with inventory of Roles/Frame elements and associated lemmas
  - ➢ Frame-to-Frame Relations
- A lexicon with
  - ➢ Frame information
  - ➢ Grammatical realisation patterns (Role Linking)
  - ➢ Annotations of example sentences (from BNC) for all use variants of words

The Berkeley FrameNet Database

- Current release: 700 frames, about 8000 lexical units (mostly verbs)
- Planned: A total of 15000 verb descriptions
- http://framenet.icsi.berkeley.edu/

FrameNet: Advantages

- A unified modeling of the core lexicon of English (relational) expressions, mostly verbs, but also deverbal nouns and relational adjectives, which supports
  - ➢ semantic representation at an appropriate level of granularity and abstraction
  - ➢ semantic construction via grammatical realization patterns
  - ➢ inference based on role information
  - ➢ An almost ideal platform for cross-lingual lexical-semantic resources

FrameNet: Disadvantages

- Lack of coverage (only 50% of the English Core Lexicon described, several years required for completion)
- Few and rather unsystematic information about Frame-to-Frame Relations (hierachical relations, causation etc.)
- Some WordNet information is lost (cf. good/bad in MORALITY_EVALUATION frame, believe/know in AWARENESS frame)
- Interfaces for language technology purposes are (still) lacking

- A perspective: The Saarbrücken SALSA project