Overview

1. Word senses
2. Word sense disambiguation
3. Semantic lexica

Word Senses

- Lexical semantics is the study of how and what the words of a language denote.
- Lexical semantics involves the meaning of each individual word
- A word sense is one of the meanings of a word
- A word is called ambiguous if it can be interpreted in more than one way, i.e., if it has multiple senses.
- Disambiguation determines a specific sense of an ambiguous word.
Homonymy and Polysemy

- A homonym is a word with multiple, unrelated meanings.
  A homonym is a word that is spelled and pronounced the same as another but with a different meaning.
  - bank → financial institution
  - slope of land alongside a river
- A polyseme is a word with multiple, related meanings.
  - school → I go to school every day. (institution)
  - The school has a blue facade. (building)
  - The school is on strike. (teacher)
- Regular polysemy performs a regular induction of a word sense on the basis of another, e.g. school / office.

Human Beings and Ambiguity

- What seems perfectly obvious to a human being is deeply ambiguous to the computer, and there is no easy way of resolving ambiguity.
  - I paid the money on my bank account.
  - I watched the ducks on the river bank.
- Semantic priming (psycholinguistics):
  The response time for a word is reduced when it is presented with a semantically related word.
  - doctor → nurse / butter
- If an ambiguous prime such as bank is given, it turns out that all word senses are primed for bank → money / river

Disambiguation Cues

- Probability and prototypicality → default interpretation:
  corpus-related importance of word senses
- Internal text evidence: context, in particular collocations
- One sense per discourse
- Domain
- Real-world knowledge
Word Sense Disambiguation (WSD)

- WSD: associating a word in a text with a meaning (sense) which can be distinguished from other meanings the word potentially has.
- Intermediate task: not an end in itself, but (arguably) necessary in most NLP tasks, such as machine translation, information retrieval, speech processing.
- Problems:
  1. Which are the senses?
  2. Which is the correct sense?
- Sources of information:
  1. Context of the word to be disambiguated (local, global).
  2. External knowledge sources (e.g. dictionary definitions).

Sense Inventory

- Word Sense Disambiguation needs a set of word senses to disambiguate between.
  - Word Sense Discrimination doesn’t.
- Sense inventories are found in dictionaries, thesauri or similar.
- The granularity and criteria for the set of senses differ (lumpers vs. splitters).
- There is no reason to expect a single set of word senses to be appropriate for different NLP applications.

Lexical Semantic Resources

- Sense inventory and organisation:
  - WordNet
- Sense annotation and semantic role annotation:
  - Prague Dependency Treebank
  - FrameNet
  - PropBank
  - OntoBank / OntoNotes
WordNet

- Online lexical reference system, freely available also for downloading
- The design is inspired by current psycholinguistic theories of human lexical memory.
- English nouns, verbs, adjectives and adverbs are organised into synonym sets (synsets).
- Each synset represents one underlying lexical concept.
- Different (paradigmatic) relations link the synonym sets.
- WordNet was developed by the Cognitive Science Laboratory at Princeton University under the direction of George A. Miller.
- WordNets now exist for many languages.

WordNet Synsets

- Synsets are sets of synonymous words ("literals").
- Polysemous words appear in multiple synsets.
- Examples:
  - noun example: (coffee, java) (coffee, coffee tree) (coffee bean, coffee berry, coffee)
  - adjective : (chocolate, coffee, deep brown, umber, burnt umber)
- adjective example:
  - (cold)
    - cold, cold
    - cold, dry, uncordial
    - cold, unaffectionate, uncaring
    - cold, old

More about synsets

Synsets also include:
- glosses (definitions)
- examples of usage
- e.g.
  - (n) glass (glassware collectively) "She collected old glass"
- recently added by ITC, Italy: semantic domains
  - Example: Bank
    - forward financial institution, bank, banking company, finance institution
    - bank (business, firm, ...)
    - bank (as a supply or stock held in reserve)
    - bank, bank building (in building ...)
    - Bank, Banking, Industries
    - Economy, Geography, Sociology
    - Economics, Business, Architecture, Economy
### WordNet Relations

- Within synsets:
  - Synonymy, such as (coffee, java)
- Between synsets / parts of synsets:
  - Antonymy: opposition, e.g. (cold) - (hot)
  - Hyponymy / Hypernymy: is-a relation, e.g. (coffee, java) - (beverage, drink, potable)
  - Meronymy / Holonymy: part-of relation, e.g. (coffee bean, coffee berry, coffee) - (coffee, coffee tree)
- Morphology:
  - Derivations: appealing - appealingness

### WordNet Hierarchy

- Depending on the part-of-speech, different relations are defined for a word. For example, the core relation for nouns is hypernymy, the core relation for adjectives is antonymy.
- Hypernymy imposes a hierarchical structure on the synsets.
- The most general synsets in the hierarchy consist of a number of pre-defined disjunctive top-level synsets:
  - nouns → (entity), (abstraction), (psychological), ...
  - verbs → (move), (change), (get), (feel), ...

### WordNet Hierarchy: Examples

{entity}       (abstraction)
| {object, inanimate object, physical object} | {attribute} |
| {substance, matter} | {property} |
| {food, nutrient} | (visual property) |
| {beverage, drink, potable} | (color, coloring) |
| {coffee, java} | {brown, brownness} |
| | (chocolate, coffee, deep brown, umber, burnt umber) |
WordNet Family

- Current status: WordNets for 38 languages
- WordNets in the world:
- Integration of WordNets into multi-lingual resources:
  - EuroWordNet: English, Dutch, Italian, Spanish, German, French, Czech and Estonian
  - BalkaNet: Bulgarian, Czech, Greek, Romanian, Turkish, Serbian
- An inter-lingual index connects the synsets of the WordNets
- ~ multilingual lexicon; machine translation

WordNet annotated corpora

- SemCor: created at Princeton University, a subset Brown corpus (700,000 words), 200,000 content words are WordNet sense-tagged
- MultiSemCor: created at ITC, Italy, consists of SemCor + translation into Italian, which is also sense-tagged http://multisemcor.itc.it/
- DSO Corpus of Sense-Tagged English (National University of Singapore)
- etc.

Prague Dependency Treebank

- Three-level annotation scenario:
  - 1. morphological level
  - 2. syntactic annotation at the analytical level
  - 3. linguistic meaning at the tectogrammatical level
- Corpus data: newspaper articles (60%), economic news and analyses (20%), popular science magazines (20%)
- 1 million tokens are annotated on the tectogrammatical level.
Tectogrammatical Level of the PDT

- Annotation: dependency, functor, ellipsis resolution, coreference, ...
- 39 attributes
- Similar to the surface (analytical) level, but:
  - certain nodes deleted (auxiliaries, non-autosemantic words, punctuation)
  - some nodes added (based on word - mostly verb, noun - valency)
  - some ellipsis resolution (detailed dependency relation labels: functors)


Tectogrammatical Functors

- General functors, e.g.: actor/bearer, addressee, patient, origin, effect, cause, regard, concession, aim, manner, extent, substitution, accompaniment, locative, means, temporal, attitude, cause, regard, directional, benefactive, comparison
- Specific functors for dependents on nouns, e.g.: material, appurtenance, restrictive, descriptive, identity
- Subtle differentiation of syntactic relations, e.g.: temporal (before, after, on), accompaniment, regard, benefactive (for/against)


Tectogrammatical Example

- Example: (he) gave him a book

  dal    mu   knihu

  The “Obj” goes into ACT, PAT, ADDR, EFF or ORIG, as based on the governor’s valency frame.
Analytical vs. Tectogrammatical Level

Other semantic lexica/corpora

- FrameNet
- PropBank
- OntoNotes
- ...
