# Taal- en spraaktechnologie
# J&M. Chapter 17.

Sophia Katrenko

Utrecht University, the Netherlands
May 30, 2012

# Outline

# Focus

This part of the course focuses on

- meaning representation
- lexical semantics
- distributional similarity
- intro to machine learning
- word sense disambiguation
- information extraction

## Today

- Chapter 17 (Representing meaning)
- Chapter 19 (Lexical semantics)

# Terminology

## Meaning representation

- the meaning of linguistic utterances can be captured in formal structures

## Meaning representation languages

- frameworks that are used to specify the syntax and semantics of these representations

# Terminology

Why yet another representation?

Linguistic input needs to be combined with world knowledge:

- how to recognize humor?
- how to follow a recipe?
- how to learn the use of software given its manual?

## Terminology

What is semantic analysis?

- The process of creating meaning representations and assigning them to linguistic inputs
- These representations are made up of the same-kind-of-stuff that is used to represent this kind of everyday commonsense knowledge of the world

# Terminology

The representation languages we consider here are

- First-Order Logic (FOL)
- Semantic Network (SN)
- Conceptual Dependency (CD)
- Frame-Based representation (FB)

# Terminology

## Example

I have a car

**FOL**: $\exists x, y \; Having(x) \wedge Haver(Speaker, x) \wedge HadThing(y, x) \wedge Car(y)$

**SN**: $Car \longleftarrow HadThing \longleftarrow Having \longrightarrow Haver \longrightarrow Speaker$

**CD**: $Car \Uparrow (Poss - By) \; Speaker$

**FB**:

Having
Haver: Speaker
HadThing: Car

# Terminology

These approaches share

- the notion that a meaning representation consists of structures composed from a set of symbols, or **representational vocabulary**.
- when appropriately arranged, these symbol structures are taken to correspond to the objects, properties of objects and relations among objects in some state of affairs being represented.

## Terminology

Meaning can be

- **literal** (conventional meaning, the one we discuss here)
- **implied**
- **figurative** (e.g., metaphors)

There is a difference between **literal** meaning and **utterance (speaker's)** meaning.

# Requirements (1)

Requirements to meaning representation

- **Verifiability** (a system should be able to compare the representation of the meaning of an input against the representations in its knowledge base)
- **Ability to deal with vagueness**
- **Unambiguity** of the final representation of an input's meaning
- **Canonical form** (inputs that mean the same thing should have the same meaning representation)

# Requirements (2)

Requirements to meaning representation

- **Inference** (a system's ability to draw valid conclusions based on the meaning representation of inputs and its store of background knowledge)
- **Expressiveness** (a meaning representation language should adequately represent the meaning of *many* sensible natural language utterances)

# Meaning structure

Languages convey meaning by

- conventional form-meaning associations
- word-order regularities, tense systems
- conjunctions and quantifiers
- fundamental predicate-argument structure

# Predicate-argument structures (1)

## Predicate-argument structure

A predicate-argument structure describes relationships (or dependencies) among concepts underlying sentential constituents.

## Example

1. I want Italian food. (*NP want NP*)
2. I want to spend less than five dollars. (*NP want Inf-VP*)
3. I want it to be close by here. (*NP want NPInf-VP*)

The semantic roles in the semantic representation can be derived using arguments of verb subcategorization frames.

## Predicate-argument structures (2)

- The semantic roles in the semantic representation can be derived using arguments of verb subcategorization frames.
- There is more than merely a syntactic restriction - not all categories can be arguments of a certain verb (*selectional preference*).
- Predicate-argument structures can be obtained from not necessarily verb phrases (e.g., **an Italian restaurant under fifteen dollars**).

# Predicate-argument structures (3)

Meaning representation language has to support

- variable arity predicate-argument structures
- the semantic labeling of arguments to predicates
- the statement of semantic constraints on the fillers of argument roles

# Model-theoretic semantics (1)

A meaning representation language is a means to describe

- objects
- properties of objects
- relations among objects

Expressions in a meaning representation language are mapped in a systematic way to the elements of the model.

# Model-theoretic semantics (2)

Vocabulary of a meaning representation:

- **non-logical vocabulary** (the open-ended set of names for the objects, properties and relations that make up the world, e.g. predicate)
- **logical vocabulary** (the closed set of symbols, operators, quantifiers).

Each element of the non-logical vocabulary has a denotation in the model (= corresponds to a fixed well-defined part of the model).

# Model-theoretic semantics (3)

Properties and relations are described extensionally:

- objects denote **elements** of the domain
- properties denote **sets of elements** of the domain
- relations denote **sets of tuples of elements** of the domain

**Interpretation:** a function that maps from the non-logical vocabulary of the meaning representation to the proper denotations in the model.

# Model-theoretic semantics (4)

## Example

**Domain** $D = \{a, b, c, \ldots\}$
Matthew, Franco, Katie and Caroline = a, b, c, d
Frasca, Med, Rio = e, f, g

**Noisy**
Frasca and Rio are noisy = $\{e, g\}$

**Likes**
Matthew likes the Med
Katie likes the Med and Rio
$Likes = \{< a, f >, < c, f >, < c, g >\}$

# Model-theoretic semantics (5)

Meaning decomposition:

- complex expressions need to be decomposed in parts whose meanings can be grounded
- operators have to be given truth-conditional semantics
- the semantics of the entire logical vocabulary of the meaning representation scheme has to be specified

# First-order logic (1)

## Definition

**terms**: $Term \rightarrow Function(Term, \ldots) | Constant | Variable$

**formulas**:
$Formula \rightarrow AtomicFormula |\ Formula\ Connective\ Formula |$
$Quantifier\ Variable, \ldots Formula |\ \neg Formula$

$AtomicFormula \rightarrow Predicate(Term, \ldots)$

$Constant \rightarrow Frasca | B | \ldots$
$Variable \rightarrow x | y | \ldots$
$Connective \rightarrow \wedge | \vee | \Rightarrow$
$Quantifier \rightarrow \exists | \forall$
$Predicate \rightarrow Likes | \ldots$
$Function \rightarrow LocationOf | \ldots$

# First-order logic (2)

## Examples

I only have five dollars and I don't have a lot of time.
$Have(Speaker, FiveDollars) \wedge \neg Have(Speaker, LotOfTime)$

Every man goes to work.
$\forall x Man(x) \Rightarrow Go(x, Work)$

# First-order logic (3)

## Semantics

- atomic formulas are true if they are literally present in the knowledge base or if they can be inferred from other formula that are in the knowledge base
- if a formula has logical connectives, then its meaning is based on the meaning of the components combined with the meanings of the connectives it contains, e.g.

| P | Q | ¬ P | ¬ Q | P ∧ Q | P ∧ Q | P ⇒ Q |
|---|---|---|---|---|---|---|
| true | false | false | true | false | true | false |
| ... | | | | | | |

# Inference (1)

How can one add valid new propositions to a knowledge base?

- **modus ponens** (if-then reasoning):

$$\frac{\begin{array}{l}\alpha \\ \alpha \Rightarrow \beta\end{array}}{\beta}$$

$$\frac{\begin{array}{l}\text{Man(Socrates)} \Rightarrow \mathit{Mortal}(\mathit{Socrates}) \\ \text{Man(Socrates)}\end{array}}{\text{Mortal(Socrates)}}$$

- **forward chaining** individual facts are added to the knowledge base and modus ponens is used to fire all applicable implication rules $\rightarrow$ new facts are added to the knowledge base and the process repeats.

# Inference (1)

How can one add valid new propositions to a knowledge base?

- **modus ponens** (if-then reasoning):

$$\frac{\begin{array}{l}\alpha \\ \alpha \Rightarrow \beta\end{array}}{\beta}$$

$$\frac{\begin{array}{l}\text{Man(Socrates)} \Rightarrow \mathit{Mortal}(\mathit{Socrates}) \\ \text{Man(Socrates)}\end{array}}{\text{Mortal(Socrates)}}$$

- **forward chaining** individual facts are added to the knowledge base and modus ponens is used to fire all applicable implication rules $\rightarrow$ new facts are added to the knowledge base and the process repeats.

## Inference (2)

How can one add valid new propositions to a knowledge base?

- **backward chaining** modus ponens is run in reverse to prove specific propositions, called queries.

To note:

1. **backward chaining** (from queries to known facts) vs. **reasoning backwards** (from known consequents to unknown antecedents; also known as *plausible reasoning* or *abduction*).
2. both backward and forward reasoning are sound (inference rules prove only formulas that are valid with respect to its semantics) but not complete (not every validity is provable).

## Categories

How to represent **categories**?

- via unary predicates, e.g. *Man*(*Socrates*) (category - *Man*, *Socrates* - a member of this category).
- **reification** (category represented as an object), e.g. *ISA*(*Socrates*, *Man*), can also be extended to hold between categories: *ISA*(*Man*, *Human*).

*ISA* relation is used to create hierarchies, consider WordNet (will be discussed later).

# Events (1)

How to represent **events**?

- if a verb: a predicate that represents the meaning of a verb has the same number of arguments as are present in the verb's syntactic subcategorization frame.

  but ...

  - difficult to determine the correct number of roles for any given event.
  - how to represent facts about the roles associated with an event?
  - all the correct inferences have to be derived directly from the representation of an event.
  - no incorrect inferences can be derived from the representation of an event.

# Events (2)

How to represent **events**?

## Example

- Ik heb winst gemaakt op de verkoop van mijn huis.
- "Ik heb Arjen gek gemaakt voor Bayern, dat heeft mij behoorlijk wat telefoonkosten opgeleverd."
- Ik heb een remix voor hem gemaakt. ($Maken_3$)
- Ik heb een remix gemaakt. ($Maken_4$)
- Geld wegschenken maakt vrouwen gelukkig.

# Events (3)

## Example

- Ik heb een remix voor hem gemaakt. (*Maken*$_3$)
- Ik heb een remix gemaakt. (*Maken*$_4$)

But predicates have fixed arity!

Solutions:
(a) create as many predicates as there are different uses of *te maken* (costly!),
(b) use meaning postulates (scalability problems), e.g.
$\forall w, x \ Maken_3(w, x) \Rightarrow Maken_4(w)$
(c) allow as many arguments in the definition of the predicate as ever appear with it (but may be missing): events are not individuated!

# Events (3)

**Example**

- Ik heb een remix voor hem gemaakt. ($Maken_3$)
- Ik heb een remix gemaakt. ($Maken_4$)

But predicates have fixed arity!

Solutions:
(a) create as many predicates as there are different uses of *te maken* (costly!),
(b) use meaning postulates (scalability problems), e.g.
$\forall w, x \; Maken_3(w, x) \Rightarrow Maken_4(w)$
(c) allow as many arguments in the definition of the predicate as ever appear with it (but may be missing): events are not individuated!

# Events (4)

## Example

- Ik heb een remix voor hem gemaakt. ($Maken_3$)
- Ik heb een remix gemaakt. ($Maken_4$)

Solutions:

(d) Reification: e.g.,

$\exists w \ ISA(w, Making) \wedge Maker(w, Speaker) \wedge Made(w, Remix)$

Pros of (d):

- no need to specify a fixed number of arguments for a given surface predicate
- no more roles are postulated

# Time (1)

How to represent **time**?

- events are associated with either points or intervals in time.
- distinct events can be ordered given the timeline (one event may follow or precede another).

**Example**

1. Ik zal mijn huiswerk doen.
2. Ik doe mijn huiswerk.
3. Ik heb mijn huiswerk gedaan.

? $\exists w \; ISA(w, Doen) \wedge Dader(w, Spreker) \wedge Daad(w, Huiswerk)$

# Time (2)

How to represent **time**?

Add temporal variables representing the interval corresponding to the event, the end point of the event, and temporal predicates relating this end point to the current time as indicated by the tense of the verb:

1. $\exists i, e, w, t \; ISA(w, Doen) \wedge Dader(w, Spreker) \wedge$
   $\wedge Daad(w, Huiswerk) \wedge IntervalVan(w, i) \wedge Eindpunt(i, e) \wedge$
   $Volgt(e, Nu)$
2. $\exists i, e, w, t \; ISA(w, Doen) \wedge Dader(w, Spreker) \wedge$
   $\wedge Daad(w, Huiswerk) \wedge IntervalVan(w, i) \wedge ElementVan(i, Nu)$
3. $\exists i, e, w, t \; ISA(w, Doen) \wedge Dader(w, Spreker) \wedge$
   $\wedge Daad(w, Huiswerk) \wedge IntervalVan(w, i) \wedge Eindpunt(i, e) \wedge$
   $Volgt(Nu, e)$

# Time (3)

How to represent **time**?

But what about *Ik ga naar New York*, or *My flight arrived late* vs. *My flight had arrived late*?

Reichenbach (1947): reference point (the notion of the reference point is separated out from the utterance time and the event time).

1. When Mary's flight departed, I ate lunch.
2. When Mary's flight departed, I had eaten lunch.

Mary's departure = reference point

## Aspect

Based on aspect, event expressions can be

- **statives** (an event participant is in a particular state at some point in time): *I want to go first class. I know you are in Utrecht now.* Not used in the progressive form, not modified by *carefully* and alike, not used in imperative.
- **activities** (event with no particular end point): *I listen to jazz.* Allow progressive and imperative, not modified by *in*.
- **accomplishments** (events with an end point and resulting in some state): *You took an exam.*
- **achievements** (also result in a state but no particular activity leads to it): *I reached Utrecht.*

# Aspect

Based on aspect, event expressions can be

- **statives** (an event participant is in a particular state at some point in time): *I want to go first class. I know you are in Utrecht now.* Not used in the progressive form, not modified by *carefully* and alike, not used in imperative.
- **activities** (event with no particular end point): *I listen to jazz.* Allow progressive and imperative, not modified by *in*.
- **accomplishments** (events with an end point and resulting in some state): *You took an exam.*
- **achievements** (also result in a state but no particular activity leads to it): *I reached Utrecht.*

# Beliefs

Expressions may refer not to the actual world, but to some hypothetical world, which would require modeling of beliefs.

- can be done using reification, e.g. *I believe that Jan wrote a poem.*: $ISA(u, Believing) \wedge Believer(u, Speaker \ldots)$ but if I believe anything, it does not make it true.
- and $Believing(Speaker, Writing(Jan, Poem))$ is not a valid FOL formula.
- introducing operator *Believes*:
  $Believes(Speaker, \exists v$
  $ISA(v, Writing) \wedge Writer(v, Jan) \wedge Product(v, Poem))$

# Description Logic

Description Logics correspond to varying subsets of FOL (to ensure the tractability of inferences).

- **Terminology (TBox)**: concepts of some domain, organized hierarchically (using subsumption relation), e.g.
- **ABox**: facts about individuals.

**Example**

ItalianRestaurant $\sqsubseteq$ Restaurant
GreekRestaurant $\sqsubseteq$ Restaurant
ItalianRestaurant $\sqsubseteq$ Restaurant $\sqcap \exists$ hasCuisine.ItalianCuisine

# Lexical semantics

# Lexical semantics

We move from meaning representations of sentences to those of words.

- **lexicon**: fixed list of lexemes.
- **lemma**: grammatical form that represents a lexeme (e.g., *maken* for *gemaakt*, *maakt*, *maakte* and *maken*).
- **lemmatization**: mapping from wordforms to lemmata (not always deterministic - why?; may be larger than morphological stems - when?).

# Word senses (1)

The meaning of a lemma depends on the context:

- **bank**: ‘financial institution’ (*I have money in a bank*) vs. ‘sloping mound’ (*We were at the river’s bank*).

A **word sense** is a representation of one aspect of the meaning of a word. Word senses can be

- unrelated, e.g. **homonymy** as in the *bank* example above
- semantically related, **polysemy** (the relation is structured and systematic): **bank** as a building of a financial institution (*This bank is located on Frederiksplein*) is related to the sense of **bank** as financial institution.

# Word senses (2)

Other relations between word senses:

- **metonymy**: one aspect of a concept is used to refer to the entire concept or its other aspects:
  - *I have Mulisch at home.*
  - *The chicken walk in the garden* vs. *He ordered the chicken.*
- **zeugma** (conjunction of antagonistic readings):
  - *This flight serves breakfast.*
  - *KLM serves Krakow.*
  - *KLM serves Krakow and breakfast.*
- **homonymy** (two sense with the same pronunciation and ortography)
  - **homophones** (same pronunciation, different spelling): *bye-by*
  - **homographs** (same spelling, different pronunciation)

# Word senses (3)

How to define a word sense?

- for humans: dictionary - but there is circularity in definitions (self-references or two definitions referencing each other)
- for computational purposes
  - through sense relationships (e.g., WordNet, EuroWordNet or Cyc).
  - by selecting a set of semantic primitives whose combination defines a sense (e.g., semantic roles)

# Word senses (4)

Adam Kilgariff. *I don't believe in word senses*. In Computers and the Humanities 31: 91-113, 1997.

`http://www.kilgarriff.co.uk/Publications/1997-K-CHum-believe.pdf`

*"an alternative conception of the word sense, in which it corresponds to a cluster of citations for a word ... Citations are clustered together where they exhibit similar patterning and meaning."*

*"there is no reason to expect a single set of word senses to be appropriate for different NLP applications. Different corpora, and different purposes, will lead to different senses."*

# Word senses (5)

Word sense relations: **synonymy**

- Two senses of two different lemmas are (nearly) identical (*car/automobile*).
- For words, synonymy is defined via substitutability one for the other in any sentence such that the truth condition of the sentence remains the same (*propositional meaning*).
- Some sense of words may be synonymous while others are not (e.g., *large sister* vs. *big sister*)

# Word senses (6)

Word sense relations: **antonymy**

- Two words with opposite meaning: *cold/hot*, *up/down*, *rich/poor*, *day/night*.
- Two senses are ambiguous if there is binary opposition, as in *long/short*.
- **Reversibles**: antonyms that describe a change or movement in opposite directions: *fall/rise*.
- It is difficult to distinguish between antonyms and synonyms automatically (why?).

# Word senses (7)

Word sense relations: **hyponymy**

- One sense is a subclass of another

university is an institution

*university* - **hyponym**
*institution* - **hypernym, superordinate**

- The class denoted by the hypernym *extensionally* includes the class denoted by the hyponym.
- It can defined via entailment: $\forall x \; A(x) \Rightarrow B(x)$.
- Hyponymy is transitive
  $(hyponymy(X, Y), hyponymy(Y, Z) \Rightarrow hyponymy(X, Z))$.

# Word senses (8)

Word sense relations: **meronymy** or **part-whole**

- relation between a part and the whole

a car has four wheels and two doors

*car* - **holonym**
*wheel* - **meronym**
*door* - **meronym**

- Usually, not considered transitive.
- It differs from *content-container* relation, compare *apples in a basket* vs. *trees in a forest*. Which one is an example of meronymy?

# Word senses (9)

Winston et al. (1987): there are six types of meronymy

1. component - integral object, e.g.: *handle - cup*
2. member - collection, e.g.: *tree - forest*
3. portion - mass, e.g.: *grain - salt*
4. stuff - object, e.g.: *steel - bike*
5. feature - activity, e.g.: *dating - adolescence*
6. place - area, e.g.: *oasis - desert*

Three "relation elements": functional, homeomerous (the part is identical to the other parts making up the whole), and separable.

*handle - cup*:

# Word senses (10)

Winston et al. (1987): there are six types of meronymy

1. component - integral object, e.g.: *handle - cup*
2. member - collection, e.g.: *tree - forest*
3. portion - mass, e.g.: *grain - salt*
4. stuff - object, e.g.: *steel - bike*
5. feature - activity, e.g.: *dating - adolescence*
6. place - area, e.g.: *oasis - desert*

Three "relation elements": functional, homeomerous (the part is identical to the other parts making up the whole), and separable.

*handle - cup*: functional (+), homeomerous (-), and separable (+)

# WordNet (1)

So what is WordNet (Miller et al., 1990)?

- A wide-coverage computational lexicon of English which exploits psycholinguistic theories (Fellbaum, 1998).
- Concepts are expressed as sets of synonyms (synsets)
  { $\text{bank}^7_n$, $\text{cant}^2_n$, $\text{camber}^2_n$ }
- A word sense is a word occurring in a synset, e.g. $\text{bank}^7_n$ is the seventh sense of noun *bank*
- There are also *semantic relations* between synsets (e.g., hypernymy, meronymy, entailment), and *lexical relations* between word senses (e.g., antonymy, nominalization).

# WordNet (2)

WordNet 3.0 stats:

- databases for *nouns*, *verbs* and *adjectives*.
- no closed class words.
- the number of word-sense pairs: 206, 941 (nouns: 146,312, verbs: 25,047, adjectives: 30002, adverbs: 5,580)
- average polysemy:

| | Including Monosemous W. | Excluding Monosemous W. |
|---|---|---|
| Noun | 1.24 | 2.79 |
| Verb | 2.17 | 3.57 |
| Adjective | 1.40 | 2.71 |
| Adverb | 1.25 | 2.50 |

# WordNet (3)

## Noun

- S: (n) **bank** (sloping land (especially the slope beside a body of water)) *"they pulled the canoe up on the bank"; "he sat on the bank of the river and watched the currents"*
- S: (n) depository financial institution, **bank**, banking concern, banking company (a financial institution that accepts deposits and channels the money into lending activities) *"he cashed a check at the bank"; "that bank holds the mortgage on my home"*
- S: (n) **bank** (a long ridge or pile) *"a huge bank of earth"*
- S: (n) **bank** (an arrangement of similar objects in a row or in tiers) *"he operated a bank of switches"*
- S: (n) **bank** (a supply or stock held in reserve for future use (especially in emergencies))
- S: (n) **bank** (the funds held by a gambling house or the dealer in some gambling games) *"he tried to break the bank at Monte Carlo"*
- S: (n) **bank**, cant, camber (a slope in the turn of a road or track; the outside is higher than the inside in order to reduce the effects of centrifugal force)
- S: (n) savings bank, coin bank, money box, **bank** (a container (usually with a slot in the top) for keeping money at home) *"the coin bank was empty"*
- S: (n) **bank**, bank building (a building in which the business of banking transacted) *"the bank is on the corner of Nassau and Witherspoon"*
- S: (n) **bank** (a flight maneuver; aircraft tips laterally about its longitudinal axis (especially in turning)) *"the plane went into a steep bank"*

# WordNet (4)

Sentence: Utrecht University has concentrated its leading research into fifteen research focus areas.

| | | | | | | |
|---|---|---|---|---|---|---|
| Utrecht | University | has | concentrated | its | leading | research |
| 1 × | 3 × | 19 × | 8 × | 1 × | 4 × | 2 × |

| | | | | |
|---|---|---|---|---|
| into | fifteen | research | focus | areas. |
| 1 × | 1 × | 2 × | 7 × | 6 |

$= \mathbf{306{,}432}$ interpretations!

Note that I already assumed the correct PoS tags here!
**Utrecht** has only 1 sense, and is therefore *monosemous*, while **focus** is *polysemous.*

# WordNet (5)

WordNet online: `http://wordnetweb.princeton.edu/perl/webwn`

## Summary

Today, we

- have reviewed several approaches to meaning representation
- started discussing lexical semantics

Next class: Friday, June 1