ANLP Lecture 19
Lexical Semantics:
Word senses, relations, and thematic roles

Sharon Goldwater
(based on slides by Philipp Koehn)

27 October 2014
Meaning

• The grand goal of artificial intelligence
  – machines that do not mindlessly process data
  – ... but that ultimately understand its meaning

• But what is meaning?
Eliza

Joseph Weizenbaum wrote the computer program Eliza in 1969 to demonstrate how easily people can be fooled into thinking that machines have some deep understanding.

young woman: Men are all alike.
eliza: In what way?
young woman: They’re always bugging us about something specific or other.
eliza: Can you think of a specific example?
young woman: Well, my boyfriend made me come here.
eliza: Your boyfriend made you come here?
young woman: He says I’m depressed much of the time.
eliza: I am sorry to hear you are depressed.
young woman: It’s true. I’m unhappy.
eliza: Do you think coming here will help you not to be unhappy?

Online demo: http://www-ai.ijs.si/eliza-cgi-bin/eliza_script
What is meaning? What is understanding?

- These are deep philosophical questions

- NLP usually takes a more pragmatic view: can the computer behave as though it understands (in order to do what we want)?
  - Dialogue systems (e.g., Eliza)
  - Machine translation
  - Question answering

- What issues will we face in building such systems?
A Concrete Goal

• We would like to build
  – a machine that answers questions in natural language.
  – may have access to knowledge bases
  – may have access to vast quantities of English text

• Basically, a smarter Google

• This is typically called Question Answering
Semantics

- To build our QA system we will need to deal with issues in **semantics**, i.e., meaning.

- Lexical semantics: the meanings of individual words (next few lectures)

- Sentential semantics: how word meanings combine (after that)

- Consider some examples to highlight problems in lexical semantics
Example Question

• Question
  When was Barack Obama born?

• Text available to the machine
  Barack Obama was born on August 4, 1961

• This is easy.
  – just phrase a Google query properly:
    "Barack Obama was born on *"
  – syntactic rules that convert questions into statements are straight-forward
Example Question (2)

• Question
  What plants are native to Scotland?

• Text available to the machine
  A new chemical plant was opened in Scotland.

• What is hard?
  – words may have different meanings (senses)
  – we need to be able to disambiguate between them
Example Question (3)

• Question
  Where did David Cameron go on vacation?

• Text available to the machine
  David Cameron spent his holiday in Cornwall

• What is hard?
  – words may have the same meaning (synonyms)
  – we need to be able to match them
Example Question (5)

- Question
  Which animals love to swim?

- Text available to the machine
  Polar bears love to swim in the freezing waters of the Arctic.

- What is hard?
  - words can refer to a subset (hyponym) or superset (hypernyms) of the concept referred to by another word
  - we need to have database of such A is-a B relationships, called an ontology
Example Question (5)

• Question
  What is a good way to remove wine stains?

• Text available to the machine
  Salt is a great way to eliminate wine stains

• What is hard?
  – words may be related in other ways, including similarity and gradation
  – we need to be able to recognize these to give appropriate responses
Example Question (6)

• Question
  Did Poland reduce its carbon emissions since 1989?

• Text available to the machine
  Due to the collapse of the industrial sector after the end of communism in 1989, all countries in Central Europe saw a fall in carbon emissions.

  Poland is a country in Central Europe.

• What is hard?
  – we need to do inference
  – a problem for sentential, not lexical, semantics
WordNet

- Some of these problems can be solved with a good ontology, e.g., **WordNet**

- WordNet (English) is a hand-built resource containing 117,000 **synsets**: sets of synonymous words (See [http://wordnet.princeton.edu/](http://wordnet.princeton.edu/))

- Synsets are connected by relations such as
  - hyponym/hypernym (IS-A: chair-furniture)
  - meronym (PART-WHOLE: leg-chair)
  - antonym (OPPOSITES: good-bad)

- [globalwordnet.org](http://globalwordnet.org) now lists wordnets in over 50 languages (but variable size/quality/licensing)
Word Sense Ambiguity

- Not all problems can be solved by WordNet alone.

- Two completely different words can be spelled the same (homonyms):
  
  I put my money in the *bank*. vs. He rested at the *bank* of the river.
  You *can* do it! vs. She bought a *can* of soda.

- More generally, words can have multiple (related or unrelated) senses (polysemes)

- Polysemous words often fall into (semi-)predictable patterns: see next slides (from Hugh Rabagliati in PPLS).
<table>
<thead>
<tr>
<th>Pattern</th>
<th>Participating Senses</th>
<th>Example Sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Animal for fur</td>
<td>Mink, chinchilla, rabbit, beaver, raccoon*, alpaca*, crocodile*</td>
<td>The <em>mink</em> drank some water / She likes to wear <em>mink</em></td>
</tr>
<tr>
<td>Animal/Object for personality</td>
<td>Chicken, sheep, pig, snake, star*, rat*, doll*</td>
<td>The <em>chicken</em> drank some water / <em>He is a chicken</em></td>
</tr>
<tr>
<td>Animal for meat</td>
<td>Chicken, lamb, fish, shrimp, salmon*, rabbit*, lobster*</td>
<td>The chicken drank some water / The <em>chicken</em> is tasty</td>
</tr>
<tr>
<td>Artifact for activity</td>
<td>Shower, bath, sauna, baseball,</td>
<td>The shower was leaking / The shower was relaxing</td>
</tr>
<tr>
<td>Body part for object part</td>
<td>Arm, leg, hand, face, back*, head*, foot*, shoulder*, lip*</td>
<td>John’s arm was tired / The <em>arm</em> was reupholstered</td>
</tr>
<tr>
<td>Building for people</td>
<td>Church, factory, school, airplane,</td>
<td>The church was built 20 years ago / The church sang a song</td>
</tr>
<tr>
<td>Complement Coercion</td>
<td>Begin, start, finish, try</td>
<td>John began reading the book / John <em>began</em> the book</td>
</tr>
<tr>
<td>Container for contents</td>
<td>Bottle, can, pot, pan, bowl*, plate*, box*, bucket*</td>
<td>The bottle is made of steel / He drank half of the bottle</td>
</tr>
<tr>
<td>Word for question</td>
<td>Price, weight, speed</td>
<td>The <em>price</em> of the coffee was low / John asked the <em>price</em> of the coffee</td>
</tr>
<tr>
<td>Pattern</td>
<td>Participating Senses</td>
<td>Example Sentences</td>
</tr>
<tr>
<td>---------------------------------------</td>
<td>---------------------------------</td>
<td>----------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Figure for Ground</td>
<td>Window, door, gate, goal</td>
<td>The window is broken / The cat walked through the window</td>
</tr>
<tr>
<td>Grinding</td>
<td>Apple, chair, fly</td>
<td>The apple was tasty / There is apple all over the table</td>
</tr>
<tr>
<td>Instrument for action</td>
<td>Hammer, brush, shovel, tape,</td>
<td>The hammer is heavy / She hammered the nail into the wall</td>
</tr>
<tr>
<td></td>
<td>lock*, bicycle*, comb*, saw*</td>
<td></td>
</tr>
<tr>
<td>Instance of an entity for kind</td>
<td>Tennis, soccer, cat, dog, class*</td>
<td>Tennis was invented in England / Tennis was fun today</td>
</tr>
<tr>
<td></td>
<td>dinner*, chair*, table*</td>
<td></td>
</tr>
<tr>
<td>Location / Place at location</td>
<td>Bench, land, floor, ground,</td>
<td>The bench was made of pine / The coach benched the player</td>
</tr>
<tr>
<td></td>
<td>box*, bottle*, jail*</td>
<td></td>
</tr>
<tr>
<td>Object for placing at goal</td>
<td>Water, paint, salt, butter,</td>
<td>The water is cold / He watered the plant.</td>
</tr>
<tr>
<td></td>
<td>frame*, dress*, oil*</td>
<td></td>
</tr>
<tr>
<td>Object for taking from source</td>
<td>Milk, dust, weed, peel, pit*,</td>
<td>The milk tastes good / He milked the cow</td>
</tr>
<tr>
<td></td>
<td>skin*, juice*</td>
<td></td>
</tr>
<tr>
<td>Material for artifact</td>
<td>Tin, iron, china, glass, linen*</td>
<td>Watch out for the broken glass / He filled the glass with water</td>
</tr>
<tr>
<td></td>
<td>rubber*, nickel*, fur*</td>
<td></td>
</tr>
<tr>
<td>Occupation for role in action</td>
<td>Boss, nurse, guard, tutor</td>
<td>My boss is nice / He bossed me around</td>
</tr>
<tr>
<td>Pattern</td>
<td>Participating Senses</td>
<td>Example Sentences</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>-----------------------------------------------------------</td>
<td>-----------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Place for an event</td>
<td>Vietnam, Korea, Waterloo, Iraq</td>
<td>It is raining in Vietnam / John was shot during Vietnam</td>
</tr>
<tr>
<td>Place for an institution</td>
<td>White House, Washington, Hollywood, Pentagon, Wall Street*, Supreme Court</td>
<td>The White House is being repainted / The White House made an announcement</td>
</tr>
<tr>
<td>Plant for food or material</td>
<td>Corn, broccoli, coffee, cotton, lettuce*, eggs*, oak*, pine*</td>
<td>The large field of corn / The corn is delicious</td>
</tr>
<tr>
<td>Portioning</td>
<td>Water, beer, jam</td>
<td>She drank some water / She bought three waters</td>
</tr>
<tr>
<td>Publisher for product</td>
<td>Newspaper, magazine, encyclopedia, Wall Street Journal*, New York Times*</td>
<td>The newspaper is badly printed / The newspaper fired three employees</td>
</tr>
<tr>
<td>Artist for product</td>
<td>Writer, artist, composer, Shakespeare, Dickens*, Mozart*, Picasso*</td>
<td>The writer drank a lot of wine / The writer is hard to understand</td>
</tr>
<tr>
<td>Visual Metaphor</td>
<td>Beam, belt, column, stick, bug*, leaf*</td>
<td>Most of the weight rests on the beam / There was a beam of light</td>
</tr>
</tbody>
</table>
How many senses?

• How many senses does the word *interest* have?
  
  – She pays 3% *interest* on the loan.
  – He showed a lot of *interest* in the painting.
  – Microsoft purchased a controlling *interest* in Google.
  – It is in the national *interest* to invade the Bahamas.
  – I only have your best *interest* in mind.
  – Playing chess is one of my *interests*.
  – Business *interests* lobbied for the legislation.

• Are these seven different senses? Four? Three?

• Also note: distinction between polysemy and homonymy not always clear!
**WordNet senses for interest**

- **S1:** a sense of concern with and curiosity about someone or something, Synonym: involvement
- **S2:** the power of attracting or holding one’s interest (because it is unusual or exciting etc.), Synonym: interestingness
- **S3:** a reason for wanting something done, Synonym: sake
- **S4:** a fixed charge for borrowing money; usually a percentage of the amount borrowed
- **S5:** a diversion that occupies one’s time and thoughts (usually pleasantly), Synonyms: pastime, pursuit
- **S6:** a right or legal share of something; a financial involvement with something, Synonym: stake
- **S7:** (usually plural) a social group whose members control some field of activity and who have common aims, Synonym: interest group
Polysemy in WordNet

• Polysemous words are part of multiple synsets

• This is why relationships are defined between synsets, not words

• On average,
  – nouns have 1.24 senses (2.79 if excluding monosemous words)
  – verbs have 2.17 senses (3.57 if excluding monosemous words)

• Is Wordnet too fine grained?

Stats from: http://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html
Different sense = different translation

- Another way to define senses: if occurrences of the word have different translations, these indicate different sense

- Example interest translated into German
  
  - Zins: financial charge paid for load (Wordnet sense 4)
  - Anteil: stake in a company (Wordnet sense 6)
  - Interesse: all other senses

- Other examples might have distinct words in English but a polysemous word in German.
Word sense disambiguation (WSD)

- For many applications, we would like to disambiguate senses
  - we may be only interested in one sense
  - searching for chemical plant on the web, we do not want to know about chemicals in bananas

- Task: Given a polysemous word, find the sense in a given context

- Popular topic, data driven methods perform well
WSD as classification

- Given a word token in context, which sense (class) does it belong to?

- We can train a supervised classifier, assuming sense-labeled training data:
  - She pays 3% **interest/**INTEREST-MONEY on the loan.
  - He showed a lot of **interest/**INTEREST-CURIOSITY in the painting.
  - Playing chess is one of my **interests/**INTEREST-HOBBY.

- **SensEval** and later **SemEval** competitions provide such data
  - held every 1-3 years since 1998
  - provide annotated corpora in many languages for WSD and other semantic tasks
What kind of classifier?

Lots of options available:

• Naive Bayes (see Lecture 10)

• Maximum entropy model (see next lecture)

• Decision lists (see J&M, 20.2.2)

• Decision trees (see any ML textbook)
Naive Bayes for WSD

- Naive Bayes requires estimates of:
  - The prior probability of each class (sense)
  - The probability of each feature given each class

- These can be estimated from the training data.

- But what features to use? (Same question for other classifiers!)
Simple features

• Directly neighboring words
  – interest paid
  – rising interest
  – lifelong interest
  – interest rate
  – interest piqued

• Any content words in a 50 word window
  – pastime
  – financial
  – lobbied
  – pursued
More features

- Syntactically related words
- Syntactic role in sense
- Topic of the text
- Part-of-speech tag, surrounding part-of-speech tags

Of course, with NB we have the usual problem with correlated features... stay tuned for next lecture.
Evaluation

- **Extrinsic**: test as part of IR, QA, or MT system

- **Intrinsic**: evaluate classification accuracy or precision/recall against gold-standard senses

- **Baseline**: choose the most frequent sense (sometimes hard to beat)
Issues with WSD

- Not always clear how fine-grained the gold-standard should be

- Classifiers must be trained separately for each word
  - Hard to learn anything for infrequent or unseen words
  - Requires new annotations for each new word
  - Motivates unsupervised and semi-supervised methods (see J&M 20.5, 20.10)
Semantic roles

- Often we want to know *who did what to whom?*

- But the same event and participants can have different syntactic realizations:
  
  Sandy broke the glass.    vs.    The glass was broken by Sandy.
  She gave the boy a book.  vs.    She gave a book to the boy.

- Instead of focusing on syntax, consider the **semantic roles** (also called **thematic roles**) defined by each event.
## Commonly used thematic roles

<table>
<thead>
<tr>
<th>Role</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agent</td>
<td><em>The boy kicked his toy</em></td>
</tr>
<tr>
<td>Theme</td>
<td><em>The boy kicked his toy</em></td>
</tr>
<tr>
<td>Experiencer</td>
<td><em>The boy felt sad</em></td>
</tr>
<tr>
<td>Result</td>
<td><em>The girl built a shelf with power tools</em></td>
</tr>
<tr>
<td>Instrument</td>
<td><em>The girl built a shelf with power tools</em></td>
</tr>
<tr>
<td>Source</td>
<td><em>She came from home</em></td>
</tr>
</tbody>
</table>

- J&M give definitions and additional roles
Issues with thematic roles

• No universally agreed-upon set of roles

• Items with the “same” role (e.g., Instrument) may not behave quite the same

  Sandy opened the door with a key  The key opened the door
  Sandy ate the salad with a fork   *The fork ate the salad

• The two main NLP resources for thematic roles avoid these problems by defining very fine-grained roles:
  – Specific to individual verbs only (PropBank)
  – Specific to small groups of verbs (FrameNet)
Semantic role labeling

• The NLP task of identifying which words/phrases play which roles in an event.

• Supervised techniques similar to other classification tasks:
  – Training data from FrameNet or PropBank
  – Features are mostly related to syntactic structure and the particular words involved
  – Use one of many standard classifiers from machine learning

• Current research focuses on reducing the need for training data (e.g., to work on non-English languages)