Algorithms for Natural Language Processing

Lexical Semantics:
Word senses, relations, and classes

Nathan Schneider
(based on slides by Philipp Koehn and Sharon Goldwater)

7 February 2018
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**
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 (4)

• 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 (**hypernym**) 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/)

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

• 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 / He is a <em>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 <em>arm</em> 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 <em>church</em> sang a song</td>
</tr>
<tr>
<td>Complement Coercion</td>
<td>Begin, start, finish, try</td>
<td>John <em>began</em> 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, lock*, bicycle*, comb*, saw*</td>
<td>The hammer is heavy / She hammered the nail into the wall</td>
</tr>
<tr>
<td>Instance of an entity for kind</td>
<td>Tennis, soccer, cat, dog, class*, dinner*, chair*, table*</td>
<td>Tennis was invented in England / Tennis was fun today</td>
</tr>
<tr>
<td>Location / Place at location</td>
<td>Bench, land, floor, ground, box*, bottle*, jail*</td>
<td>The bench was made of pine / The coach benched the player</td>
</tr>
<tr>
<td>Object for placing at goal</td>
<td>Water, paint, salt, butter, frame*, dress*, oil*</td>
<td>The water is cold / He watered the plant</td>
</tr>
<tr>
<td>Object for taking from source</td>
<td>Milk, dust, weed, peel, pit*, skin*, juice*</td>
<td>The milk tastes good / He milked the cow</td>
</tr>
<tr>
<td>Material for artifact</td>
<td>Tin, iron, china, glass, linen*, rubber*, nickel*, fur*</td>
<td>Watch out for the broken glass / He filled the glass with water</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?

• 5 min. exercise: How many senses does the word interest have?
How many senses?
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!
Lexicography requires data
Lumping vs. Splitting

- For any given word, lexicographer faces the choice:
  - **Lump** usages into a small number of senses? or
  - **Split** senses to reflect fine-grained distinctions?
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
• **Synsets** ("synonym sets", effectively senses) are the basic unit of organization in WordNet.
  
  – Each synset is specific to nouns (.n), verbs (.v), adjectives (.a, .s), or adverbs (.r).
  – Synonymous words belong to the same synset: \( \text{car}^1 \) (car.n.01) = \{car, auto, automobile\}.
  – Polysemous words belong to multiple synsets: \( \text{car}^1 \) vs. \( \text{car}^4 \) = \{car, elevator car\}. Numbered roughly in descending order of frequency.

• Synsets are organized into a **network** by several kinds of relations, including:
  
  – **Hypernymy** (Is-A): hyponym \{ambulance\} is a kind of hypernym \( \text{car}^1 \)
  – **Meronymy** (Part-Whole): meronym \{air bag\} is a part of holonym \( \text{car}^1 \)
Visualizing WordNet
Using WordNet

- NLTK provides an excellent API for looking things up in WordNet:
  ```python
  >>> from nltk.corpus import wordnet as wn
  >>> wn.synsets('car')
  [Synset('car.n.01'), Synset('car.n.02'),
   Synset('car.n.03'),
   Synset('car.n.04'), Synset('cable_car.n.01')]
  >>> wn.synset('car.n.01').definition()
  u'a motor vehicle with four wheels; usually propelled by an internal combustion engine'
  >>> wn.synset('car.n.01').hypernyms()
  [Synset('motor_vehicle.n.01')]
  ```

- (WordNet uses an obscure custom file format, so reading the files directly is not recommended!)
Polysemy and Coverage in WordNet

- Online stats:
  - 155k unique strings, 118k unique synsets, 207k pairs
  - nouns have an average 1.24 senses (2.79 if excluding monosemous words)
  - verbs have an average 2.17 senses (3.57 if excluding monosemous words)

- Too fine-grained?

- WordNet is a snapshot of the English lexicon, but by no means complete.
  - E.g., consider multiword expressions (including noncompositional expressions, idioms): hot dog, take place, carry out, kick the bucket are in WordNet, but not take a break, stress out, pay attention
  - Neologisms: hoodie, facepalm
  - Names: Microsoft
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.
SemCor in NLTK

In the SemCor corpus, words and multiword units are annotated with their part of speech:

```python
>>> semcor.tagged_sents()[0]
[Tree('DT', ['The']),
 Tree('NNP', ['Fulton', 'County', 'Grand', 'Jury']),
 Tree('VB', ['said']),
 Tree('NN', ['Friday']),
 Tree('DT', ['an']),
 Tree('NN', ['investigation']),
 Tree('IN', ['of']),
 Tree('NN', ['Atlanta']), ...]
```

Each sentence consists of a series of chunks with 1 or more words.

In the tagset used in SemCor, DT = determiner, NN = common noun, NNP = proper noun, VB = verb, etc.
In addition, nouns, verbs, adjectives, and adverbs are annotated with a WordNet synset:

```python
>>> semcor.tagged_sents(tag='sem')[0]
[['The'],
  Tree(Lemma('group.n.01.group'), [Tree('NE',
    ↦ ['Fulton', 'County', 'Grand', 'Jury'])]),
  Tree(Lemma('state.v.01.say'), ['said']),
  Tree(Lemma('friday.n.01.Friday'), ['Friday']),
  ['an'],
  Tree(Lemma('probe.n.01.investigation'),
    ↦ ['investigation']),
  ['of'],
  Tree(Lemma('atlanta.n.01.Atlanta'), ['Atlanta'])],
```

Note that *Fulton County Grand Jury* is a named entity (NE) not in WordNet, so it receives a high-level synset group.n.01.
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
Semantic Classes

• Other approaches, such as named entity recognition and supersense tagging, define coarse-grained semantic categories like PERSON, LOCATION, ARTIFACT.

• Like senses, can disambiguate: APPLE as ORGANIZATION vs. FOOD.

• Unlike senses, which are refinements of particular words, classes are typically larger groupings.

• Unlike senses, classes can be applied to words/names not listed in a lexicon.
Named Entity Recognition

• Recognizing and classifying proper names in text is important for many applications. A kind of information extraction.

• Different datasets/named entity recognizers use different inventories of classes.
  – Smaller: PERSON, ORGANIZATION, LOCATION, MISCELLANEOUS
  – Larger: sometimes also PRODUCT, WORK_OF_ART, HISTORICAL_EVENT, etc., as well as numeric value types (TIME, MONEY, etc.)

• NER systems typically use some form of feature-based sequence tagging, with features like capitalization being important.

• Lists of known names called gazetteers are also important.
Supersenses

- As a practical measure, WordNet noun and verb synset entries were divided into multiple files ("lexicographer files") on a semantic basis.

- Later, people realized these provided a nice inventory of high-level semantic classes, and called them **supersenses**.

- Supersenses offer an alternative, broad-coverage, language-neutral approach to corpus annotation.
### Supersenses

<table>
<thead>
<tr>
<th>N:Tops</th>
<th>N:Object</th>
<th>V:Cognition</th>
</tr>
</thead>
<tbody>
<tr>
<td>N:Act</td>
<td>N:Person</td>
<td>V:Communication</td>
</tr>
<tr>
<td>N:Animal</td>
<td>N:Phenomenon</td>
<td>V:Competition</td>
</tr>
<tr>
<td>N:Artifact</td>
<td>N:Plant</td>
<td>V:Consumption</td>
</tr>
<tr>
<td>N:Attribute</td>
<td>N:Possession</td>
<td>V:Contact</td>
</tr>
<tr>
<td>N:Body</td>
<td>N:Process</td>
<td>V:Creation</td>
</tr>
<tr>
<td>N:Cognition</td>
<td>N:Quantity</td>
<td>V:Emotion</td>
</tr>
<tr>
<td>N:Communication</td>
<td>N:Relation</td>
<td>V:Motion</td>
</tr>
<tr>
<td>N:Event</td>
<td>N:Shape</td>
<td>V:Perception</td>
</tr>
<tr>
<td>N:Feeling</td>
<td>N:State</td>
<td>V:Possession</td>
</tr>
<tr>
<td>N:Food</td>
<td>N:Substance</td>
<td>V:Social</td>
</tr>
<tr>
<td>N:Group</td>
<td>N:Time</td>
<td>V:Stative</td>
</tr>
<tr>
<td>N:Location</td>
<td></td>
<td>V:Body</td>
</tr>
<tr>
<td>N:Motive</td>
<td></td>
<td>V:Change</td>
</tr>
</tbody>
</table>

- The **supersense tagging** goes beyond NER to cover all nouns and verbs.
• In order to support technologies like question answering, we need ways to reason computationally about meaning. **Lexical semantics** addresses meaning at the word level.

  – Words can be ambiguous (**polysemy**), sometimes with related meanings, and other times with unrelated meanings (**homonymy**).
  – Different words can mean the same thing (**synonymy**).

• Computational lexical databases, notably WordNet, organize words in terms of their meanings.

  – **Synsets** and relations between them such as hypernymy and meronymy.
• **Word sense disambiguation** is the task of choosing the right sense for the context.
  
  – Classification with contextual features
  – Relying on dictionary senses has limitations in granularity and coverage

• **Semantic classes**, as in NER and supersense tagging, are a coarser-grained representation for semantic disambiguation and generalization.