Algorithms for Natural Language Processing
Lexical Semantics: Word senses, relations, and classes

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

3 February 2022

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 (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).

Pattern	Participating Senses	Example Sentences
Animal for fur	Mink, chinchilla, rabbit, beaver, raccoon*, alpaca*, crocodile*	The *mink* drank some water / She likes to wear *mink*
Animal/Object for personality	Chicken, sheep, pig, snake, star*, rat*, doll*	The *chicken* drank some water / He is a *chicken*
Animal for meat	Chicken, lamb, fish, shrimp, salmon*, rabbit*, lobster*	The chicken drank some water / The *chicken* is tasty
Artifact for activity	Shower, bath, sauna, baseball,	The *shower* was leaking / The *shower* was relaxing
Body part for object part	Arm, leg, hand, face, back*, head*, foot*, shoulder*, lip*,	John's *arm* was tired / The *arm* was reupholstered
Building for people	Church, factory, school, airplane,	The *church* was built 20 years ago / The *church* sang a song
Complement Coercion	Begin, start, finish, try	John *began* reading the book / John *began* the book
Container for contents	Bottle, can, pot, pan, bowl*, plate*, box*, bucket*	The *bottle* is made of steel / He drank half of the *bottle*
Word for question	Price, weight, speed	The *price* of the coffee was low / John asked the *price* of the coffee

Pattern	Participating Senses	Example Sentences
Figure for Ground	Window, door, gate, goal	The window is broken / The cat walked through the window
Grinding	Apple, chair, fly	The apple was tasty / There is apple all over the table
Instrument for action	Hammer, brush, shovel, tape, lock*, bicycle*, comb*, saw*	The hammer is heavy / She hammered the nail into the wall
Instance of an entity for kind	Tennis, soccer, cat, dog, class*, dinner*, chair*, table*	Tennis was invented in England / Tennis was fun today
Location / Place at location	Bench, land, floor, ground, box*, bottle*, jail*	The bench was made of pine / The coach benched the player
Object for placing at goal	Water, paint, salt, butter, frame*, dress*, oil*	The water is cold / He watered the plant.
Object for taking from source	Milk, dust, weed, peel, pit*, skin*, juice*	The milk tastes good / He milked the cow
Material for artifact	Tin, iron, china, glass, linen*, rubber*, nickel*, fur*	Watch out for the broken glass / He filled the glass with water
Occupation for role in action	Boss, nurse, guard, tutor	My boss is nice / He bossed me around

Pattern	Participating Senses	Example Sentences
Place for an event	Vietnam, Korea, Waterloo, Iraq	It is raining in *Vietnam* / John was shot during *Vietnam*
Place for an institution	White House, Washington, Hollywood, Pentagon, Wall Street*, Supreme Court	The *White House* is being repainted / The *White House* made an announcement
Plant for food or material	Corn, broccoli, coffee, cotton, lettuce*, eggs*, oak*, pine*	The large field of *corn* / The *corn* is delicious
Portioning	Water, beer, jam	She drank some *water* / She bought three *waters*
Publisher for product	Newspaper, magazine, encyclopedia, Wall Street Journal*, New York Times*,	The *newspaper* is badly printed / The *newspaper* fired three employees
Artist for product	Writer, artist, composer, Shakespeare, Dickens*, Mozart*, Picasso*	The *writer* drank a lot of wine / The *writer* is hard to understand
Object for contents	Book, CD, DVD, TV*, magazine*, newspaper*	The heavy, leather- bound *book* / The *book* is funny.
Visual Metaphor	Beam, belt, column, stick, bug*, leaf*	Most of the weight rests on the *beam* / There was a *beam* of light

How many senses?

- 5 min. exercise: How many senses does the word interest have?

How many senses?

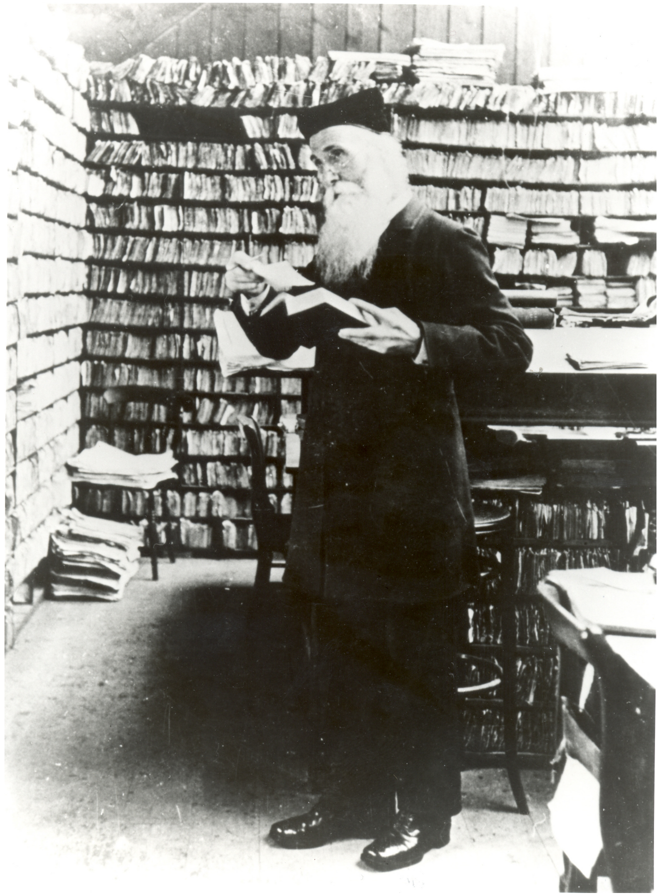
How many senses?

- 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

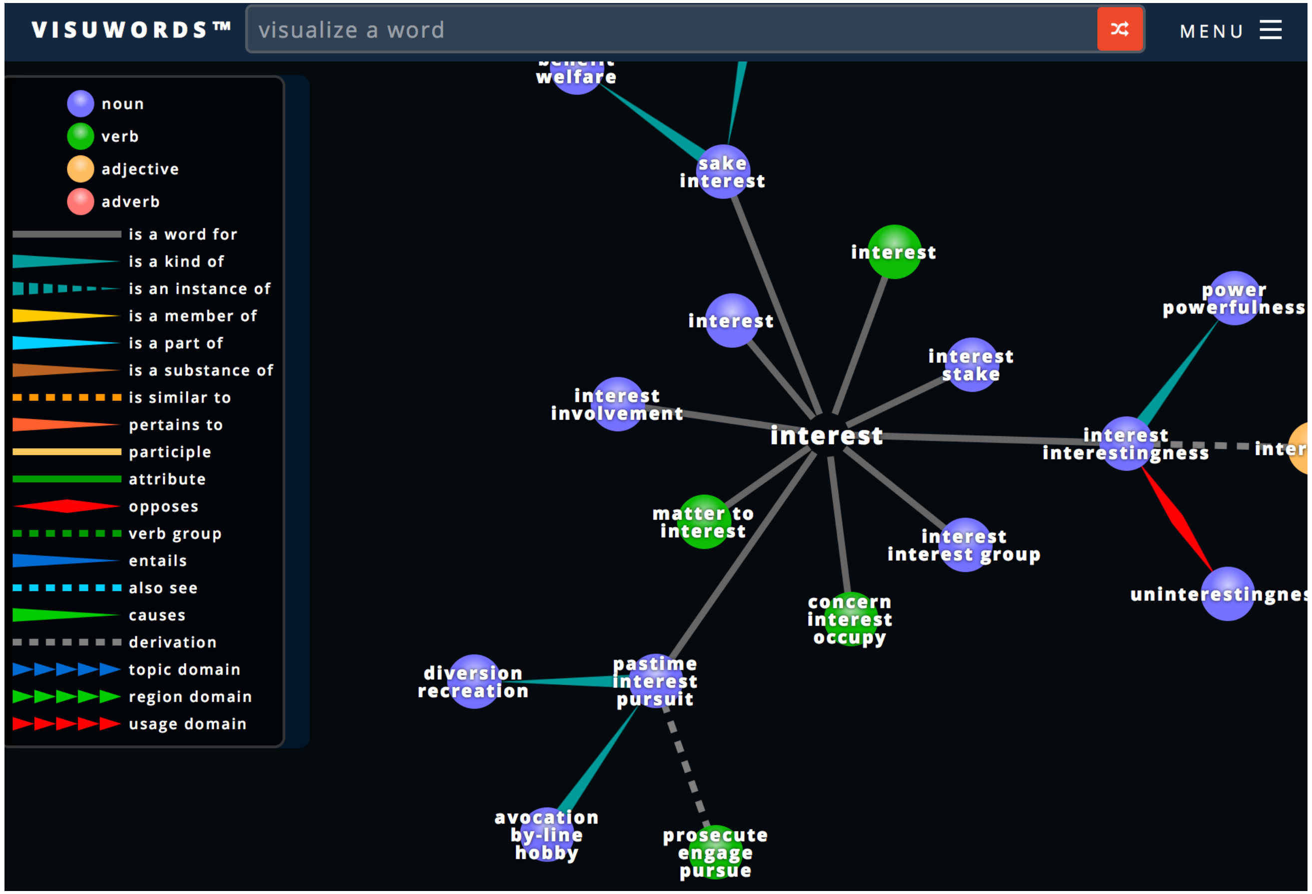
- S1: a sense of concern with and curiosity about someone or something, Synonym: involvement
- S2: the power of attracting or holding one's interest (because it is unusual or exciting etc.), Synonym: interestingness
- S3: a reason for wanting something done, Synonym: sake
- S4: a fixed charge for borrowing 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 and Relations in WordNet

- **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: car^1 (`car.n.01`) $= \{\mathrm{car},\mathrm{auto},\mathrm{automobile}\}$.
  - Polysemous words belong to multiple synsets: car^1 vs. $\mathrm{car}^4 = \{\mathrm{car},\mathrm{elevator\ car}\}$. Numbered roughly in descending order of frequency.

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

Visualizing WordNet



Using WordNet

- NLTK provides an excellent API for looking things up in WordNet:

```
>>> 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 exluding monosemous words)
  - verbs have an average 2.17 senses (3.57 if exluding 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**:

```
>>> 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.

SemCor in NLTK

In addition, nouns, verbs, adjectives, and adverbs are annotated with a **WordNet synset**:

```
>>> 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

N:TOPS	N:OBJECT	V:COGNITION
N:ACT	N:PERSON	V:COMMUNICATION
N:ANIMAL	N:PHENOMENON	V:COMPETITION
N:ARTIFACT	N:PLANT	V:CONSUMPTION
N:ATTRIBUTE	N:POSSESSION	V:CONTACT
N:BODY	N:PROCESS	V:CREATION
N:COGNITION	N:QUANTITY	V:EMOTION
N:COMMUNICATION	N:RELATION	V:MOTION
N:EVENT	N:SHAPE	V:PERCEPTION
N:FEELING	N:STATE	V:POSSESSION
N:FOOD	N:SUBSTANCE	V:SOCIAL
N:GROUP	N:TIME	V:STATIVE
N:LOCATION	V:BODY	V:WEATHER
N:MOTIVE	V:CHANGE	

- The **supersense tagging** goes beyond NER to cover all nouns and verbs.

Summary (1)

- 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.

Summary (2)

- **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.