Orientation

- So far, we have focused on linguistics, models and algorithms for:
  - Words and sequences
  - Syntactic structure

- We also motivated syntactic structure with reference to meaning (specifically, compositional semantics).

- But we haven’t talked much about meaning itself.

- So, let’s get started! For the rest of the course:
  - semantics of words, sentences, and documents
  - also lectures about the bigger picture: data, annotation, evaluation, and real issues in research (including guest lectures).

Meaning

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

- But how do we know if we succeeded?

Eliza

A famous computer program from 1969 shows people can be easily 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: [https://www.masswerk.at/elizabot/](https://www.masswerk.at/elizabot/)
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 (QA for short)

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 (later on)
- 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
    · Not just different parts of speech
    · But also different (senses) for the same PoS
  – we need to be able to disambiguate between them

Example Question (3)

• Question
  Where did Theresa May go on vacation?

• Text available to the machine
  Theresa May spent her holiday in Cornwall

• What is hard?
  – different 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?
  – one word can refer to a subclass (hyponym) or superclass (hypernym) of the concept referred to by another word
  – we need to have database of such A is-a-kind-of 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 lots of facts
  - we need to do inference
    - a problem for sentential, not lexical, semantics

---

**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)
  - '*' is for words where the non-literal reading is a bit harder to get without some context

---

**WordNet**

- Some of these problems can be solved with a good ontology.

- **WordNet** (for English: see http://wordnet.princeton.edu/) is a hand-built ontology containing 117,000 synsets: sets of synonymous words.

- 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)
Another name for one of those

- Instance of an entity for kind is a kind of abstraction
- So common we barely notice it
- Some examples, using the call sign of an airplane flight:

  EZY386 will depart from gate E17 at 2010 [announcement]
  Just arrived on EZY386 [text message]
  EZY386 flies from Stansted to Avalon
  EZY386 is easyJet’s 3rd most popular flight to Avalon
  I prefer EZY386 to EZY387
  EZY386 has an 102% on-time record
  EZY386 was cancelled yesterday
  EZY386 was delayed because of a problem with one of its engines

How many senses?

- How many senses does the noun 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!

Henry S. Thompson  Word senses and relations  18

Henry S. Thompson  Word senses and relations  19
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: (usu. 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, that’s evidence for multiple senses

• Example interest translated into German
  – Zins: financial charge paid for loan (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

• As we’ve seen, this can be formulated as a classification task.
**WSD as classification**

- Given word token in context, which sense (class) is it?
- Just train a classifier, if we have 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

**Evaluation of WSD**

- 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)

**Classifiers for WSD**

As usual, lots of options:

- We've discussed Naive Bayes, logistic regression, neural nets; many others available...

For many of these, need to choose relevant features. For example,

- Directly neighboring words:
  - interest paid, rising interest, lifelong interest, interest rate

- Any content words in a 50 word window
  - pastime, financial, lobbied, pursued

- Syntactically related words, topic of the text, part-of-speech tag, surrounding part-of-speech tags, etc ...

**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 JM3 C.7-C.8: optional)
Summary

• Aspects of lexical semantics:
  – Word senses, and methods for disambiguating.
  – Lexical semantic relationships, like synonymy, hyponymy, and meronymy.
  – Disambiguation: Different senses need to be distinguished

• Resources that provide annotated data for lexical semantics:
  – WordNet (senses, relations)
  – SensEval datasets