WSTA Lecture 17
Word Sense Disambiguation

1. Word meaning
2. WordNet Lexical Resource
3. Word sense disambiguation

Slide credits: Steven Bird
Polysemy and Homonymy

- **Polysemy:**
  - e.g. *plain* - clear, undecorated, unattractive, level area of land

- **Homonymy:**
  - Homograph: different words with same orthography
    - e.g. *dove* - dive into water, white bird
    - e.g. *deal* – distribute cards (verb), an agreement (noun)
  - Homophone: different words with same sound
    - e.g. *see, sea*
    - e.g. French: *vert, verre, vers, ver* (green, glass, towards, worm)
What is a word’s meaning?

- Meaning hard to pin down
  - as referent: words stand for objects/concepts in the world
  - as mental image: words convey conceptual grouping
  - as context: meaning as the set of contexts in which they can occur
  - as dictionary entry: let a lexicographer decide

- Dictionary definitions
  - all definitions ultimately circular: dictionaries just give paraphrases
  - what meaning is really contained in a dictionary entry?
    - cf. bilingual dictionary, giving foreign translations

- Explicit semantic databases
  - SQL statements, first order logic, ...
Hypernyms and Hyponyms

- Hypernym/Hyponym = generic/specific
- e.g. fork is a kind of cutlery
- “fork” is a hyponym of “cutlery”
- “cutlery” is a hypernym of “fork”
- Induces forest structure on our set of words
- Also gives a measure of semantic distance
Holonyms and Meronyms

- Holonym/Meronym (whole/part): 3 subtypes:
  1. Part: bone is part of arm
  2. Member: arm is member of body
  3. Substance: bone is substance of horn
Other Lexical Relationships

- **Synonym/Antonym:**
  - same vs complementary referential meanings

- **Hypernym/Troponym:**
  - *walk* is a hypernym of *stroll*
    - To walk *is one way to stroll*
  - *stroll* is a troponym of *walk*
    - To stroll *is a particular way to walk*

- **Entails:**
  - *Walking* entails *stepping*
  - *Snoring* entails *sleeping*

- Many more lexical relationships exist...
Data for Exploring Lexical Semantics

- **Thesaurus:**
  - Synonyms and Antonyms

- **Wordnet:**
  - Synonyms and Antonyms
  - Hypernyms and Hyponyms, Hypernyms and Troponyms
  - Meronyms and Holonyms
  - Entails
WordNet: Introduction

- A lexical database
  - Inspired by psycholinguistic theories of human lexical memory
  - Establishes a massive network of lexical items and lexical relationships
  - English wordnet
    - Four categories: noun, verb, adjective, adverb
    - Nouns: 120,000; Verbs: 12,000; Adjectives: 21,000; Adverbs: 4,000
  - Wordnet in other languages [www.globalwordnet.org]
    - Wordnets exist or are in preparation for: Afrikaans Albanian Arabic Assamese Bantu Basque Bengali Bodo Bulgarian Burmese Catalan Chinese Croatian Czech Danish Dutch English Estonian Finnish French French German Greek Gujarati Hebrew Hindi Hungarian Icelandic Indonesian Irish Italian Japanese Kannada Kashmiri Konkani Korean Kurdish Lao Latin Latvian Macedonian Malayalam Malaysian Maltese Marathi Meitei Moldavian Mongolian Nepali Norwegian Oriya Persian Polish Portuguese Punjabi Romanian Russian Sanskrit Serbian Sinhala Slovenian Spanish Swedish Tamil Telugu Thai Turkish Urdu Vietnamese
Synonym Sets - Synsets

- Words are ambiguous
  - e.g. “fork” in earlier slide
  - the different senses participate in different lexical relations

- Nodes in Wordnet represent “synonym sets”, or synsets.
  - e.g. {chump, fish, fool, gull, mark, patsy, fall guy, sucker, schlemiel, shlemiel, soft touch, mug} (a person who is gullible and easy to take advantage of)

- Applications:
  - Overcome limitations in other data (e.g. NLU)
  - Implement selectional restrictions (use WordNet categories on grammar productions, e.g., can only “eat” with a certain sense of “fork”)

NLTK WordNet Interface

```python
>>> import nltk

>>> nltk.corpus.wordnet.synsets('fork')

[Synset('fork.n.01'), Synset('branching.n.01'), Synset('fork.n.03'),
  Synset('fork.n.04'), Synset('crotch.n.02'), Synset('pitchfork.v.01'),
  Synset('fork.v.02'), Synset('branch.v.02'), Synset('fork.v.04')]

>>> nltk.corpus.wordnet.synsets('fork')[1].lemma_names()

[u'branching', u'ramification', u'fork', u'forking']

>>> nltk.corpus.wordnet.synsets('fork')[1].definition()

u'the act of branching out or dividing into branches'

>>> nltk.corpus.wordnet.synsets('fork')[1].hypernyms()

[Synset('division.n.03')]
```
Word Sense Disambiguation

- Applications:
  - semantic analysis, machine translation, information retrieval, homograph resolution in text-to-speech, sentence boundary detection, restoring accents and capitals, automatic diacritics while typing

- Example

- Definition and application

- Training data – SENSEVAL, SEMCOR

- Methods for robust WSD
  - Supervised classifiers
  - Semisupervised method
  - Unsupervised clustering

- Issues
Example

- “The US puts a new face on the chase for Saddam”
  - US/n
    - First person plural inclusive pronoun
    - Abbrev. The United States of America
  - Put/v.t.
    - Transfer to a specified place
    - Express in words
    - Propel from hand with pushing motion
  - Put/n
    - Throw of shot
    - Option of selling stock at a certain date
Example (cont)

- "The US puts a new face on the chase for Saddam”
  - new/a
    - Invented, discovered, previously unknown
    - Fresh, further, additional
    - Different, changed, substituted for old
    - Of recent growth, origin or manufacture
  - new/adv ...
  - face/n
    - Front of head
    - Expression, grimace
    - Aspect (on the face of it...)
    - ....
The Context

- An avalanche of competing interpretations
  - 5,760 different sense combinations in example
  - as sentences grow
  - exponential growth of interpretations

- *disambiguate two or more semantically distinct forms which have been conflated into the same representation in some medium* (Yarowsky)
Methods for Robust WSD

- Simple n-gram methods won't work - why not?
  - disambiguating context
  - the tag on the word

- Popular approaches
  - Feature based classifiers
  - Unsupervised methods

- Supervised training data
  - lexical sample: many labelled examples for single polysemous word
  - all-words: sense annotations for all ambiguous words in documents
Training Data for WSD
SENSEVAL ‘lexical sample’

\begin{verbatim}
<instance id="hard-a.br-a06:">
<answer instance="hard-a.br-a06:" senseid="HARD1"/>
<context>
</context>
</instance>
\end{verbatim}

- Competition data from SENSEVAL events
  - SENSEVAL-1: 35 words, 12k instances
  - SENSEVAL-2: 73 words, 12k instances
  - SENSEVAL-3, SEMEVAL…
  - Many datasets available http://www.d.umn.edu/~tpederse/data.html
Training Data for WSD SEMCOR ‘all-words’

Semcor

- 352 documents from Brown corpus manually tagged for WordNet senses
- Several versions available from http://web.eecs.umich.edu/~mihalcea/downloads.html
Sense tagged corpora in NLTK

```python
>>> from nltk.corpus import senseval, semcor
>>> senseval.fileids()
['serve.pos', 'interest.pos', 'hard.pos', 'line.pos']
>>> senseval.instances('serve.pos')
[SensevalInstance(word=u'serve-v', position=42, context=[('some', 'DT'), ('tart', 'JJ'), ('fruits', 'NNS'), ...], senses=('SERVE10',)), ...]
>>> semcor.tagged_sents(tag='sem')[0]
[u'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']), [u'an'], Tree(Lemma('probe.n.01.investigation'), ['investigation']), [u'of'], Tree(Lemma('atlanta.n.01.Atlanta'), ['Atlanta']), [u''s''], Tree(Lemma('late.s.03.recent'), ['recent']), Tree(Lemma('primary.n.01.primary_election'), ['primary', 'election']), Tree(Lemma('produce.v.04.produce'), ['produced']), [u'''], [u'no'], Tree(Lemma('evidence.n.01.evidence'), ['evidence']), [u'''], [u'that'], [u'any'], Tree(Lemma('abnormality.n.04.irregularity'), ['irregularities']), Tree(Lemma('happen.v.01.take_place'), ['took', 'place']), [u'.']]
Inputs: Feature Vectors 1

- What contextual have good predictive value?
- Local context
  - E.g. Reid saw me looking at the iron bars .
    NNP VBD PRP VBG IN DT NN NNS .
  - local POS around the word
    - P₀ = NNS; P₋₁ = NN; P₁ = .; P₋₂ = DT; ...
  - unigrams and collocations
    - nearby word ‘iron’; nearby word ‘gin’; ...
    - ‘the iron X’; ‘iron X .’; ‘the __ X’
Inputs: Feature Vectors 2

- **Syntactic relations**
  - E.g., he *turned* his attention to the workbench
    - subject = he/PRP; object = attention/NN, active tense
  - E.g., he turned his *attention* to the workbench
    - head = turned/VBD; active tense; head to the left
  - E.g., the modern tram is a *green* machine.
    - head = machine/NN

- **Combine all these inputs in a classifier**
WSD as Classification

- **Supervised classification**
  - given feature vectors for each occurrence of our word in context
  - and its label (e.g., bass/fish)
  - use to train a classifier, e.g.,
    - logistic regression
    - support vector machine
    - naïve Bayes, k-NN, etc

- **Evaluate performance on test data**
  - baseline = most frequent sense (hard to beat!)
  - measure accuracy, precision and recall

- **See Lee & Ng (2002) for overview**
  - best accuracy for SVMs vs several other classifiers compared on SENSEVAL-1 and 2
Other approaches

• Lesk’s dictionary based method
  • one of the earliest approaches
  • look for terms from the dictionary definition
    • difficult, hard -- (not easy; requiring great physical or mental effort to accomplish or comprehend or endure; "a difficult task"; "nesting places on the cliffs are difficult of access"; "difficult times"; "a difficult child"; "found himself in a difficult situation"; "why is it so hard for you to keep a secret?")
  • e.g., search for effort, endure, easy, etc in the context
  • no need for explicit supervision, but depends on comprehensive definitions
Yarowsky ‘boot-strap’ algorithm

- Assumes two properties
  - One sense per collocation: Local contexts (collocations) highly informative of word sense (this is exploited by the classifier’s features)
  - One sense per discourse: Documents tend to use only one sense of an ambiguous word (holds 99% of the time)

- Example
  - \(d_1\) living... close-up studies of plant life and natural... living... many dangers to plant and animal life... ???... cell types found in the plant kingdom are...
  - \(d_2\) factory... discovered at a St. Louis manufacturing plant... ???... computer disk drive plant located in...

- OSPC highly informative context words: *life, manufacturing*
- OSPD resolves the difficult instances based on easy ones in same document
Yarowsky algorithm

- Proceed as follows, for a given focus word
  1. Sense label a small “seed” collection, to use as training
     - automatically from dictionary
     - dream up one or two good examples per sense
     - annotate a few corpus examples
  2. Repeat
     1. learn a classifier on the training set
     2. predict senses for the remaining unlabelled text
     3. add the most confident predictions to training, excluding documents that don’t obey the ‘one-sense-per-document’ heuristic
  3. Apply final classifier to test data
     - use OSPD to vote for the best sense
- Rivals supervised classification (Yarowsky, 1995)
Unsupervised 1

- Represent context of target word 'suit' as vectors $\langle F_1, F_2, \ldots, F_v \rangle$
  
  - The suit against the union was successful and many workers lost their homes to pay off the judgment. [1]
  
  - Mantle, more concerned with dress, buys his suits four at a time at Neiman-Marcus in Dallas and pays as much as $250 each. [2]
Unsupervised 2

- Cluster context vectors
  - Cosine, Euclidian distance
  - Hierarchical clustering, K-means, EM
  - Dimensionality reduction, e.g., latent semantic indexing (LSI)
Problems

- Knowledge acquisition bottleneck
  - Supervised methods need marked up data

- Performance measurement?
  - Non-uniform confusability

- Incorporation into larger tasks
  - e.g. parsing, translation, document retrieval

- What is a word sense?
  - Dictionaries as sense inventories
  - Bilingual texts as sense inventories
  - Dependent on contextual usage

- Dealing with rapid language change...
Readings

One of the following:

MS 7.1, 7.3-7.4: Word Sense Disambiguation

JM 20.1-20.4.1, 20.6-20.7: Word Sense Disambiguation

Optional, for more recent overview of field


Optional, for more details on lexical semantics:

JM 19.1-19.3 Lexical Semantics