CS 6120/CS4120: Natural Language Processing
Instructor: Prof. Lu Wang
College of Computer and Information Science
Northeastern University
Webpage: www.ccs.neu.edu/home/luwang

Logistics
• Quiz solution will be released by the end of each Friday.
• Make-up can be taken place at TA’s office hours for that week.

Today’s Outline
• Word Senses and Word Relations
• Word Similarity
• Word Sense Disambiguation

Terminology: lemma and wordform
• A lemma or citation form
  • Same stem, part of speech, rough semantics
• A wordform
  • The inflected word as it appears in text

<table>
<thead>
<tr>
<th>Wordform</th>
<th>Lemma</th>
</tr>
</thead>
<tbody>
<tr>
<td>banks</td>
<td>bank</td>
</tr>
<tr>
<td>sung</td>
<td>sing</td>
</tr>
<tr>
<td>duermes</td>
<td>dormir</td>
</tr>
</tbody>
</table>

Lemmas have senses
• One lemma “bank” can have many meanings:
  Sense 1: “a bank can hold the investments in a custodial account…”
  Sense 2: “…as agriculture burgeons on the east bank, the river will shrink even more”
• Sense (or word sense)
  • A discrete representation of an aspect of a word’s meaning.
  • The lemma bank here has two senses

Homonymy
Homonyms: words that share a form but have unrelated, distinct meanings:
• bank₁: financial institution, bank₂: sloping land
• bat₁: club for hitting a ball, bat₂: nocturnal flying mammal
1. Homographs (bank/bank, bat/bat)
2. Homophones:
   1. Write and right
   2. Piece and peace
Homonymy causes problems for NLP applications

- Information retrieval
- "bat care"
- Machine Translation
  - bat: murciélago (animal) or bate (for baseball)
- Text-to-Speech
  - baas (stringed instrument) vs. baas (fish)

Polysemy

1. The bank was constructed in 1875 out of local red brick.
2. I withdrew the money from the bank

- Are those the same sense?
  - Sense 2: "A financial institution"
  - Sense 1: "The building belonging to a financial institution"

- A polysemous word has related meanings
- Most non-rare words have multiple meanings

Metonymy or Systematic Polysemy: A systematic relationship between senses

- Lots of types of polysemy are systematic
  - School, university, hospital
  - All can mean the institution or the building.
- A systematic relationship:
  - Building Organization
- Other such kinds of systematic polysemy:
  - Author (Jane Austen wrote Emma)
  - Work of Author (I love Jane Austen)
  - Tree (Plums have beautiful blossoms)
  - Fruit (I ate a preserved plum)

How do we know when a word has more than one sense?

- The “zeugma” test: Two senses of serve?
  - Which flights serve breakfast?
  - Does Lufthansa serve Philadelphia?
  - Does Lufthansa serve breakfast and San Jose?
  - Since this conjunction sounds weird,

- we say that these are two different senses of "serve"

Synonyms

- Word that have the same meaning in some or all contexts.
  - filbert / hazelnut
  - couch / sofa
  - big / large
  - automobile / car
  - vomit / throw up
  - Water / H₂O
- Two lexemes are synonyms
  - if they can be substituted for each other in all situations
  - if so they have the same propositional meaning

- But there are few (or no) examples of perfect synonymy.
  - Even if many aspects of meaning are identical
  - Skill may not preserve the acceptability based on notions of politeness, slang, register, genre, etc.

- Example:
  - Water / H₂O
  - Big/large
  - Brave/courageous
Synonymy is a relation between senses rather than words

- Consider the words big and large
- Are they synonyms?
- How big is that plane?
- Would I be flying on a large or small plane?
- How about here:
  - Miss Nelson became a kind of big sister to Benjamin.
  - Miss Nelson became a kind of large sister to Benjamin.
- Why?
  - big has a sense that means being older, or grown up
  - large lacks this sense

Antonyms

- Senses that are opposites with respect to one feature of meaning
- Otherwise, they are very similar!
- dark/light short/long fast/slow rise/fall hot/cold up/down
- More formally: antonyms can
  - define a binary opposition or be at opposite ends of a scale
- Be reversives:
  - rise/fall, fast/slow

Hyponymy and Hypernymy

- One sense is a hyponym of another if the first sense is more specific, denoting a subclass of the other
  - car is a hyponym of vehicle
  - mango is a hyponym of fruit
- Conversely hypernym/superordinate ("hyper is super")
  - vehicle is a hypernym of car
  - fruit is a hypernym of mango

Hyponymy more formally

- Extensional:
  - The class denoted by the superordinate extensionally includes the class denoted by the hyponym
- Entailment:
  - A sense A is a hyponym of sense B if being an A entails being a B
- Hyponymy is usually transitive
  - (A hypo B and B hypo C entails A hypo C)
- Another name: the IS-A hierarchy
  - A IS-A B (or A ISA B)
  - B subsumes A

Hyponyms and Instances

- WordNet has both classes and instances.
- An instance is an individual, a proper noun that is a unique entity
  - San Francisco is an instance of city
  - But city is a class
  - city is a hyponym of municipality...location...

Meronymy

- The part-whole relation
  - A leg is part of a chair; a wheel is part of a car.
  - Wheel is a meronym of car, and car is a holonym of wheel.
WordNet 3.0

- A hierarchically organized lexical database
- On-line thesaurus + aspects of a dictionary
  - Some other languages available or under development
    - Arabic, Finnish, German, Portuguese...

<table>
<thead>
<tr>
<th>Category</th>
<th>Unique Strings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noun</td>
<td>117,798</td>
</tr>
<tr>
<td>Verb</td>
<td>11,529</td>
</tr>
<tr>
<td>Adjective</td>
<td>22,479</td>
</tr>
<tr>
<td>Adverb</td>
<td>4,481</td>
</tr>
</tbody>
</table>

How is “sense” defined in WordNet?

- The synset (synonym set), the set of near-synonyms, instantiates a sense or concept, with a gloss
- Example: chump as a noun with the gloss:
  “a person who is gullible and easy to take advantage of”
- This sense of “chump” is shared by 9 words:
  - chump, fool', gull', mark', patsy', fall guy', sucker', soft touch', mug'
  - Each of these senses have this same gloss
  - (Not every sense; sense 2 of gull is the aquatic bird)

WordNet Noun Relations

<table>
<thead>
<tr>
<th>Relation</th>
<th>Also Called</th>
<th>Definition</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypernym</td>
<td>Superordinate</td>
<td>From concepts to superordinates</td>
<td>bass → melody</td>
</tr>
<tr>
<td>Hypnym</td>
<td>Subordinate</td>
<td>From concepts to subconcepts</td>
<td>snore → sleep</td>
</tr>
<tr>
<td>Instance Hypernym</td>
<td>Instance</td>
<td>From instances to their concepts</td>
<td>pluck → pluck</td>
</tr>
<tr>
<td>Instance Ham-Instance</td>
<td>Instance</td>
<td>From concepts to concept instances</td>
<td>compose → Bach</td>
</tr>
<tr>
<td>Member Holonym</td>
<td>Has-Member</td>
<td>From members to their groups</td>
<td>operate → engine</td>
</tr>
<tr>
<td>Part Holonym</td>
<td>Has-Part</td>
<td>From parts to wholes</td>
<td>course → meal</td>
</tr>
<tr>
<td>Substance Holonym</td>
<td>From substances to their subparts</td>
<td>water → swamp</td>
<td></td>
</tr>
<tr>
<td>Substance</td>
<td>Holonym</td>
<td>From parts of substances to whole</td>
<td>book → library</td>
</tr>
<tr>
<td>Antonym</td>
<td>Semantic opposition between lemmas</td>
<td>leader → follower</td>
<td></td>
</tr>
<tr>
<td>Derivationally Related Form</td>
<td>Lemmas w/same morphological root</td>
<td>destruction → destroy</td>
<td></td>
</tr>
</tbody>
</table>

WordNet Verb Relations

<table>
<thead>
<tr>
<th>Relation</th>
<th>Definition</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypernym</td>
<td>From events to superordinate events</td>
<td>fly → soar</td>
</tr>
<tr>
<td>Trionym</td>
<td>From events to subordinate event (often via specific manner)</td>
<td>snore → sleep</td>
</tr>
<tr>
<td>Entails</td>
<td>From verbs (events) to the verbs (events) they entail</td>
<td>move → sleep</td>
</tr>
<tr>
<td>Antonym</td>
<td>Semantic opposition between lemmas</td>
<td>increase → decrease</td>
</tr>
<tr>
<td>Derivationally Related Form</td>
<td>Lemmas with same morphological root</td>
<td>destroy → destruction</td>
</tr>
</tbody>
</table>

Senses of “bass” in WordNet

Noun

- 1. bass (the lowest part of the musical range)
- 2. bass, bass part (the lowest part in polyphonic music)
- 3. bass, bass part (the lowest part in polychoral music)
- 4. bass, (it is a tailless fish of the family Teleostei)
- 5. bass, bass (any of various North American freshwater fish with long fins, especially of the genus Micropterus)
- 6. bass, bas(e) (the lowest male siring voice)
- 7. bass (the member with the lowest range of a family of musical instruments)

Adjective

- 8. (of) bass, deep (having or denoting a low vocal or instrumental range) "a deep voice," "a bass voice is lower than a baritone voice," "a bass clarinet"
WordNet: Viewed as a graph

WordNet 3.0
- Where it is:
  - http://wordnetweb.princeton.edu/perl/webwn
- Libraries
  - Python: WordNet from NLTK
    - http://www.nltk.org/Home
  - Java:
    - JWNL, extJWNL on sourceforge

Other (domain specific) thesauri

MeSH: Medical Subject Headings
thesaurus from the National Library of Medicine
- MeSH (Medical Subject Headings)
  - 177,000 entry terms that correspond to 26,142 biomedical "headings"
- Hemoglobins
  - Entry Terms: Erythhem, Ferrous Hemoglobin, Hemoglobin
  - Definition: The oxygen-carrying proteins of ERYTHROCYTES. They are found in all vertebrates and some invertebrates. The number of globin subunits in the hemoglobin quaternary structure differs between species. Structures range from monomeric to a variety of multimeric arrangements

The MeSH Hierarchy
- Amino Acids, Peptides, and Proteins (D12)
  - Peptides (D12.778)
- Acute Phase Proteins (D12.779.124)
- Atrial Natriuretic Peptide (D12.779.124.100)
- Antithrombin (D12.779.124.050)
- Blood Coagulation Factors (D12.779.124.125)
- Blood Coagulation Factors, von Willebrand Factor (D12.779.124.184.187)
- Cholesterol (D12.779.124.270)
- Cysteine Proteases (D12.779.124.280)
- Haptoglobin (D12.779.124.337)
- Hemoglobins (D12.779.124.431)
- Carboxyhemoglobin (D12.779.124.400)
- Hemoglobin (D12.779.124.400.141)
- Hemoglobinuria (D12.779.124.400.220)

Uses of the MeSH Ontology
- Provide synonyms ("entry terms")
  - E.g., glucose and dextrose
- Provide hypernyms (from the hierarchy)
  - E.g., glucose ISA monosaccharide
- Indexing in MEDLINE/PubMED database
- NLM's bibliographic database:
  - 20 million journal articles
  - Each article hand assigned 10-20 MeSH terms
Today’s Outline

- Word Senses and Word Relations
- Word Similarity
- Word Sense Disambiguation

Word Similarity

- Synonymy: a binary relation
  - Two words are either synonymous or not
- Similarity (or distance): a looser metric
  - Two words are more similar if they share more features of meaning
- Similarity is properly a relation between senses
  - The word “bank” is not similar to the word “slope”
  - Bank\(^1\) is similar to fund\(^1\)
  - Bank\(^1\) is similar to slope\(^2\)
  - But we’ll compute similarity over both words and senses

Why word similarity

- A practical component in lots of NLP tasks
  - Question answering
  - Natural language generation
  - Automatic essay grading
  - Plagiarism detection
- A theoretical component in many linguistic and cognitive tasks
  - Historical semantics
  - Models of human word learning
  - Morphology and grammar induction

Word similarity and word relatedness

- We often distinguish word similarity from word relatedness
- Similar words: near-synonyms
- Related words: can be related any way
  - car, bicycle: similar
  - car, gasoline: related, not similar

Two classes of similarity algorithms

- Thesaurus-based algorithms
  - Are words “nearby” in hypernym hierarchy?
  - Do words have similar glosses (definitions)?
- Distributional algorithms
  - Do words have similar distributional contexts?

Path based similarity

- Two concepts (senses/synsets) are similar if they are near each other in the thesaurus hierarchy
  - have a short path between them
  - concepts have path 1 to themselves
Refinements to path-based similarity

- \( \text{pathlen}(c_1, c_2) = 1 + \text{number of edges in the shortest path in the hypernym graph between sense nodes } c_1 \text{ and } c_2 \)
- ranges from 0 to 1 (identity)

- \( \text{simpath}(c_1, c_2) = \frac{1}{\text{pathlen}(c_1, c_2)} \)

- \( \text{wordsim}(w_1, w_2) = \max_{c_1 \in \text{senses}(w_1), c_2 \in \text{senses}(w_2)} \text{sim}(c_1, c_2) \)

Example: path-based similarity

\[ \text{simpath}(\text{nickel}, \text{coin}) = \frac{1}{2} = 0.5 \]
\[ \text{simpath}(\text{fund}, \text{budget}) = \frac{1}{2} = 0.5 \]
\[ \text{simpath}(\text{nickel}, \text{currency}) = \frac{1}{4} = 0.25 \]
\[ \text{simpath}(\text{nickel}, \text{money}) = \frac{1}{6} = 0.17 \]
\[ \text{simpath}(\text{coinage}, \text{Richter scale}) = \frac{1}{6} = 0.17 \]

Problem with basic path-based similarity

- Assumes each link represents a uniform distance
- But nickel to money seems to us to be closer than nickel to standard
- Nodes high in the hierarchy are very abstract
- We instead want a metric that
  - Represents the cost of each edge independently
  - Words connected only through abstract nodes
    * are less similar

Information content similarity metrics

- Let's define \( P(c) \) as:
  - The probability that a randomly selected word in a corpus is an instance of concept \( c \)
  - Formally: there is a distinct random variable, ranging over words, associated with each concept in the hierarchy
    * for a given concept, each observed noun is either
      - a member of that concept with probability \( P(c) \)
      - not a member of that concept with probability \( 1 - P(c) \)
  - All words are members of the root node (Entity)
  - \( P(\text{root}) = 1 \)
  - The lower a node in hierarchy, the lower its probability

Information content similarity

- WordNet hierarchy augmented with probabilities \( P(c) \)

Train by counting in a corpus

- Each instance of hill counts toward frequency of natural elevation, geological formation, entity, etc
- Let \( \text{words}(c) \) be the set of all words that are children of node \( c \)
- \( \text{words}(\text{geo-formation}) = \{ \text{hill, ridge, grotto, coast, cave, shore, natural elevation} \} \)
- \( \text{words}(\text{natural elevation}) = \{ \text{hill, ridge} \} \)

\[
P(c) = \frac{\sum_{w \in \text{words}(c)} \text{count}(w)}{N}
\]
Information content: definitions

- Information content:
  \[ \text{IC}(c) = -\log P(c) \]
- Most informative subsumer
  (Lowest common subsumer)
  \( \text{LCS}(c_1, c_2) = \)
  The most informative (lowest) subsumer that is in both the hierarchy subsuming both \( c_1 \) and \( c_2 \)

Using information content for similarity: the Resnik method


- The similarity between two words is related to their common information
- The more two words have in common, the more similar they are
- Resnik: measure common information as:
  - The information content of the most informative (lowest) subsumer (MIS/LCS) of the two nodes
  - \( \text{sim}_{\text{Resnik}}(c_1, c_2) = -\log P(\text{LCS}(c_1, c_2)) \)

Dekang Lin method

Dekang Lin. 1998. An Information-Theoretic Definition of Similarity. ICML.

- Intuition: Similarity between A and B is not just what they have in common
- The more differences between A and B, the less similar they are:
  - Commonality: the more A and B have in common, the more similar they are
  - Difference: the more differences between A and B, the less similar
- Commonality: \( IC(\text{common}(A, B)) \)
- Difference: \( IC(\text{description}(A, B)) - IC(\text{common}(A, B)) \)

Lin similarity function

- \( \text{sim}_{\text{LCS}}(A, B) = \frac{2 \log P(\text{LCS}(c_1, c_2))}{\log P(c_1) + \log P(c_2)} \)
- \( \text{sim}_{\text{LCS}}(\text{hill, coast}) = \frac{2 \log P(\text{geological-formation})}{\log P(\text{hill}) + \log P(\text{coast})} \)

- \( \text{sim}_{\text{LCS}}(\text{hill, coast}) = \frac{2 \log 0.0000189}{\log 0.0000059 + \log 0.0000189} \)
- Lin similarity function:
  \( \text{sim}_{\text{LCS}}(A, B) = \frac{2 \log P(\text{LCS}(c_1, c_2))}{\log P(c_1) + \log P(c_2)} \)

The (extended) Lesk Algorithm

- A thesaurus-based measure that looks at glosses
- Two concepts are similar if their glosses contain similar words
  - Drawing paper: paper that is specially prepared for use in drafting
  - Decal: the art of transferring designs from specially prepared paper to a wood or glass or metal surface
- For each n-word phrase that’s in both glosses
  - Add a score of \( n^2 \)
  - Paper and specially prepared for 1 + 2^2 = 5
  - Compute overlap also for other relations
    - glosses of hypernyms and hyponyms

Information content:

- Information content:
  \[ \text{IC}(c) = -\log P(c) \]
- Most informative subsumer
  (Lowest common subsumer)
  \( \text{LCS}(c_1, c_2) = \)
  The most informative (lowest) subsumer that is in both the hierarchy subsuming both \( c_1 \) and \( c_2 \)

Using information content for similarity: the Resnik method


- The similarity between two words is related to their common information
- The more two words have in common, the more similar they are
- Resnik: measure common information as:
  - The information content of the most informative (lowest) subsumer (MIS/LCS) of the two nodes
  - \( \text{sim}_{\text{Resnik}}(c_1, c_2) = -\log P(\text{LCS}(c_1, c_2)) \)

Dekang Lin method

Dekang Lin. 1998. An Information-Theoretic Definition of Similarity. ICML.

- Intuition: Similarity between A and B is not just what they have in common
- The more differences between A and B, the less similar they are:
  - Commonality: the more A and B have in common, the more similar they are
  - Difference: the more differences between A and B, the less similar
- Commonality: \( IC(\text{common}(A, B)) \)
- Difference: \( IC(\text{description}(A, B)) - IC(\text{common}(A, B)) \)

Lin similarity function

- \( \text{sim}_{\text{LCS}}(A, B) = \frac{2 \log P(\text{LCS}(c_1, c_2))}{\log P(c_1) + \log P(c_2)} \)
- \( \text{sim}_{\text{LCS}}(\text{hill, coast}) = \frac{2 \log P(\text{geological-formation})}{\log P(\text{hill}) + \log P(\text{coast})} \)

- \( \text{sim}_{\text{LCS}}(\text{hill, coast}) = \frac{2 \log 0.0000189}{\log 0.0000059 + \log 0.0000189} \)
- Lin similarity function:
  \( \text{sim}_{\text{LCS}}(A, B) = \frac{2 \log P(\text{LCS}(c_1, c_2))}{\log P(c_1) + \log P(c_2)} \)

The (extended) Lesk Algorithm

- A thesaurus-based measure that looks at glosses
- Two concepts are similar if their glosses contain similar words
  - Drawing paper: paper that is specially prepared for use in drafting
  - Decal: the art of transferring designs from specially prepared paper to a wood or glass or metal surface
- For each n-word phrase that’s in both glosses
  - Add a score of \( n^2 \)
  - Paper and specially prepared for 1 + 2^2 = 5
  - Compute overlap also for other relations
    - glosses of hypernyms and hyponyms
### Summary: thesaurus-based similarity

\[
\begin{align*}
\text{sim}_\text{path}(c_1, c_2) &= \frac{1}{\text{pathlen}(c_1, c_2)} \\
\text{sim}_\text{mbw}(c_1, c_2) &= \frac{-\log P(\text{LCS}(c_1, c_2))}{\log P(c_1) + \log P(c_2)} \\
\text{sim}_\text{exponential}(c_1, c_2) &= \frac{1}{\log P(c_1) + \log P(c_2) + 2\log P(\text{LCS}(c_1, c_2))} \\
\text{sim}_\text{glob}(c_1, c_2) &= \sum_{g \in \text{gloss}(c_1), \text{gloss}(c_2)} \text{overlap}(\text{gloss}(c_1), \text{gloss}(c_2)) \\
\text{sim}_\text{Lem}(c_1, c_2) &= \exp\left(\frac{\log P(c_1) + \log P(c_2) - 2\log P(\text{LCS}(c_1, c_2))}{2}\right)
\end{align*}
\]

### Libraries for computing thesaurus-based similarity

- **NLTK**
  - nltk.corpus.reader.WordNetCorpusReader.res_similarity

- **WordNet::Similarity**
  - Web-based interface:
    - [http://marimba.d.umn.edu/cgi-bin/similarity/similarity.cgi](http://marimba.d.umn.edu/cgi-bin/similarity/similarity.cgi)

### Evaluating similarity

- **Extrinsic (task-based, end-to-end) Evaluation**:
  - Question answering
  - Spell checking
  - Essay grading
  - Word sense disambiguation

- **Intrinsic Evaluation**:
  - Correlation between algorithm and human word similarity ratings
  - Wordsim353 353 noun pairs rated 0-10. simplest Corp=5.77
  - Taking multiple-choice vocabulary tests:
    - *Lexicon* is closest in meaning to: imposed, believed, requested, correlated

### Today’s Outline

- Word Senses and Word Relations
- Word Similarity
- Word Sense Disambiguation

### Word Sense Disambiguation (WSD)

- **Given**
  - A word in context
  - A fixed inventory of potential word senses
  - Decide which sense of the word this is

- **Why?** Machine translation, QA, speech synthesis

- **What set of senses?**
  - English-to-Spanish MT: set of Spanish translations
  - Speech Synthesis: homographs like bass and bow
  - In general: the senses in a thesaurus like WordNet

### Two variants of WSD task

- ** Lexical Sample task**
  - Small pre-selected set of target words (*line, plant*)
  - And inventory of senses for each word
  - Supervised machine learning: train a classifier for each word

- ** All-words task**
  - Every word in an entire text
  - A lexicon with senses for each word
  - Data sparseness: can’t train word-specific classifiers
WSD Methods

• Supervised Machine Learning
• Thesaurus/Dictionary Methods
• Semi-Supervised Learning

Supervised Machine Learning Approaches

• Supervised machine learning approach:
  • a training corpus of words tagged in context with their sense
  • used to train a classifier that can tag words in new text
• Summary of what we need:
  • the tag set (“sense inventory”)
  • the training corpus
  • A set of features extracted from the training corpus
  • A classifier

Supervised WSD 1: WSD Tags

• What’s a tag?
  A dictionary sense?
• For example, for WordNet an instance of “bass” in a text has 8 possible tags or labels (bass1 through bass8, as noun).

8 senses of “bass” in WordNet

1. bass - (the lowest part of the musical range)
2. bass, bass part - (the lowest part in polyphonic music)
3. bass, basso - (an adult male singer with the lowest voice)
4. sea bass, bass - (flesh of lean-fleshed saltwater fish of the family Serranidae)
5. freshwater bass, bass - (any of various North American lean-fleshed freshwater fishes especially of the genus Micropterus)
6. bass, bass voice, basso - (the lowest adult male singing voice)
7. bass - (the member with the lowest range of a family of musical instruments)
8. bass - (nontechnical name for any of numerous edible marine and freshwater spiny-finned fishes)

Supervised WSD 2: Get a corpus

• Lexical sample task:
  • Line hard-serve corpus - 4000 examples of each
  • Interest corpus - 2369 sense-tagged examples
• All words:
  • Semantic concordance: a corpus in which each open-class word is labeled with a sense from a specific dictionary/thesaurus.
    • SemCor: 234,000 words from Brown Corpus, manually tagged with WordNet senses
    • SENSEVAL-3 competition corpora - 2081 tagged word tokens

Supervised WSD 3: Extract feature vectors

Intuition from Warren Weaver (1955):
“If one examines the words in a book, one at a time as through an opaque mask with a hole in it one word wide, then it is obviously impossible to determine, one at a time, the meaning of the words...

But if one lengthens the slit in the opaque mask, until one can see not only the central word in question but also say N words on either side, then if N is large enough one can unambiguously decide the meaning of the central word...
The practical question is: "What minimum value of N will, at least in a tolerable fraction of cases, lead to the correct choice of meaning for the central word?"
Feature vectors

- A simple representation for each observation (each instance of a target word)
- Vectors of sets of feature/value pairs
- Represented as an ordered list of values
- These vectors represent, e.g., the window of words around the target

Two kinds of features in the vectors

- Collocational features and bag-of-words features
  - Collocational
    - Features about words at specific positions near target word
    - Often limited to just word identity and POS
  - Bag-of-words
    - Features about words that occur anywhere in the window (regardless of position)
    - Typically limited to frequency counts

Examples

- Example text (WSJ):
  An electric guitar and **bass** player stand off to one side not really part of the scene
  - Assume a window of +/- 2 from the target

Examples

- Example text (WSJ): **An electric** [guitar] **and** [bass] **player stand** off to one side not really part of the scene,
  - Assume a window of +/- 2 from the target

Collocational features

- Position-specific information about the words and collocations in window
  - **guitar** and **bass** **player stand**
  - \([w_{-3}, \text{POS}_{-3}, w_{-2}, \text{POS}_{-1}, w_{-1}, \text{POS}_{0}, w_{0}, \text{POS}_{1}, w_{1}, \text{POS}_{2}, w_{2}, w_{3}]\)
  - [guitar, \text{NN}, and, \text{CC}, player, \text{NN}, stand, \text{VB}, and guitar, player stand]

Bag-of-words features

- “an unordered set of words” – position ignored
- Counts of words occur within the window.
- First choose a vocabulary
- Then count how often each of those terms occurs in a given window
  - sometimes just a binary “indicator” 1 or 0
Co-Occurrence Example

• Assume we’ve settled on a possible vocabulary of 12 words in “bass” sentences:

[fishing, big, sound, player, fly, rod, pound, double, runs, playing, guitar, band]

• The vector for:
  
guitar and bass player stand
  
[0,0,0,1,0,0,0,0,0,1,0]

Classification: definition

• Input:
  
  • a word w and some features f
  
  • a fixed set of classes C = \{c_1, c_2, \ldots, c_J\}

• Output: a predicted class c \in C

Classification Methods: Supervised Machine Learning

• Input:
  
  • a word w in a text window d (which we’ll call a “document”)
  
  • a fixed set of classes C = \{c_1, c_2, \ldots, c_J\}
  
  • A training set of m hand-labeled text windows again called “documents” (d_1, c_1), \ldots, (d_m, c_m)

• Output:
  
  • a learned classifier γ: d \rightarrow c

Applying Naive Bayes to WSD

• P(c) is the prior probability of that sense
  
  • Counting in a labeled training set.
  
  • P(w|c): conditional probability of a word given a particular sense
  
  • We get both of these from a tagged corpus like SemCor

\[ \hat{P}(c) = \frac{N_c}{N} \]

\[ \hat{P}(w|c) = \frac{\text{count}(w,c) + 1}{\text{count}(c) + |V|} \]

Choosing a class:

\[ \hat{P}(\text{lilf}|\text{fsh smocl fsh}) = \frac{3/4 \times 2/14 \times 0.5}{2/14 \times 1/4} = 0.03003 \]

\[ \hat{P}(\text{gtr}|\text{gtr}) = \frac{1/4 \times 2/9 \times 1/29}{2/9 \times 1/29} = 0.00006 \]
WSD Evaluations and baselines

- **Best evaluation:** extrinsic ("end-to-end", "task-based") evaluation
- Embed WSD algorithm in a task and see if you can do the task better!
- **What we often do for convenience:** intrinsic evaluation
  - Exact match sense accuracy
    - % of words tagged identically with the human-manual sense tags
  - Usually evaluate using held-out data/test data from same labeled corpus
- **Baselines**
  - Most frequent sense
  - The Lesk algorithm

Most Frequent Sense

- WordNet senses are ordered in frequency order
- So "most frequent sense" in WordNet = "take the first sense"
- Sense frequencies come from the SemCor corpus

<table>
<thead>
<tr>
<th>Freq</th>
<th>Synset</th>
<th>Gloss</th>
</tr>
</thead>
<tbody>
<tr>
<td>338</td>
<td>plant$^1$, works, industrial plant</td>
<td>buildings for carrying on industrial labor</td>
</tr>
<tr>
<td>207</td>
<td>plant$^2$, flora, plant life</td>
<td>a living organism lacking the power of locomotion</td>
</tr>
<tr>
<td>2</td>
<td>plant$^3$</td>
<td>something planted secretly for discovery by another</td>
</tr>
<tr>
<td>0</td>
<td>plant$^4$</td>
<td>an actor assumed in the audience whose acting is enhanced by sense spontaneity to the audience</td>
</tr>
</tbody>
</table>

Ceiling

- **Human inter-annotator agreement**
  - Compare annotations of two humans
  - On same data
  - Given same tagging guidelines
- Human agreements on all-words corpora with WordNet style senses
  - 75%-80%

The Simplified Lesk algorithm

Choose sense with most word overlap between gloss and context (not counting function words)

The bank can guarantee deposits will eventually cover future tuition costs because it invests in adjustable-rate mortgage securities.

<table>
<thead>
<tr>
<th>Sense</th>
<th>Gloss</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bank$^1$</td>
<td>a financial institution that accepts deposits and channels the money into lending activities</td>
</tr>
<tr>
<td>Examples:</td>
<td>&quot;he cashed a check at the bank&quot;, &quot;that bank holds the mortgage on my home&quot;</td>
</tr>
<tr>
<td>Bank$^2$</td>
<td>sloping land (especially the slope beside a body of water)</td>
</tr>
<tr>
<td>Examples:</td>
<td>&quot;they pulled the canoe up on the bank&quot;, &quot;he sat on the bank of the river and watched the current&quot;</td>
</tr>
</tbody>
</table>

The Corpus Lesk algorithm

- Assumes we have some sense-labeled data (like SemCor)
- Take all the sentences with the relevant word sense:
  - These short, "streamlined" meetings usually are sponsored by local banks | Chambers of Commerce, trade associations, or other civic organizations.
- Now add these to the gloss + examples for each sense, call it the "signature" of a sense.
- Choose sense with most word overlap between context and signature.
Corpus Lesk: IDF weighting

- Instead of just removing function words
  - Down-weights words that occur in every 'document' (gloss, example, etc)
  - These are generally function words, but is a more fine-grained measure
- Weigh each overlapping word by inverse document frequency

\[
\text{idf}_i = \log \left( \frac{N}{df_i} \right)
\]

score\(sense_i, \text{context}_j\) = \(\sum_{w \in \text{overlap}(\text{signature}_i, \text{context}_j)} \text{idf}_w\)

Semi-Supervised Learning

**Problem:** supervised and dictionary-based approaches require large hand-built resources

What if you don't have so much training data?

**Solution:** Bootstrapping

- Generalize from a very small hand-labeled seed-set.

Bootstrapping

- For bass
  - Rely on "One sense per collocation" rule
    - A word reoccurring in collocation with the same word will almost surely have the same sense.
  - the word play occurs with the music sense of bass
  - the word fish occurs with the fish sense of bass

Sentences extracting using “fish” and “play”

We need more good teachers – right now, there are only a half a dozen who can play the free bass with ease.

An electric guitar and bass player stand off to one side, not really part of the scene, just as a sort of nod to grunge expectations perhaps.

The researchers said the worms spend part of their life cycle in such fish as Pacific salmon and striped bass and Pacific rockfish or snapper.

And it all started when fishermen decided the striped bass in Lake Mead were too skinny.

Summary: generating seeds

1) Hand labeling
2) "One sense per collocation":
   - A word reoccurring in collocation with the same word will almost surely have the same sense.
3) "One sense per discourse":
   - The sense of a word is highly consistent within a document - Yarowsky (1995)
   - (At least for non-function words, and especially topic-specific words)
Stages in the Yarowsky bootstrapping algorithm for the word “plant”

Summary

- Word Sense Disambiguation: choosing correct sense in context
- Applications: MT, QA, etc.
- Three classes of Methods
  - Supervised Machine Learning: Naive Bayes classifier
  - Thesaurus/Dictionary Methods
  - Semi-Supervised Learning
- Main intuition
  - There is lots of information in a word’s context.
  - Simple algorithms based just on word counts can be surprisingly good