Computational Linguistics 1

CMSC/LING 723, LBSC 744

Kristy Hollingshead Seitz
Institute for Advanced Computer Studies
University of Maryland

Lecture 19: 8 & 10 November 2011

Agenda

- Midterms handed back today
  - Discussion
- Questions, comments, concerns?
- Evaluating parser accuracy for HW5
- Finish context-sensitive grammars discussion
  - Combinatory Categorial Grammars (CCG)
- Semantics
  - Meaning
  - Word sense
  - Semantic similarity

Midterm Discussion

- Follow directions!
  - name on every page
- Counting bigrams
  - Don't count across sentences
  - Include <s> and </s> as tokens
- Function composition
  - FST2 $\circ$ FST1 = FST2(FST1(input))
  - epsilon transitions
  - mismatch of output to input (e.g., "BROWN")
- Viterbi & Forward
  - **show your work**
  - include </s> transition
  - assumptions for $b_{</s>}$
- Perplexity: N=4

Evaluating Parses

- Unlike in tagging, parsing results in a variable number of tags being annotated
- Example: *systems analyst arbitration chef*
  (NP (NNS systems) (NN analyst) (NN arbitration) (NN chef))
  (NP (NP (NNS systems) (NN analyst)) (NN arbitration) (NN chef))
  (NP (NP (NP (NNS systems) (NN analyst)) (NN arbitration)) (NN chef))
- How do we score a parse relative to the true parse?
- Need to penalize a parser that guesses too many constituents, as well as a parser that guesses too few
- Guessing both label and span of constituent

Precision and Recall

- Each constituent has a label and a span
- For each constituent in the guessed parse, we can try to match it to a constituent in the true parse with the same label and span
- Each constituent in the true parse can only match with one in the guessed parse
- A constituent is counted correct if it matches
- A parser has high *precision* if most of the constituents it guessed were correct

$$\text{labeled precision (LP)} = \frac{\textbf{Number of constituents correct}}{\textbf{Number of constituents in guess}}$$

- A parser has high *recall* if it guesses most of the true constituents correctly

$$\text{labeled recall (LR)} = \frac{\textbf{Number of constituents correct}}{\textbf{Number of constituents in truth}}$$

F-score

- Suppose we don't care about recall...how could we get very high precision (nearly 100%)?
  - Put just a flat S category spanning the whole string
  - Precision would be high; recall low
- Suppose we don't care about precision...how could we get very high recall (100%)?
  - Guess every category for every span
  - Recall would be high; precision low
- Must measure *both* for evaluation purposes
- For those who insist on a single score, the F-measure is common:

$$F = \frac{2(LR)(LP)}{LR + LP}$$

Agenda

- Midterms handed back today
  - Discussion
- Questions, comments, concerns?
- Evaluating parser accuracy for HW5
- Finish context-sensitive grammars discussion
  - Combinatory Categorial Grammars (CCG)
- Semantics
  - Meaning
  - Word sense
  - Semantic similarity

Agenda

- Midterms handed back today
  - Discussion
- Questions, comments, concerns?
- Evaluating parser accuracy for HW5
- Finish context-sensitive grammars discussion
  - Combinatory Categorial Grammars (CCG)
- Semantics
  - Meaning
  - Word sense
  - Semantic similarity

What's meaning?

- Let's start at the word level…
- How do you define the meaning of a word?
- Look it up in the dictionary!

right	*adj.* located nearer the right hand esp. being on the right when facing the same direction as the observer.
left	*adj.* located nearer to this side of the body than the right.
red	*n.* the color of blood or a ruby.
blood	*n.* the red liquid that circulates in the heart, arteries and veins of animals.

Well, that really doesn't help…

Approaches to Meaning

- Truth conditional
- Semantic network

Word Senses

- "Word sense" = distinct meaning of a word
- Same word, different senses
  - Homonyms (homonymy): unrelated senses; identical orthographic form is coincidental
    - Example: "financial institution" vs. "side of river" for bank
  - Polysemes (polysemy): related, but distinct senses
    - Example: "financial institution" vs. "sperm bank"
  - Metonyms (metonymy): "stand in", technically, a sub-case of polysemy
    - Examples: author for works or author, building for organization, capital city for government
- Different word, same sense
  - Synonyms (synonymy)

Just to confuse you…

- Homophones: same pronunciation, different orthography, different meaning
  - Examples: would/wood, to/too/two
- Homographs: distinct senses, same orthographic form, different pronunciation
  - Examples: bass (fish) vs. bass (instrument)

Relationship Between Senses

- IS-A relationships
  - From specific to general (up): hypernym (hypernymy)
    - Example: bird is a hypernym of robin
  - From general to specific (down): hyponym (hyponymy)
    - Example: robin is a hyponym of bird
- Part-Whole relationships
  - wheel is a meronym of car (meronymy)
  - car is a holonym of wheel (holonymy)

WordNet Tour

Material drawn from slides by Christiane Fellbaum

What is WordNet?

- A large lexical database developed and maintained at Princeton University
- Includes most English nouns, verbs, adjectives, adverbs
- Electronic format makes it amenable to automatic manipulation: used in many NLP applications
- "WordNets" generically refers to similar resources in other languages

WordNet: History

- Research in artificial intelligence:
  - How do humans store and access knowledge about concept?
  - Hypothesis: concepts are interconnected via meaningful relations
  - Useful for reasoning
- The WordNet project started in 1986
  - Can most (all?) of the words in a language be represented as a semantic network where words are interlinked by meaning?
  - If so, the result would be a **large** semantic network…

Synonymy in WordNet

- WordNet is organized in terms of "synsets"
  - Unordered set of (roughly) synonymous "words" (or multi-word phrases)
- Each synset expresses a distinct meaning/concept

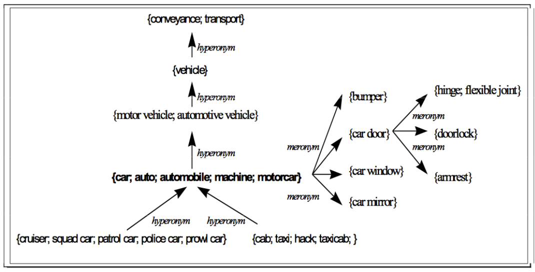
WordNet: Example

Noun
{pipe, tobacco pipe} (a tube with a small bowl at one end; used for smoking tobacco)
{pipe, pipage, piping} (a long tube made of metal or plastic that is used to carry water or oil or gas etc.)
{pipe, tube} (a hollow cylindrical shape)
{pipe} (a tubular wind instrument)
{organ pipe, pipe, pipework} (the flues and stops on a pipe organ)

Verb
{shriek, shrill, pipe up, pipe} (utter a shrill cry)
{pipe} (transport by pipeline) "pipe oil, water, and gas into the desert"
{pipe} (play on a pipe) "pipe a tune"
{pipe} (trim with piping) "pipe the skirt"

Observations about sense granularity?

The “Net” Part of WordNet



Computational Linguistics 1 19

WordNet: Size

Part of speech	Word form	Synsets
Noun	117,798	82,115
Verb	11,529	13,767
Adjective	21,479	18,156
Adverb	4,481	3,621
Total	155,287	117,659

Computational Linguistics 1 20

Word Sense Disambiguation

- Task: automatically select the correct sense of a word
  - Lexical sample
  - All-words
- Theoretically useful for many applications:
  - Semantic similarity (remember from last time?)
  - Information retrieval
  - Machine translation
  - …
- Solution in search of a problem? Why?

Computational Linguistics 1 21

How big is the problem?

- Most words in English have only one sense
  - 62% in Longman’s Dictionary of Contemporary English
  - 79% in WordNet
- But the others tend to have several senses
  - Average of 3.83 in LDOCE
  - Average of 2.96 in WordNet
- Ambiguous words are more frequently used
  - In the British National Corpus, 84% of instances have more than one sense
- Some senses are more frequent than others

Computational Linguistics 1 22

Ground Truth

- Which sense inventory do we use?
- Issues there?
- Application specificity?

Computational Linguistics 1 23

Corpora

- Lexical sample
  - *line-hard-serve* corpus (4k sense-tagged examples)
  - *interest corpus* (2,369 sense-tagged examples)
  - …
- All-words
  - SemCor (234k words, subset of Brown Corpus)
  - Senseval-3 (2081 tagged content words from 5k total words)
  - …
- Observations about the size?

Computational Linguistics 1 24

Evaluation

- Intrinsic
  - Measure accuracy of sense selection wrt ground truth
- Extrinsic
  - Integrate WSD as part of a bigger end-to-end system, e.g., machine translation or information retrieval
  - Compare ±WSD

End of lecture, 8 Nov.

Agenda

- HW4 handed back today
  - Grades are reported out of 100, so -20 for true grade
- Questions, comments, concerns?
- Semantics
  - Meaning
  - Word sense disambiguation
  - Semantic similarity

Word Sense Disambiguation (WSD)

- Take a word in context and resolve which sense of the word is being used
- Example: He is washing the *dishes* versus
  He is cooking three *dishes*
- In some past competitions, just given verb and object pairs, goal to disambiguate object
  - Selectional restrictions of verbs drive disambiguation
  - (How do we learn selectional restrictions?)

Evaluation of WSD

- Different words have a different degree of difficulty
  - As far as I know, *aardvark* has one sense
  - The word *goal* has many senses
- Some differences in senses are relatively subtle
  - e.g. financial bank versus blood bank versus river bank
- How to provide partial credit for 'close' answers
- Senseval is a competition that has addressed many of these questions

Baseline + Upper Bound

- Baseline: most frequent sense
  - Equivalent to "take first sense" in WordNet
  - Does surprisingly well!

62% accuracy in this case!

Freq	Synset	Gloss
338	$plant^1$, works, industrial plant	buildings for carrying on industrial labor
207	$plant^2$, flora, plant life	a living organism lacking the power of locomotion
2	$plant^3$	something planted secretly for discovery by another
0	$plant^4$	an actor situated in the audience whose acting is rehearsed but seems spontaneous to the audience

- Upper bound:
  - Fine-grained WordNet sense: 75-80% human agreement
  - Coarser-grained inventories: 90% human agreement possible
- What does this mean?

WSD Approaches

- Depending on use of manually created knowledge sources
  - Knowledge-lean
  - Knowledge-rich
- Depending on use of labeled data
  - Supervised
  - Semi- or minimally supervised
  - Unsupervised

Lesk's Algorithm

- Intuition: note word overlap between context and dictionary entries
  - Unsupervised, but knowledge rich

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

WordNet

$bank^1$	Gloss:	a financial institution that accepts deposits and channels the money into lending activities
	Examples:	"he cashed a check at the bank", "that bank holds the mortgage on my home"
$bank^2$	Gloss:	sloping land (especially the slope beside a body of water)
	Examples:	"they pulled the canoe up on the bank", "he sat on the bank of the river and watched the currents"

Lesk's Algorithm

- Simplest implementation:
  - Count overlapping content words between glosses and context
- Lots of variants:
  - Include the examples in dictionary definitions
  - Include hypernyms and hyponyms
  - Give more weight to larger overlaps (e.g., bigrams)
  - Give extra weight to infrequent words (e.g., *idf* weighting)
  - …
- Works reasonably well!

Supervised WSD

- WSD as a supervised classification task
  - Train a separate classifier for each word
- Three components of a machine learning problem:
  - Training data (corpora)
  - Representations (features)
  - Learning method (algorithm, model)

Features

- Possible features
  - POS and surface form of the word itself
  - Surrounding words and POS tag
  - Positional information of surrounding words and POS tags
  - Same as above, but with *n*-grams
  - Grammatical information
  - …
- Richness of the features?
  - Richer features = ML algorithm does less of the work
  - More impoverished features = ML algorithm does more of the work

Classifiers

- Once we cast the WSD problem as supervised classification, many learning techniques are possible:
  - Naïve Bayes (the thing to try first)
  - Decision lists
  - Decision trees
  - MaxEnt
  - Support vector machines
  - Nearest neighbor methods
  - …

Classifiers Tradeoffs

- Which classifier should I use?
- It depends:
  - Number of features
  - Types of features
  - Number of possible values for a feature
  - Noise
  - …
- General advice:
  - Start with Naïve Bayes
  - Use decision trees/lists if you want to understand what the classifier is doing
  - SVMs often give state of the art performance
  - MaxEnt methods also work well

Naïve Bayes

- Extract features Φ, predict word based on features
- Common features: POS-tag and word collocations, word co-occurrence
  - should use subcategorization if available!
- Simplest approach (common baseline): Naive Bayes:
  - Given a set of senses s ∈ S, pick the sense that is most probable given the context ("context" represented by feature vector):

$$\hat{s} = \operatorname*{argmax}_{s \in S} P(s|\Phi) = \operatorname*{argmax}_{s \in S} \frac{P(\Phi|s)P(s)}{P(\Phi)} = \operatorname*{argmax}_{s \in S} P(\Phi|s)P(s)$$

- Problem: data sparsity!

The "Naïve" Part

- Feature vectors are too sparse to estimate directly
  - So… assume features are conditionally independent given the word sense
  - This is naïve because?
- Putting everything together:

$$\hat{s} \approx \operatorname*{argmax}_{s \in S} \left(\prod_{j=1}^{|\Phi|} P(\Phi_j|s) \right) P(s)$$

Naïve Bayes: Training

- How do we estimate the probability distributions?

$$\hat{s} = \arg\max_{s \in S} P(s) \prod_{j=1}^{n} P(\vec{f} \mid s)$$

- Maximum-Likelihood Estimates (MLE):

$$P(s_i) = \frac{\text{count}(s_i, w_j)}{\text{count}(w_j)}$$

$$P(f_j \mid s) = \frac{\text{count}(f_j, s)}{\text{count}(s)}$$

- What else do we need to do?

Well, how well does it work? (later…)

Decision Lists

- Used for binary classification problems, e.g. **bass**[1] (fish) versus **bass**[2] (guitar)
- A list like a **case** statement in programming:
  - test condition 1; if true, set sense and break
  - otherwise, test condition 2, . . .
- Learn by generating and ordering tests
  - Order by, e.g., log likelihood ratio

Decision List: Example

- Example decision list, discriminating between bass[1] (fish) and bass[2] (music) :

Context	Sense
fish in ±*k* words	FISH
striped bass	FISH
guitar in ±*k* words	MUSIC
bass player	MUSIC
piano in ±*k* words	MUSIC
sea bass	FISH
play bass	MUSIC
river in ±*k* words	FISH
on bass	MUSIC
bass are	FISH

Decision List: Example

Rule		Sense
fish within window	⇒	**bass**[1]
striped bass	⇒	**bass**[1]
guitar within window	⇒	**bass**[2]
bass player	⇒	**bass**[2]
piano within window	⇒	**bass**[2]
tenor within window	⇒	**bass**[2]
sea bass	⇒	**bass**[1]
play/V *bass*	⇒	**bass**[2]
river within window	⇒	**bass**[1]
violin within window	⇒	**bass**[2]
salmon within window	⇒	**bass**[1]
on bass	⇒	**bass**[2]
bass are	⇒	**bass**[1]

Building Decision Lists

- Simple algorithm:
  - Compute how discriminative each feature is:

$$\left|\log\left(\frac{P(S_1 \mid f_i)}{P(S_2 \mid f_i)}\right)\right|$$

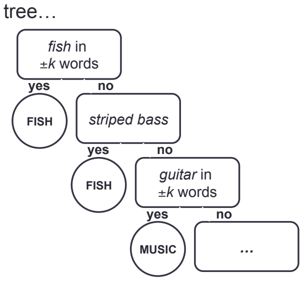
  - Create ordered list of tests from these values
- Limitation?
- How do you build *n*-way classifiers from binary classifiers?
  - One vs. rest (sequential vs. parallel)
  - Another learning problem

Well, how well does it work? (later…)

Decision Trees

- Instead of a list, imagine a tree…

Context	Sense
fish in ±*k* words	FISH
striped bass	FISH
guitar in ±*k* words	MUSIC
bass player	MUSIC
piano in ±*k* words	MUSIC
sea bass	FISH
play bass	MUSIC
river in ±*k* words	FISH
on bass	MUSIC
bass are	FISH



Using Decision Trees

- Given an instance (= list of feature values)
  - Start at the root
  - At each interior node, check feature value
  - Follow corresponding branch based on the test
  - When a leaf node is reached, return its category

Building Decision Trees

- Basic idea: build tree top down, recursively partitioning the training data at each step
  - At each node, try to split the training data on a feature (could be binary or otherwise)
- What features should we split on?
  - Small decision tree desired
  - Pick the feature that gives the most information about the category
- Example: 20 questions
  - I'm thinking of a number from 1 to 1,000
  - You can ask any yes no question
  - What question would you ask?

Evaluating Splits via Entropy

- Entropy of a set of events *E*:

$$H(E) = -\sum_{c \in C} P(c)\log_2 P(c)$$

  - Where *P(c)* is the probability that an event in *E* has category *c*
- How much information does a feature give us about the category (sense)?
  - *H(E)* = entropy of event set *E*
  - *H(E|f)* = expected entropy of event set *E* once we know the value of feature *f*
  - Information Gain: *G(E, f)* = *H(E)* – *H(E|f)* = amount of new information provided by feature *f*
- Split on feature that maximizes information gain

Well, how well does it work? (later…)

WSD Accuracy

- Generally:
  - Naïve Bayes provides a reasonable baseline: ~70%
  - Decision lists and decision trees slightly lower
  - State of the art is slightly higher
- However:
  - Accuracy depends on actual word, sense inventory, amount of training data, number of features, etc.
  - Remember caveat about baseline and upper bound

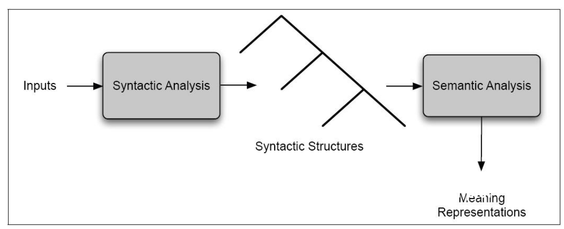
WSD with Parallel Text

- But annotations are expensive!
- What's the "proper" sense inventory?
  - How fine or coarse grained?
  - Application specific?
- Observation: multiple senses translate to different words in other languages!
  - A "bill" in English may be a "pico" (bird jaw) in or a "cuenta" (invoice) in Spanish
  - Use the foreign language as the sense inventory!
  - Added bonus: annotations for free! (Byproduct of word-alignment process in machine translation)

Beyond Lexical Semantics

Syntax-Semantics Pipeline

- Interaction between lexical semantic and syntax



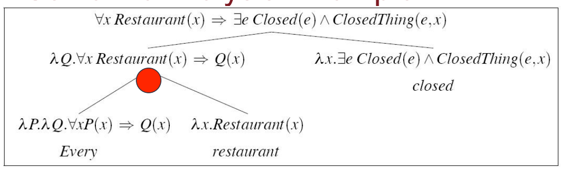
Semantic Attachments

- Basic idea:
  - Associate λ-expressions with lexical items
  - At branching node, apply semantics of one child to another (based on synctatic rule)
- Refresher in λ-calculus…

Augmenting Syntactic Rules

Grammar Rule
$S \rightarrow NP\ VP$
$NP \rightarrow Det\ Nominal$
$NP \rightarrow ProperNoun$
$Nominal \rightarrow Noun$
$VP \rightarrow Verb$
$VP \rightarrow Verb\ NP$
$Det \rightarrow every$
$Det \rightarrow a$
$Noun \rightarrow restaurant$
$ProperNoun \rightarrow Matthew$
$ProperNoun \rightarrow Franco$
$ProperNoun \rightarrow Frasca$
$Verb \rightarrow closed$
$Verb \rightarrow opened$

Semantic Analysis: Example



NP → Det Nominal {Det.*sem*(Nominal.*sem*)}

$$\lambda P.\lambda Q.\forall x\ P(x) \Rightarrow Q(x)(\lambda x.\text{Restaurant}(x))$$
$$\lambda Q.\forall x \lambda x.\text{Restaurant}(x)(x) \Rightarrow Q(x)$$
$$\lambda Q.\forall x\ \text{Restaurant}(x) \Rightarrow Q(x)$$

Complexities

- Oh, there are many…
- Classic problem: quantifier scoping
  - Every restaurant has a menu
- Issues with this style of semantic analysis?

Semantics in NLP Today

- Can be characterized as "shallow semantics"
- Verbs denote events
  - Represent as "frames"
- Nouns (in general) participate in events
  - Types of event participants = "slots" or "roles"
  - Event participants themselves = "slot fillers"
  - Depending on the linguistic theory, roles may have special names: agent, theme, etc.
- Semantic analysis: semantic role labeling
  - Automatically identify the event type (i.e., frame)
  - Automatically identify event participants and the role that each plays (i.e., label the semantic role)

Semantic Role Labeling: Thematic Roles

- Syntactically, verbs call for arguments
- The arguments play semantic roles, dictated by the verb
- For example, *the dog bit the postman*
  - the dog is the bit**er**
  - the postman is the bit**ee**
- Range of complicated roles that arise

Common Thematic Roles

Thematic Role	Definition
AGENT	The volitional causer of an event
EXPERIENCER	The experiencer of an event
FORCE	The non-volitional causer of the event
THEME	The participant most directly affected by an event
RESULT	The end product of an event
CONTENT	The proposition or content of a propositional event
INSTRUMENT	An instrument used in an event
BENEFICIARY	The beneficiary of an event
SOURCE	The origin of the object of a transfer event
GOAL	The destination of an object of a transfer event

Thematic Roles: Examples

Thematic Role	Example
AGENT	*The waiter* spilled the soup.
EXPERIENCER	*John* has a headache.
FORCE	*The wind* blows debris from the mall into our yards.
THEME	Only after Benjamin Franklin broke *the ice*...
RESULT	The French government has built a *regulation-size baseball diamond*...
CONTENT	Mona asked *"You met Mary Ann at a supermarket"?*
INSTRUMENT	He turned to poaching catfish, stunning them *with a shocking device*...
BENEFICIARY	Whenever Ann Callahan makes hotel reservations *for her boss*...
SOURCE	I flew in *from Boston*.
GOAL	I drove *to Portland*.

Constraints on Thematic Roles

- Verbs impose constraints on what fills their roles
  - Refresher: selectional restrictions
- Example: agent of *imagine* must be animate
- These constraints can aid interpretation
  - John would like to eat downtown tonight
  - John would like to eat sushi tonight
- In the case of violated constraints, features can be coerced, such as animacy
  - The thumbtack took *revenge* on the unruly poster

PropBank: Two Examples

- agree.01
  - Arg0: Agreer
  - Arg1: Proposition
  - Arg2: Other entity agreeing
  - Example: [$_{Arg0}$ John] *agrees* [$_{Arg2}$ with Mary] [$_{Arg1}$ on everything]
- fall.01
  - Arg1: Logical subject, patient, thing falling
  - Arg2: Extent, amount fallen
  - Arg3: Start point
  - Arg4: End point
  - Example: [$_{Arg1}$ Sales] fell [$_{Arg4}$ to $251.2 million] [$_{Arg3}$ from $278.7 million]

How do we do it?

- Short answer: supervised machine learning
- One approach: classification of each tree constituent
  - Features can be words, phrase type, linear position, tree position, etc.
  - Apply standard machine learning algorithms

Agenda

- Midterms handed back today
  - Discussion
- Questions, comments, concerns?
- Evaluating parser accuracy for HW5
- Finish context-sensitive grammars discussion
  - Combinatory Categorial Grammars (CCG)
- Semantics
  - Meaning
  - Word sense
  - Semantic similarity

Agenda

- HW4 handed back today
  - Grades are reported out of 100, so -20 for true grade
- Questions, comments, concerns?
- Semantics
  - Meaning
  - Word sense disambiguation
  - Semantic similarity

Intuition of Semantic Similarity

Semantically close
- bank–money
- apple–fruit
- tree–forest
- bank–river
- pen–paper
- run–walk
- mistake–error
- car–wheel

Semantically distant
- doctor–beer
- painting–January
- money–river
- apple–penguin
- nurse–bottle
- pen–river
- clown–tramway
- car–algebra

Why?

- Meaning
  - The two concepts are close in terms of their meaning
- World knowledge
  - The two concepts have similar properties, often occur together, or occur in similar contexts
- Psychology
  - We often think of the two concepts together

Two Types of Relations

- Synonymy: two words are (roughly) interchangeable


- Semantic similarity (distance): somehow "related"
  - Sometimes, explicit lexical semantic relationship, often, not


Validity of Semantic Similarity

- Is semantic distance a valid linguistic phenomenon?
- Experiment (Rubenstein and Goodenough, 1965)
  - Compiled a list of word pairs
  - Subjects asked to judge semantic distance (from 0 to 4) for each of the word pairs
- Results:
  - Rank correlation between subjects is ~0.9
  - People are consistent!

Compute Semantic Similarity?

- Task: automatically compute semantic similarity between words
- Theoretically useful for many applications:
  - Detecting paraphrases (i.e., automatic essay grading, plagiarism detection)
  - Information retrieval
  - Machine translation
  - …
- Solution in search of a problem?

Agenda: Summary

- Midterms handed back today
  - Discussion
- Questions, comments, concerns?
- Evaluating parser accuracy for HW5
- Finish context-sensitive grammars discussion
  - Combinatory Categorial Grammars (CCG)
- Semantics
  - Meaning
  - Word sense
  - Semantic similarity

Agenda

- HW4 handed back today
  - Grades are reported out of 100, so -20 for true grade
- Questions, comments, concerns?
- Semantics
  - Meaning
  - Word sense disambiguation
  - Semantic similarity
- HW5 due on Tuesday!