CS474 Natural Language Processing

- Today
  - Lexical semantic resources: WordNet
  - Word sense disambiguation
    - Dictionary-based approaches
    - Supervised machine learning methods
    - Issues for WSD evaluation

Word sense disambiguation

- Given a *fixed* set of senses associated with a lexical item, determine which of them applies to a particular instance of the lexical item
- Two fundamental approaches
  - WSD occurs during semantic analysis as a side-effect of the elimination of ill-formed semantic representations
  - Stand-alone approach
    - WSD is performed independent of, and prior to, compositional semantic analysis
    - Makes minimal assumptions about what information will be available from other NLP processes
    - Applicable in large-scale practical applications

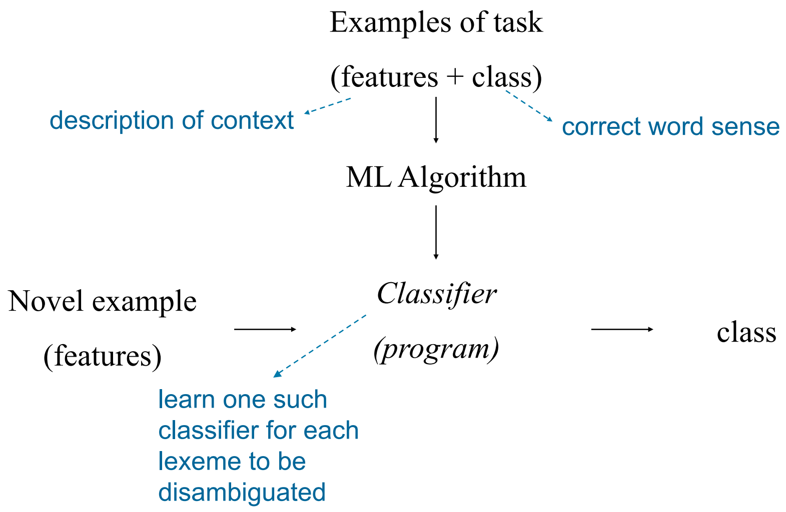
Dictionary-based approaches

- Rely on machine readable dictionaries
- Initial implementation of this kind of approach is due to Michael Lesk (1986)
  - Given a word *W* to be disambiguated in context C
    - Retrieve all of the sense definitions, *S*, for *W* from the MRD
    - Compare each *s* in *S* to the dictionary definitions D of all the remaining words c in the context C
    - Select the sense *s* with the most overlap with D (the definitions of the context words C)

Machine learning approaches

- Machine learning methods
  - Supervised inductive learning
  - Bootstrapping
  - Unsupervised
- Emphasis is on acquiring the knowledge needed for the task from data, rather than from human analysts.

Inductive ML framework



Examples of task

(features + class)

description of context

correct word sense

ML Algorithm

Novel example

(features)

Classifier

(program)

class

learn one such classifier for each lexeme to be disambiguated

Running example

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

1 Fish sense

2 Musical sense

3 …

Feature vector representation

- **target:** the word to be disambiguated
- **context** : portion of the surrounding text
  - Select a "window" size
  - Tagged with part-of-speech information
  - Stemming or morphological processing
  - Possibly some partial parsing
- Convert the context (and target) into a set of features
  - Attribute-value pairs
    - Numeric, boolean, categorical, …

Collocational features

- Encode information about the lexical inhabitants of *specific* positions located to the left or right of the target word.
  - E.g. the word, its root form, its part-of-speech
  - *An electric guitar and **bass** player stand off to one side, not really part of the scene, just as a sort of nod to gringo expectations perhaps.*

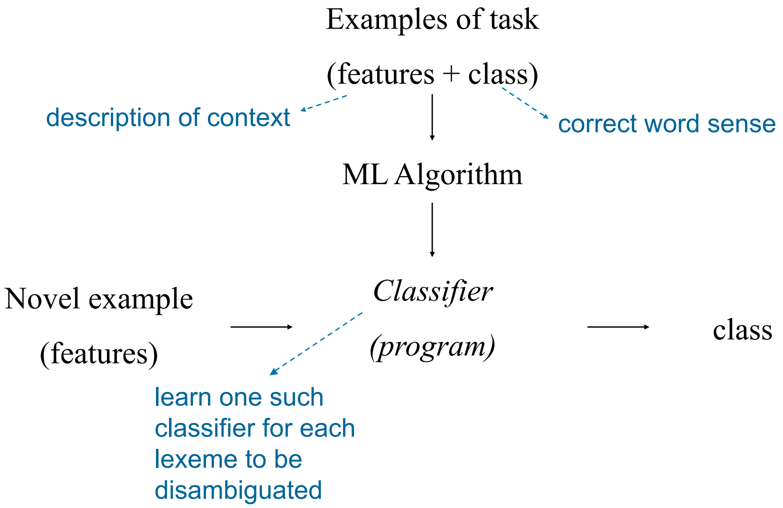
pre2-word	pre2-pos	pre1-word	pre1-pos	fol1-word	fol1-pos	fol2-word	fol2-pos
guitar	NN1	and	CJC	player	NN1	stand	VVB

Co-occurrence features

- Encodes information about neighboring words, ignoring exact positions.
  - Select a small number of frequently used content words for use as features
    - 12 most frequent content words from a collection of *bass* sentences drawn from the WSJ: *fishing, big, sound, player, fly, rod, pound, double, runs, playing, guitar, band*
    - Co-occurrence vector (window of size 10)
  - **Attributes**: the words themselves (or their roots)
  - **Values**: number of times the word occurs in a region surrounding the target word

fishing?	big?	sound?	player?	fly?	rod?	pound?	double?	…	guitar?	band?
0	0	0	1	0	0	0	0		1	0

Inductive ML framework



Examples of task

(features + class)

description of context

correct word sense

ML Algorithm

Novel example

(features)

Classifier

(program)

class

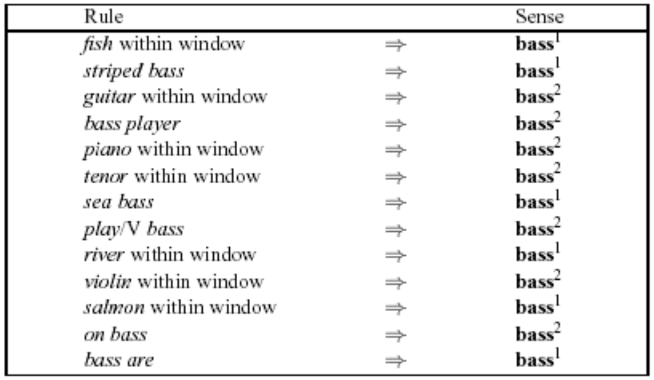
learn one such classifier for each lexeme to be disambiguated

Decision list classifiers

- Decision lists: equivalent to simple case statements.
  - Classifier consists of a sequence of tests to be applied to each input example/vector; returns a word sense.
- Continue only until the first applicable test.
- Default test returns the majority sense.

Decision list example

- Binary decision: fish *bass* vs. musical *bass*



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}

Learning decision lists

- Consists of *generating* and *ordering* individual tests based on the characteristics of the training data
- Generation: every feature-value pair constitutes a test
- Ordering: based on accuracy on the training set

$$abs\left(\log\frac{P(Sense_1 \mid f_i = v_j)}{P(Sense_2 \mid f_i = v_j)}\right)$$

- Associate the appropriate sense with each test

WSD Evaluation

- Corpora:
  - *line* corpus
  - Yarowsky's 1995 corpus
    - » 12 words (plant, space, bass, …)
    - » ~4000 instances of each
  - Ng and Lee (1996)
    - » 121 nouns, 70 verbs (most frequently occurring/ambiguous); WordNet senses
    - » 192,800 occurrences
  - SEMCOR (Landes et al. 1998)
    - » Portion of the Brown corpus tagged with WordNet senses
  - SENSEVAL (Kilgarriff and Rosenzweig, 2000)
    - » Annual performance evaluation conference
    - » Provides an evaluation framework (Kilgarriff and Palmer, 2000)
- Baseline: most frequent sense

WSD Evaluation

- Metrics
  - Precision
    - » Nature of the senses used has a huge effect on the results
    - » E.g. results using coarse distinctions cannot easily be compared to results based on finer-grained word senses
  - Partial credit
    - » Worse to confuse musical sense of *bass* with a fish sense than with another musical sense
    - » Exact-sense match → full credit
    - » Select the correct broad sense → partial credit
    - » Scheme depends on the organization of senses being used

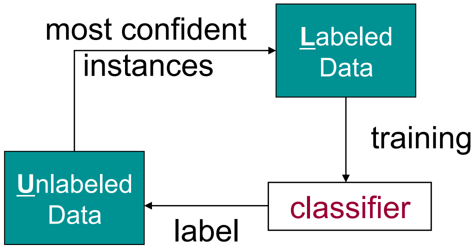
---

CS474 Natural Language Processing

- Before…
  - Lexical semantic resources: WordNet
  - Word sense disambiguation
    - » Dictionary-based approaches
- Today
  - Word sense disambiguation
    - » Supervised machine learning methods
    - » Weakly supervised (bootstrapping) methods
    - » SENSEVAL
    - » Unsupervised methods

Weakly supervised approaches

- Problem: Supervised methods require a large sense-tagged training set
- Bootstrapping approaches: Rely on a small number of labeled **seed** instances



most confident instances

Labeled Data

training

classifier

label

Unlabeled Data

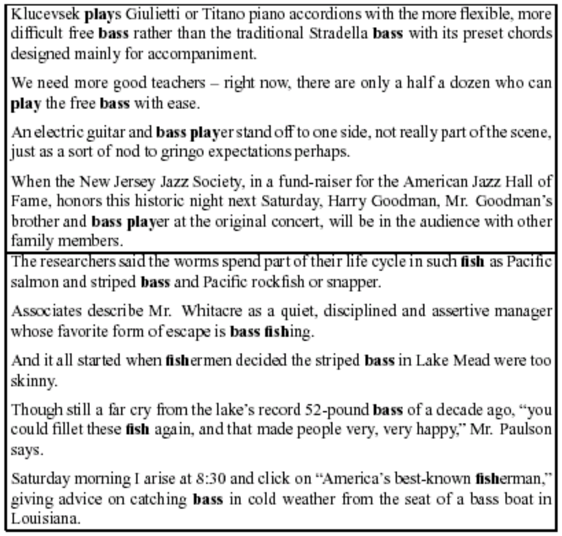
Repeat:

1. train *classifier* on *L*
2. label *U* using *classifier*
3. add *g* of *classifier*' s best *x* to *L*

Generating initial seeds

- Hand label a small set of examples
  - Reasonable certainty that the seeds will be correct
  - Can choose prototypical examples
  - Reasonably easy to do
- **One sense per collocation** constraint (Yarowsky 1995)
  - Search for sentences containing words or phrases that are strongly associated with the target senses
    - » Select *fish* as a reliable indicator of $bass_1$
    - » Select *play* as a reliable indicator of $bass_2$
  - Or derive the collocations automatically from machine readable dictionary entries
  - Or select seeds automatically using collocational statistics (see Ch 6 of J&M)

One sense per collocation



Klucevsek **play**s Giulietti or Titano piano accordions with the more flexible, more difficult free **bass** rather than the traditional Stradella **bass** with its preset chords designed mainly for accompaniment.

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** **play**er stand off to one side, not really part of the scene, just as a sort of nod to gringo expectations perhaps.

When the New Jersey Jazz Society, in a fund-raiser for the American Jazz Hall of Fame, honors this historic night next Saturday, Harry Goodman, Mr. Goodman's brother and **bass** **play**er at the original concert, will be in the audience with other family members.

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.

Associates describe Mr. Whitacre as a quiet, disciplined and assertive manager whose favorite form of escape is **bass** **fish**ing.

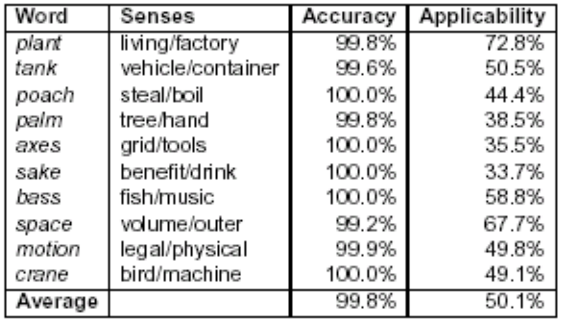
And it all started when **fish**ermen decided the striped **bass** in Lake Mead were too skinny.

Though still a far cry from the lake's record 52-pound **bass** of a decade ago, "you could fillet these **fish** again, and that made people very, very happy," Mr. Paulson says.

Saturday morning I arise at 8:30 and click on "America's best-known **fish**erman," giving advice on catching **bass** in cold weather from the seat of a bass boat in Louisiana.

Yarowsky' s bootstrapping approach

- Relies on a **one sense per discourse** constraint: The sense of a target word is highly consistent within any given document
  - Evaluation on ~37,000 examples



Word	Senses	Accuracy	Applicability
plant	living/factory	99.8%	72.8%
tank	vehicle/container	99.6%	50.5%
poach	steal/boil	100.0%	44.4%
palm	tree/hand	99.8%	38.5%
axes	grid/tools	100.0%	35.5%
sake	benefit/drink	100.0%	33.7%
bass	fish/music	100.0%	58.8%
space	volume/outer	99.2%	67.7%
motion	legal/physical	99.9%	49.8%
crane	bird/machine	100.0%	49.1%
Average		99.8%	50.1%

Yarowsky’s bootstrapping approach

To learn disambiguation rules for a polysemous word:

1. [Find all instances of the word in the training corpus and save the contexts around each instance.]

2. [For each word sense, identify a small set of training examples representative of that sense. Now we have a few labeled examples for each sense.]

3. Build a classifier (e.g. decision list) by training a supervised learning algorithm with the labeled examples.

4. Apply the classifier to all the unlabeled examples. Find instances that are classified with probability > a threshold and add them to the set of labeled examples.

5. *Optional:* Use the one-sense-per-discourse constraint to augment the new examples.

6. Go to Step 3. Repeat until the unlabelled data is stable.

CS474 Natural Language Processing

- Last class
  - Lexical semantic resources: WordNet
  - Word sense disambiguation
    - Dictionary-based approaches
    - Supervised machine learning methods
- Today
  - Word sense disambiguation
    - Supervised machine learning methods (finish)
    - Weakly supervised (bootstrapping) methods
    - SENSEVAL
    - Unsupervised methods

SENSEVAL-2 2001

- Three tasks
  - Lexical sample
  - All-words
  - Translation
- 12 languages
- Lexicon
  - SENSEVAL-1: from HECTOR corpus
  - SENSEVAL-2: from WordNet 1.7
- 93 systems from 34 teams

Lexical sample task

- Select a sample of words from the lexicon
- Systems must then tag instances of the sample words in short extracts of text
- SENSEVAL-1: 35 words
  - 700001 John Dos Passos wrote a poem that talked of `the <tag>bitter</> beat look, the scorn on the lip."
  - 700002 The beans almost double in size during roasting. Black beans are over roasted and will have a <tag>bitter</> flavour and insufficiently roasted beans are pale and give a colourless, tasteless drink.

Lexical sample task: SENSEVAL-1

Nouns		Verbs		Adjectives		Indeterminates	
-n	N	**-v**	N	**-a**	N	**-p**	N
accident	267	amaze	70	brilliant	229	band	302
behaviour	279	bet	177	deaf	122	bitter	373
bet	274	bother	209	floating	47	hurdle	323
disability	160	bury	201	generous	227	sanction	431
excess	186	calculate	217	giant	97	shake	356
float	75	consume	186	modest	270		
giant	118	derive	216	slight	218		
…	…	…	…	…	…		
TOTAL	2756	TOTAL	2501	TOTAL	1406	TOTAL	1785

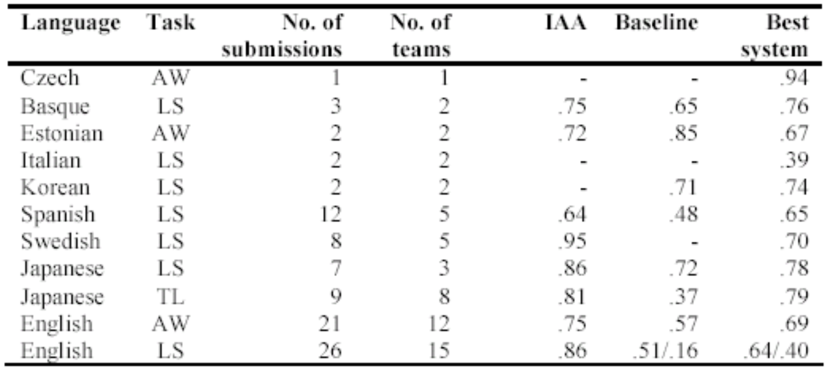
All-words task

- Systems must tag almost all of the content words in a sample of running text
  - sense-tag all predicates, nouns that are heads of noun-phrase arguments to those predicates, and adjectives modifying those nouns
  - ~5,000 running words of text
  - ~2,000 sense-tagged words

Translation task

- SENSEVAL-2 task
- Only for Japanese
- word sense is defined according to translation distinction
  - if the head word is translated differently in the given expressional context, then it is treated as constituting a different sense
- word sense disambiguation involves selecting the appropriate English word/phrase/sentence equivalent for a Japanese word

SENSEVAL-2 results



Language	Task	No. of submissions	No. of teams	IAA	Baseline	Best system
Czech	AW	1	1	-	-	.94
Basque	LS	3	2	.75	.65	.76
Estonian	AW	2	2	.72	.85	.67
Italian	LS	2	2	-	-	.39
Korean	LS	2	2	-	.71	.74
Spanish	LS	12	5	.64	.48	.65
Swedish	LS	8	5	.95	-	.70
Japanese	LS	7	3	.86	.72	.78
Japanese	TL	9	8	.81	.37	.79
English	AW	21	12	.75	.57	.69
English	LS	26	15	.86	.51/.16	.64/.40

SENSEVAL-2 de-briefing

- Where next?
  - Supervised ML approaches worked best
    - Looking at the role of feature selection algorithms
  - Need a well-motivated sense inventory
    - Inter-annotator agreement went down when moving to WordNet senses
  - Need to tie WSD to real applications
    - The translation task was a good initial attempt

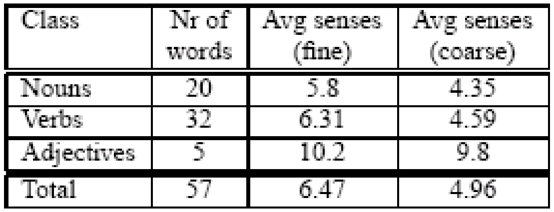
SENSEVAL-3 2004

- 14 core WSD tasks including
  - All words (Eng, Italian): 5000 word sample
  - Lexical sample (7 languages)
- Tasks for identifying semantic roles, for multilingual annotations, logical form, subcategorization frame acquisition

English lexcial sample task

- **Data collected from the Web from Web users**
- Guarantee at least two word senses per word
- 60 ambiguous nouns, adjectives, and verbs
- test data
  - ½ created by lexicographers
  - ½ from the web-based corpus
- Senses from WordNet 1.7.1 and **Wordsmyth** (verbs)
- Sense maps provided for fine-to-coarse sense mapping
- **Filter out multi-word expressions from data sets**

English lexical sample task



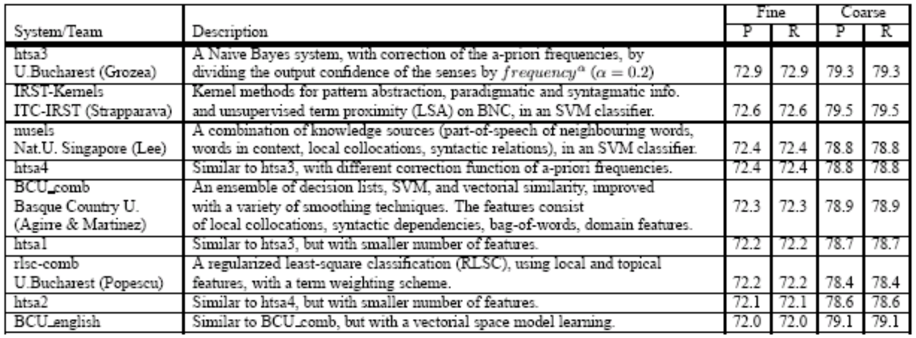
Class	Nr of words	Avg senses (fine)	Avg senses (coarse)
Nouns	20	5.8	4.35
Verbs	32	6.31	4.59
Adjectives	5	10.2	9.8
Total	57	6.47	4.96

Table 1: Summary of the sense inventory

Results

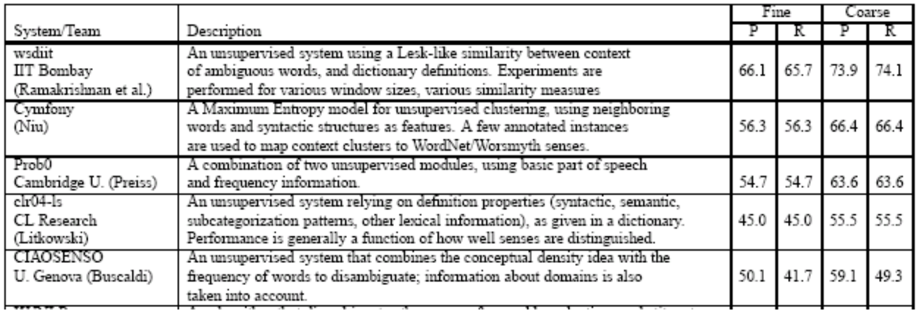
- 27 teams, 47 systems
- Most frequent sense baseline
  - 55.2% (fine-grained)
  - 64.5% (coarse)
- Most systems significantly above baseline
  - Including some unsupervised systems
- Best system
  - 72.9% (fine-grained)
  - 79.3% (coarse)

SENSEVAL-3 lexical sample results



System/Team	Description	Fine		Coarse	
		P	R	P	R
htsa3 U.Bucharest (Grozea)	A Naive Bayes system, with correction of the a-priori frequencies, by dividing the output confidence of the senses by $frequency^{\alpha}$ ($\alpha = 0.2$)	72.9	72.9	79.3	79.3
IRST-Kernels ITC-IRST (Strapparava)	Kernel methods for pattern abstraction, paradigmatic and syntagmatic info. and unsupervised term proximity (LSA) on BNC, in an SVM classifier.	72.6	72.6	79.5	79.5
nusels Nat.U. Singapore (Lee)	A combination of knowledge sources (part-of-speech of neighbouring words, words in context, local collocations, syntactic relations), in an SVM classifier.	72.4	72.4	78.8	78.8
htsa4	Similar to htsa3, with different correction function of a-priori frequencies.	72.4	72.4	78.8	78.8
BCU_comb Basque Country U. (Agirre & Martinez)	An ensemble of decision lists, SVM, and vectorial similarity, improved with a variety of smoothing techniques. The features consist of local collocations, syntactic dependencies, bag-of-words, domain features.	72.3	72.3	78.9	78.9
htsa1	Similar to htsa3, but with smaller number of features.	72.2	72.2	78.7	78.7
rlsc-comb U.Bucharest (Popescu)	A regularized least-square classification (RLSC), using local and topical features, with a term weighting scheme.	72.2	72.2	78.4	78.4
htsa2	Similar to htsa4, but with smaller number of features.	72.1	72.1	78.6	78.6
BCU_english	Similar to BCU_comb, but with a vectorial space model learning.	72.0	72.0	79.1	79.1

SENSEVAL-3 results (unsupervised)



System/Team	Description	Fine		Coarse	
		P	R	P	R
wsdiit IIT Bombay (Ramakrishnan et al.)	An unsupervised system using a Lesk-like similarity between context of ambiguous words, and dictionary definitions. Experiments are performed for various window sizes, various similarity measures	66.1	65.7	73.9	74.1
Cymfony (Niu)	A Maximum Entropy model for unsupervised clustering, using neighboring words and syntactic structures as features. A few annotated instances are used to map context clusters to WordNet/Worsmyth senses.	56.3	56.3	66.4	66.4
Prob0 Cambridge U. (Preiss)	A combination of two unsupervised modules, using basic part of speech and frequency information.	54.7	54.7	63.6	63.6
clr04-ls CL Research (Litkowski)	An unsupervised system relying on definition properties (syntactic, semantic, subcategorization patterns, other lexical information), as given in a dictionary. Performance is generally a function of how well senses are distinguished.	45.0	45.0	55.5	55.5
CIAOSENSO U. Genova (Buscaldi)	An unsupervised system that combines the conceptual density idea with the frequency of words to disambiguate; information about domains is also taken into account.	50.1	41.7	59.1	49.3

CS474 Natural Language Processing

- Last class
  - Lexical semantic resources: WordNet
  - Word sense disambiguation
    - » Dictionary-based approaches
    - » Supervised machine learning methods
- Today
  - Word sense disambiguation
    - » Supervised machine learning methods (finish)
    - » Issues for WSD evaluation
    - » SENSEVAL
    - » Weakly supervised (bootstrapping) methods
    - » Unsupervised methods

Unsupervised WSD

- Rely on **agglomerative clustering** to cluster feature-vector representations (without class/word-sense labels) according to a similarity metric
- Represent each cluster as the average of its constituent feature-vectors
- Label the cluster by hand with known word senses
- Unseen feature-encoded instances are classified by assigning the word sense of the most similar cluster
- Schuetze (1992, 1998) uses a (complex) clustering method for WSD
  - For coarse binary decisions, unsupervised techniques can achieve results approaching those of supervised and bootstrapping methods
  - In most cases approaching the 90% range
  - Tested on a small sample of words

Issues for evaluating clustering

- The correct senses of the instances used in the training data may not be known.
- The clusters are almost certainly heterogeneous w.r.t. the sense of the training instances contained within them.
- The number of clusters is almost always different from the number of senses of the target word being disambiguated.