


CS4740 Natural Language Processing

- Word sense disambiguation
 - Dictionary-based approaches
 - Supervised machine learning methods
 -  – WSD evaluation
 - Weakly supervised methods

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)
 - » Performance evaluation conference (every few years)
 - » Provides an evaluation framework (Kilgarriff and Palmer, 2000)
- Baseline: most frequent sense

Metrics

- Precision

- $\# \text{ correct} / \# \text{ of predictions}$

- Recall (== accuracy)

- $\# \text{ correct} / \# \text{ of examples in test set}$

WSD Evaluation

- Issues with the metrics
 - 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

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.

Adjective

- **S: (adj)** [acrimonious#1](#), **bitter#1** (marked by strong resentment or cynicism) *"an acrimonious dispute"; "bitter about the divorce"*
- **S: (adj)** **bitter#2** (very difficult to accept or bear) *"the bitter truth"; "a bitter sorrow"*
- **S: (adj)** [acerb#2](#), [acerbic#2](#), [acid#1](#), [acrid#2](#), **bitter#3**, [blistering#1](#), [caustic#1](#), [sulfurous#2](#), [sulphurous#2](#), [virulent#3](#), [vitriolic#1](#) (harsh or corrosive in tone) *"an acerbic tone piercing otherwise flowery prose"; "a barrage of acid comments"; "her acrid remarks make her many enemies"; "bitter words"; "blistering criticism"; "caustic jokes about political assassination, talk-show hosts and medical ethics"; "a sulfurous denunciation"; "a vitriolic critique"*
- **S: (adj)** **bitter#4** (expressive of severe grief or regret) *"shed bitter tears"*
- **S: (adj)** **bitter#5** (proceeding from or exhibiting great hostility or animosity) *"a bitter struggle"; "bitter enemies"*
- **S: (adj)** **bitter#6** (causing a sharp and acrid taste experience) *"quinine is bitter"*
- **S: (adj)** [biting#2](#), **bitter#7** (causing a sharply painful or stinging sensation; used especially of cold) *"bitter cold"; "a biting wind"*

Lexical sample task: SENSEVAL-1

Nouns		Verbs		Adjectives		Indeterminates	
-n	#	-v	#	-a	#	-p	#
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

Clickers...

- The **lexical sample** task for WSD is which type of an evaluation?
 - A. Extrinsic
 - B. Intrinsic

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

-
- predicates
 - nouns that are heads of noun-phrase arguments to those predicates
 - adjectives modifying those nouns

The twentieth century author wrote a poem that talked of “the bitter beat look, the scorn on the lip.”

Clickers...

- The **all words** task for WSD is which type of an evaluation?
 - A. Extrinsic
 - B. Intrinsic

Translation task

- SENSEVAL-2 task
- Only for Japanese
- word sense is defined according to translation distinction
 - if the target 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 equivalent for a Japanese word

English → German

WalMart is **open** from 9 to 5.

Ithaca's WalMart **opened** in 2001.

geoffnet

eröffnet

Clickers...

- The **translation** task for WSD is which type of an evaluation?
 - A. Extrinsic
 - B. Intrinsic

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-3 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

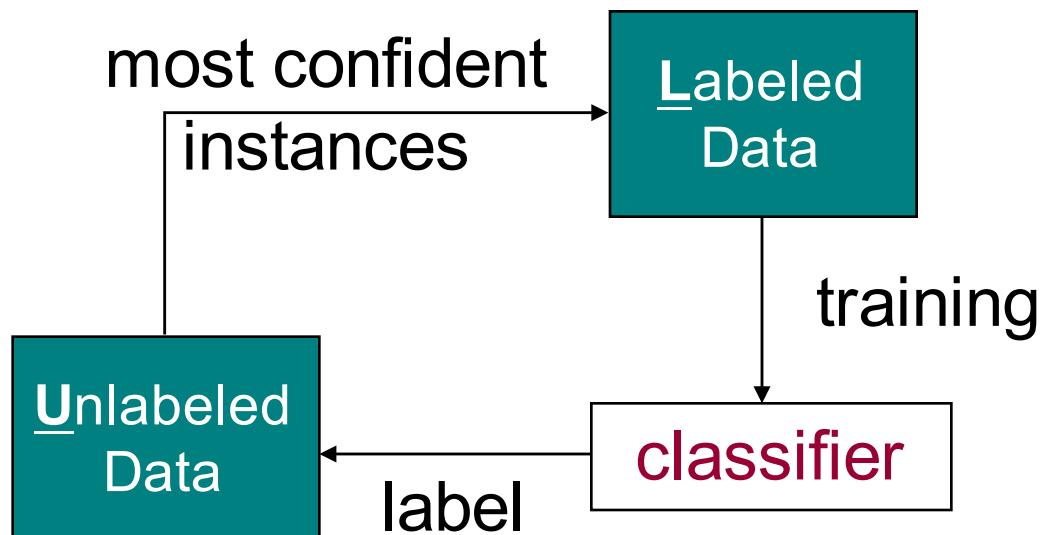
System/Team	Description	Fine		Coarse	
		P	R	P	R
htsa3 U.Bucharest (Grozea)	A Naive Bayes system, with correction of the a-priori frequencies, by dividing the output confidence of the senses by <i>frequency</i> ^{α} ($\alpha = 0.2$)	72.9	72.9	79.3	79.3
IRST-Kernels ITC-IRST (Strapparava)	Kernel methods 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. (Agiñe & 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

CS4740 Natural Language Processing

- Word sense disambiguation
 - Dictionary-based approaches
 - Supervised machine learning methods
 - WSD evaluation
 -  – Weakly supervised 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



Repeat:

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

Generating initial *seed* examples

- Hand label a small set of examples
 - Reasonable certainty that the seeds will be correct
 - Can choose prototypical examples
 - Reasonably easy to do

Generating initial *seed* examples

- Automatically via the **one sense per co-occurrence** constraint
 - 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 co-occurrence terms automatically from machine readable dictionary entries
 - Or select seeds automatically using co-occurrence statistics (see Ch 6 of J&M)

One sense per co-occurrence

Klucsevsek **plays** 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 player** 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 player** 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 fishing**.

And it all started when **fishermen** 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 **fisherman**," giving advice on catching **bass** in cold weather from the seat of a bass boat in Louisiana.

Yarowsky's bootstrapping approach

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

Yarowsky's bootstrapping approach

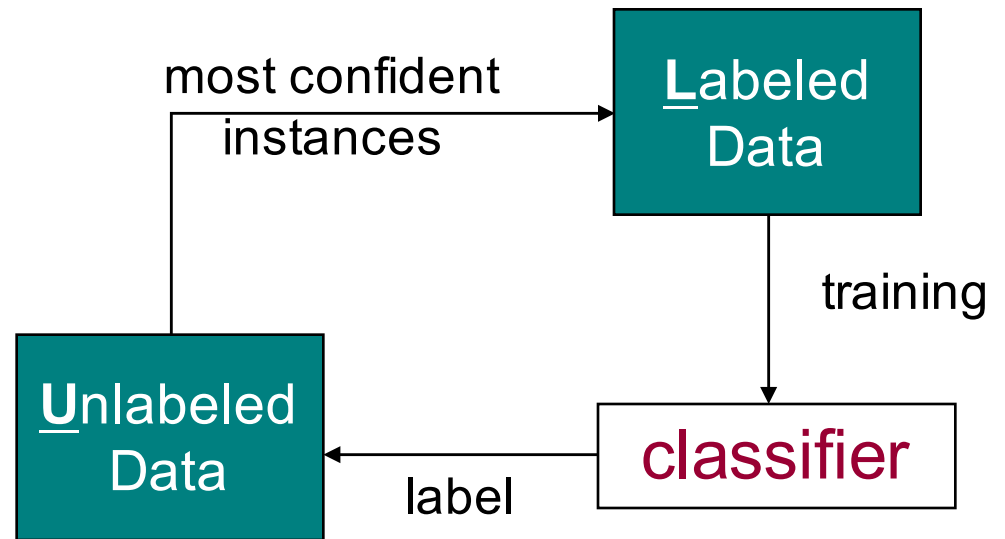
To disambiguate lexical item **W**:

L = the seed set of labeled examples

most confident instances =
instances classified with
probability $> threshold$

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

Repeat until the unlabeled data is stable.



96.5% accuracy: coarse senses involving 12 words