CS474 Natural Language Processing

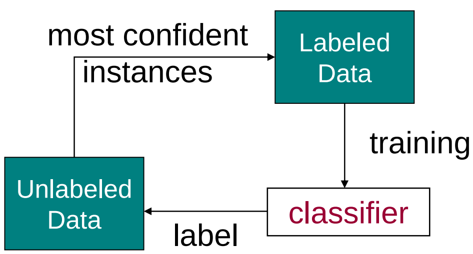
- Last class
  - Word sense disambiguation
    - » Supervised machine learning methods
    - » Issues for WSD evaluation
- Today
  - Critique paper discussion
  - Word sense disambiguation
    - » Precision and recall revisited
    - » Weakly supervised (bootstrapping) methods
    - » SENSEVAL

WSD Evaluation

- Precision
  - # of correct senses predicted / # of words in the test set for which the algorithm made a prediction
- Recall
  - # of correct senses predicted / # of words in the test set
  - recall=accuracy

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 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 **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 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. The unlabeled examples are called the *residual*.

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

4. Apply the classifier to all the examples. Find members of the residual 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 residual set is stable.

CS474 Natural Language Processing

- Last class
  - Word sense disambiguation
    - » Supervised machine learning methods
    - » Issues for WSD evaluation
- Today
  - Critique paper discussion
  - Word sense disambiguation
    - » Precision and recall revisited
    - » Weakly supervised (bootstrapping) methods
    - SENSEVAL

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 several instances of the sample words in short extracts of text
- SENSEVAL-1: 35 words, 41 tasks
  - 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)