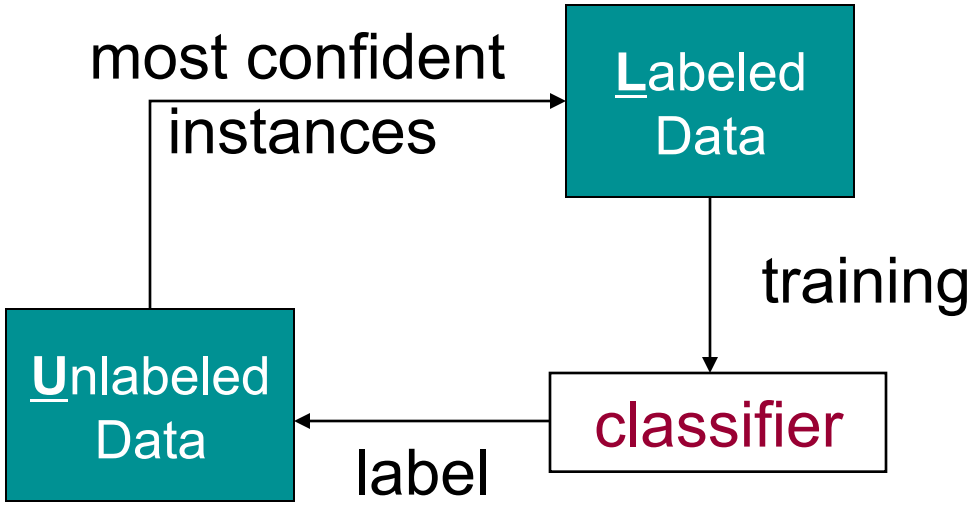
CS474 Natural Language Processing

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

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

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

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

4. Repeat until the unlabelled data is stable.

96.5% accuracy on coarse binary sense assignment involving 12 words

CS474 Natural Language Processing

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