Bootstrapping without the Boot

Jason Eisner
Damianos Karakos

JOHNS HOPKINS UNIVERSITY

HLT-EMNLP, October 2005

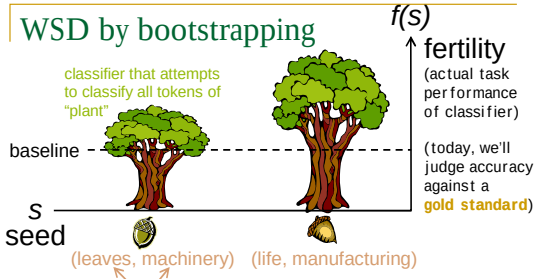
1

Executive Summary

(if you're not an executive, you may stay for the rest of the talk)

- What:
  - We like minimally supervised learning (bootstrapping).
  - Let's convert it to unsupervised learning ("strapping").
- How:
  - If the supervision is so minimal, let's just guess it!
  - Lots of guesses → lots of classifiers.
  - Try to predict which one looks plausible (!?!).
  - We can learn to make such predictions.
- Results (on WSD):
  - Performance actually goes up!
  - (Unsupervised WSD for translational senses, English Hansards, 14M words.)

WSD by bootstrapping



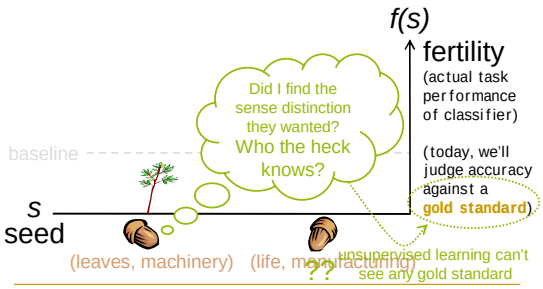
- we know "plant" has 2 senses
- we hand-pick 2 words that indicate the desired senses
- use the word pair to "seed" some bootstrapping procedure

3

How do we choose among seeds?

automatically

Want to maximize fertility but we can't measure it!



4

How do we choose among seeds?

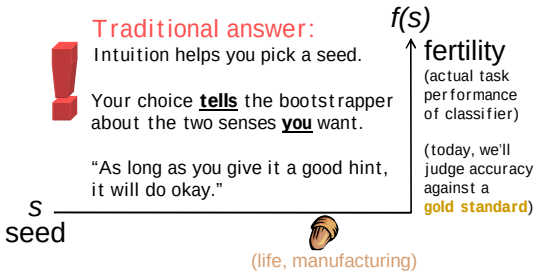
Want to maximize fertility but we can't measure it!

Traditional answer:

Intuition helps you pick a seed.

Your choice **tells** the bootstrapper about the two senses **you** want.

"As long as you give it a good hint, it will do okay."



5

Why not pick a seed by hand?

- Your intuition might not be trustworthy
  (even a sensible seed could go awry)
- You don't speak the language / sublanguage
- You want to bootstrap lots of classifiers
  - All words of a language
  - Multiple languages
  - On ad hoc corpora, i.e., results of a search query

- You're not sure that # of senses = 2
  - (life, manufacturing) vs. (life, manufacturing, sow)
    - which works better?

6

How do we choose among seeds?

Want to maximize fertility but we can't measure it!

Our answer:

Bad classifiers smell funny.
Stick with the ones that smell like real classifiers.

~~f(s)~~ h(s)

predicted ^ fertility

s
seed

"Strapping"

This name is supposed to remind you of bagging and boosting, which also train many classifiers.

(But those methods are supervised, & have theorems …)

1. Quickly pick a bunch of candidate seeds
2. For each candidate seed s:
   - grow a classifier C_s
   - compute h(s) (i.e., guess whether s was fertile)
3. Return C_s where s maximizes h(s)

Single classifier that we guess to be best.

Future work: Return a combination of classifiers?

Review: Yarowsky's bootstrapping algorithm

To test the idea, we chose to work on word-sense disambiguation and bootstrap decision-list classifiers using the method of Yarowsky (1995).

other tasks?

other classifiers?

other bootstrappers?

Possible future work

table taken from Yarowsky (1995)

Review: Yarowsky's bootstrapping algorithm

Sense	Training Examples (Keyword in Context)	
A	used to strain microscopic	*plant* **life** from the …
A	… rapid growth of aquatic	*plant* **life** in water …
A	… that divide **life** into	*plant* and animal kingdom
A	beds too salty to support	*plant* **life** . River …
A	…	…
?	… company said the	*plant* is still operating …
?	… molecules found in	*plant* and animal tissue
?	…	…
?	… Nissan car and truck	*plant* in Japan is …
?	… animal rather than	*plant* tissues can be …
B	…	…
B	automated **manufacturing**	*plant* in Fremont …
B	… vast **manufacturing**	*plant* and distribution …
B	chemical **manufacturing**	*plant* , producing viscose
B	… keep a **manufacturing**	*plant* profitable without

life (1%)

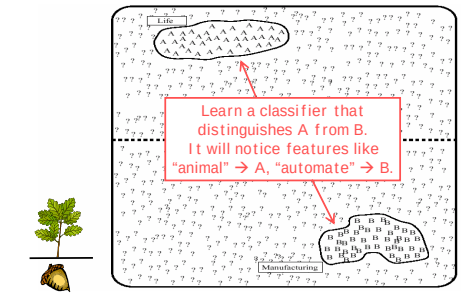
98%

manufacturing (1%)

(life, manufacturing)

figure taken from Yarowsky (1995)

Review: Yarowsky's bootstrapping algorithm

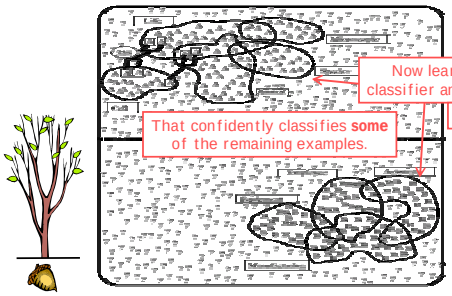


Learn a classifier that distinguishes A from B. It will notice features like "animal" → A, "automate" → B.

(life, manufacturing)

figure taken from Yarowsky (1995)

Review: Yarowsky's bootstrapping algorithm



Now learn a new classifier and repeat …

That confidently classifies **some** of the remaining examples.

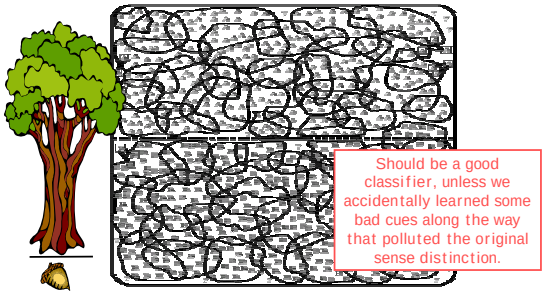
& repeat …

& repeat …

(life, manufacturing)

figure taken from Yarowsky (1995)

Review: Yarowsky's bootstrapping algorithm



Should be a good classifier, unless we accidentally learned some bad cues along the way that polluted the original sense distinction.

(life, manufacturing)

13

table taken from Yarowsky (1995)

Review: Yarowsky's bootstrapping algorithm

Final decision list for *plant* (abbreviated)

Log L	Collocation	Sense
10.12	*plant* growth	⇒ A
9.68	*car* (within ±*k* words)	⇒ B
9.64	*plant* height	⇒ A
9.61	*union* (within ±*k* words)	⇒ B
9.54	*equipment* (within ±*k* words)	⇒ B
9.51	*assembly plant*	⇒ B
9.50	*nuclear plant*	⇒ B
9.31	*flower* (within ±*k* words)	⇒ A
9.24	*job* (within ±*k* words)	⇒ B
9.03	*fruit* (within ±*k* words)	⇒ A
9.02	*plant species*	⇒ A

(life, manufacturing)

14

ambiguous words from Gale, Church, & Yarowsky (1992)

Data for this talk

- Unsupervised learning from 14M English words (transcribed formal speech).
- Focus on 6 ambiguous word types:
  - **drug, duty, land, language, position, sentence**

each has from 300 to 3000 tokens

To learn an English → French MT model, we would first hope to discover the 2 **translational senses** of each word.

$drug_1$ $drug_2$ — *medicament* *drogue*

$sentence_1$ $sentence_2$ — *peine* *phrase*

15

ambiguous words from Gale, Church, & Yarowsky (1992)

Data for this talk

- Unsupervised learning from 14M English words (transcribed formal speech).
- Focus on 6 ambiguous word types:
  - **drug, duty, land, language, position, sentence**

try to learn these distinctions **monolingually**

(assume insufficient bilingual data to learn when to use each translation)

$drug_1$ $drug_2$ — *medicament* *drogue*

$sentence_1$ $sentence_2$ — *peine* *phrase*

16

ambiguous words from Gale, Church, & Yarowsky (1992)

Data for this talk

Canadian parliamentary proceedings (Hansards)

- Unsupervised learning from 14M English words (transcribed formal speech).
- Focus on 6 ambiguous word types:
  - **drug, duty, land, language, position, sentence**

but evaluate **bilingually**:

for this corpus, happen to have a French translation → gold standard for the senses we want.

$drug_1$ $drug_2$ — *medicament* *drogue*

$sentence_1$ $sentence_2$ — *peine* *phrase*

17

Strapping word-sense classifiers

1. Quickly pick a bunch of candidate seeds
2. Fo
   -
   -
3. Re

Automatically generate 200 seeds *(x,y)*

Get *x*, *y* to select distinct senses of target *t*:
- *x* and *y* each have high MI with *t*
- **but** *x* and *y* never co-occur

Also, for safety:
- *x* and *y* are not too rare
- *x* isn't far more frequent than *y*

18

Strapping word-sense classifiers

1. Quickly pick a bunch of candidate seeds
2. For each candidate seed s:
   - grow a classifier C_s
   - c… …tile)
3. Retu…

replicate Yarowsky (1995)

(with fewer kinds of features,
and some small algorithmic differences)

19

Strapping word-sense classifiers

1. Quickly pick a bunch of candidate seeds
2. For each candidate seed s:
   - grow a classifier C_s
   - compute h(s) (i.e., guess whether s was fertile)
3. Retu…

h(s) is the interesting part.

	best	good	lousy
drug	(alcohol, medical)	(abuse, information)	(traffickers, trafficking)
sentence	(reads, served)	(quote, death)	(length, life)

Strapping word-sense classifiers

1. Quickly pick a bunch of candidate seeds
2. For each candidate seed s:
   - grow a classifier C_s
   - compute h(s) (i.e., guess whether s was fertile)
3. Return C_s where s maximizes h(s)

For comparison, hand-picked 2 seeds.

Casually selected (< 2 min.) – one author picked a reasonable *(x,y)* from the 200 candidates.

Carefully constructed (< 10 min.) – other author studied gold standard, then separately picked high-MI *x* and *y* that retrieved appropriate initial examples.

21

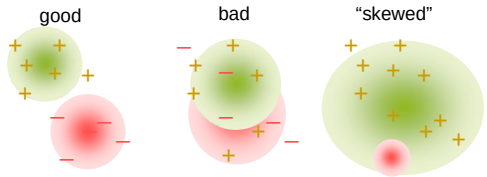
Strapping word-sense classifiers

1. Quickly pick a bunch of candidate seeds
2. For each candidate seed s:
   - grow a classifier C_s
   - compute h(s) (i.e., guess whether s was fertile)
3. Retu…

h(s) is the interesting part.

How can you possibly tell,
without supervision,
whether a classifier is any good?

Unsupervised WSD as clustering



- Easy to tell which clustering is "best"
- A good unsupervised clustering has high
  - p(data | label) – minimum-variance clustering
  - p(data) – EM clustering
  - MI(data, label) – information bottleneck clustering

23

oversimplified slide

Clue #1: **Confidence** of the classifier

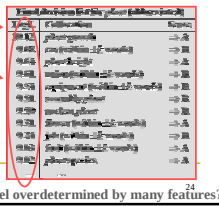
Yes! These tokens are sense A! And these are B!

though this could be overconfidence: may have found the wrong senses

Um, maybe I found some senses, but I'm not sure.

though maybe the senses are truly hard to distinguish

- Final decision list for C_s
- Does it confidently classify the training tokens, on average?
- Opens the "black box" classifier to assess confidence
  (but so does bootstrapping itself)



24

possible variants – e.g., is the label overdetermined by many features?

Clue #2: **Agreement** with other classifiers

- Intuition: for WSD, any reasonable seed *s* should find a true sense distinction.
- **So it should agree with some other reasonable seeds *r* that find the same distinction.**

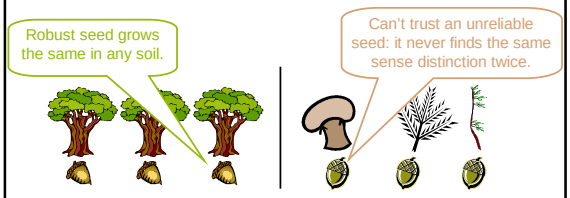
C_s + + - - + + - + + +
C_r + + - + + - - + + -

prob of agreeing this well by chance?

$$\left(\frac{1}{199}\sum_{r\neq s}\left(-\log p(\text{agr of } C_r, C_s \text{ by chance})\right)^{\alpha}\right)^{\frac{1}{\alpha}}$$

25

Clue #3: **Robustness** of the seed



- C_S was trained on the original dataset.
- Construct 10 new datasets by resampling the data ("bagging").
- Use seed s to bootstrap a classifier on each new dataset.
- How well, on average, do these agree with the original C_S? (again use prob of agreeing this well by chance)

possible variant – robustness under changes to feature space (not changes to data)

26

How well did we predict actual fertility f(s)?

Spearman rank correlation with f(s):

- 0.748 Confidence of classifier
- 0.785 Agreement with other classifiers
- 0.764 Robustness of the seed

- 0.794 Average rank of all 3 clues

27

Smarter combination of clues?

- Really want a "meta-classifier"!
  - **Output:** Distinguishes good from bad seeds.
  - **Input:** Multiple fertility clues for each seed (amount of confidence, agreement, robustness, etc.)

train: some other corpus **plant, tank** 200 seeds per word

learns "how good seeds behave" for the WSD task

we need gold standard answers so we know which seeds really were fertile

test: English Hansards **drug, duty, land, language, position, sentence** 200 seeds per word

guesses which seeds probably grew into a good sense distinction

28

Yes, the test is still unsupervised WSD ☺

no information provided about the desired sense distinctions

train: some labeled corpus **plant, tank** 200 seeds per word

learns "what good classifiers look like" for the WSD task

test: English Hansards **drug, duty, land, language, position, sentence** 200 seeds per word

Unsupervised WSD research has **always** relied on supervised WSD instances to learn about the space (e.g., what kinds of features & classifiers work).

29

How well did we predict actual fertility f(s)?

Spearman rank correlation with f(s):

- 0.748 Confidence of classifier
- 0.785 Agreement with other classifiers
- 0.764 Robustness of the seed
- 0.794 Average rank of all 3 clues
- 0.851% Weighted average of clues

Includes 4 versions of the "agreement" feature

good weights are learned from supervised instances **plant, tank**

just simple linear regression … might do better with SVM & polynomial kernel …

30

How good are the strapped classifiers???

drug
duty
sentence
land

Our top pick is the **very best** seed out of 200 seeds! Wow!
(i.e., it agreed **best** with an unknown **gold standard**)

language
position

Our top pick is the **7th best** seed of 200.
(The **very best** seed is our 2nd or 3rd pick.)

Statistically significant wins:

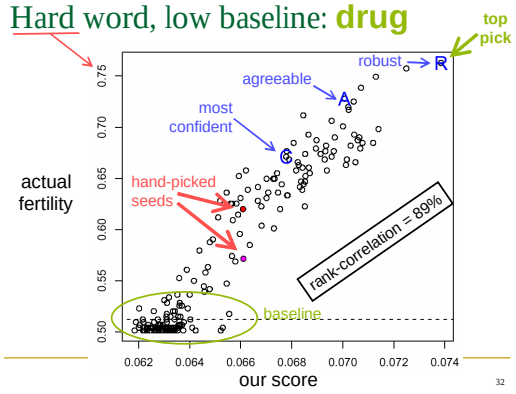
accuracy 76-90% — strapped classifier (top pick) — 12 of 12 times → classifiers bootstrapped from hand-picked seeds — accuracy 57-88%

strapped classifier (top pick) — 6 of 6 times → chance — baseline 50-87%

strapped classifier (top pick) — 5 of 12 times → classifiers bootstrapped from hand-picked seeds

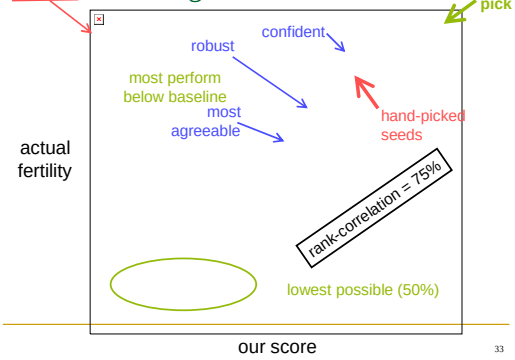
Good seeds are hard to find! Maybe because we used only 3% as much data as Yarowsky (1995), & fewer kinds of features.

Hard word, low baseline: drug



32

Hard word, high baseline: land



33

Reducing supervision for decision-list WSD

Eisner & Karakos (2005) unsupervised strapping

"beats"

Yarowsky (1995) minimally supervised bootstrapping

"rivals"

Gale et al. (1992) supervised classifiers

34

How about *no supervision at all*?

"cross-instance learning"
Each **word** is an instance of the WSD task.

train
~~some other corpus~~
~~plant, tank~~
200 seeds per word

test
English Hansards
drug, duty, land, language, position, sentence
200 seeds per word

Q: What if you had no labeled data to help you learn what a good classifier looks like?

A: Manufacture some artificial data! ... use pseudowords.

35

Automatic construction of pseudowords

Consider a target word:
sentence
Automatically pick a seed:
(death, page)
Merge into ambig. pseudoword:
deathpage

unlabeled corpus
blah sentence blah **death** blah
blah blah **page** blah sentence

labeled corpus
~~*"living"*~~
blah blah blah **plant** blah
~~*"factory"*~~
blah blah **plant** blah blah

labeled corpus
"death"
blah sentence blah **deathpage** blah
"page"
blah blah **deathpage** blah sentence

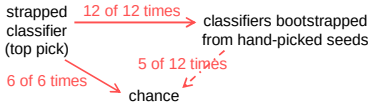
Use this to train the meta-classifier

36

pseudowords for eval.: Gale et al. 1992, Schütze 1998, Gaustad 2001, Nakov & Hearst 2003

Does pseudoword training work as well?

1. Average correlation w/ predicted fertility stays at 85%
2. **duty**, **sentence** — Our top pick is still the **very best** seed
   land — Our top pick is the 2nd best seed
   drug, **language**, **position** — Top pick works okay, but the very best seed is our 2nd or 3rd pick
3. Statistical significance diagram is unchanged:



37

Opens up lots of future work

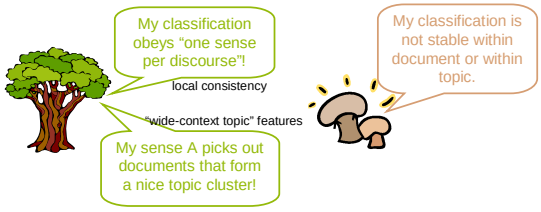
- Compare to other unsupervised methods (Schütze 1998)
- Other tasks (discussed in the paper!) ~ 10 other papers at this conference ...
  - Lots of people have used bootstrapping!
  - Seed grammar induction with basic word order facts?

- Make WSD even smarter:
  - Better seed generation (e.g., learned features ➔ new seeds)
  - Better meta-classifier (e.g., polynomial SVM)
  - Additional clues: Variant ways to measure confidence, etc.
  - Task-specific clues

38

oversimplified slide

Future work: **Task-specific clues**

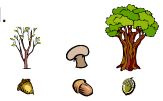


True senses have these properties.
We didn't happen to use them while bootstrapping.
So we can use them instead to validate the result.

39

Summary

- Bootstrapping requires a "seed" of knowledge.
- Strapping = try to guess this seed.
  - Try many reasonable seeds.
  - See which ones grow plausibly.
  - You can learn what's plausible.



- Useful because it eliminates the human:
  - You may need to bootstrap often.
  - You may not have a human with the appropriate knowledge.
  - Human-picked seeds often go awry, anyway.
- Works great for WSD! (Other unsup. learning too?)

40