Bootstrapping without the Boot

Jason Eisner
Damianos Karakos

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

classifier that attempts to classify all tokens of “plant”

baseline

f(s)

fertility

(0)

(0)

actual task performance of classifier

(today, we’ll judge accuracy against a gold standard)

s

seed

(leaves, machinery) (life, manufacturing)

- 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

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.”

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?
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).

Possible future work

“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?

<table>
<thead>
<tr>
<th>Sense (Training Examples)</th>
<th>Keyword in Context</th>
</tr>
</thead>
<tbody>
<tr>
<td>A used to train microscopic</td>
<td>plant life from the...</td>
</tr>
<tr>
<td>A ... rapid growth of aquatic</td>
<td>plant life in water...</td>
</tr>
<tr>
<td>A ... that divide life into</td>
<td>plant animal kingdom...</td>
</tr>
<tr>
<td>A ... beds too salty to support</td>
<td>plant life River...</td>
</tr>
<tr>
<td>? ... company said the</td>
<td>plant is still operating...</td>
</tr>
<tr>
<td>? ... molecules found in</td>
<td>plant and animal tissue...</td>
</tr>
<tr>
<td>? ... Nitrogen and truck</td>
<td>plant in Japan...</td>
</tr>
<tr>
<td>? ... animal rather than</td>
<td>plant tissues can be...</td>
</tr>
<tr>
<td>B automated manufacturing</td>
<td>plant in France...</td>
</tr>
<tr>
<td>B ... vast manufacturing</td>
<td>plant and distribution...</td>
</tr>
<tr>
<td>B chemical manufacturing</td>
<td>plant producing viscose...</td>
</tr>
<tr>
<td>B ... keep a manufacturing</td>
<td>plant profitable without...</td>
</tr>
</tbody>
</table>

life (1%)
manufacturing (1%)

Review: Yarowsky’s bootstrapping algorithm

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

- Try to learn these distinctions bilingual (assume insufficient bilingual data to learn when to use each translation)

- drug1 drug2
- medicament drgue
- sentence sentence2
- peine phrase

Strapping word-sense classifiers

1. Quickly pick a bunch of candidate seeds

2. Automatically generate 200 seeds \((x, y)\)
   - Get \(x\) and \(y\) to select distinct senses of target
   - \(x\) and \(y\) each have high MI with \(r\)
   - but \(x\) and \(y\) never co-occur

3. Refine
   - Also, for safety:
     - \(x\) and \(y\) are not too rare
     - \(x\) isn’t far more frequent than \(y\)

Each word has from 300 to 3000 tokens

Translational senses

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

Focus on 6 ambiguous word types:

- drug, duty, land, language, position, sentence

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

\(h(s)\) (i.e., guess whether \(s\) was fertile)

\(s\) maximizes \(h(s)\)

\(C_s\) where \(s\)

get \(C_s\)

of each word.

\(C_s\) where

\(s\)

\(h(s)\)

\(C_s\)

\(x\)

\(y\)

\(x\)

\(y\)

\(y\)

\(y\)

\(y\)

\(y\)

\(y\)

\(y\)

\(y\)
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 $h(s)$

replicate Yarowsky (1995)
(with fewer kinds of features, and some small algorithmic differences)

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.

Unsupervised WSD as clustering

- easy to tell which clustering is "best"
- a good unsupervised clustering has high
  - $p(\text{data} | \text{label})$ — minimum-variance clustering
  - $\text{MI}(\text{data}, \text{label})$ — information bottleneck clustering

Clue #1: Confidence of the classifier

- 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)

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.

\[ \frac{1}{|\mathcal{S}|} \sum_{s \neq r} \log \left( \frac{\text{agreement of } C_s, C_r \text{ by chance}}{\alpha} \right) \]

I like my neighbors.
I seem to be odd tree out around here …

Clue #3: Robustness of the seed

- Cs 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 Cs?

\[ \text{prob of agreeing this well by chance} \]

Can’t trust an unreliable seed: it never finds the same sense distinction twice.

Robust seed grows the same in any soil.

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

learned “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

Yes, the test is still unsupervised WSD

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

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

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 …
How good are the strapped classifiers???

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

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

Statistically significant wins:
- Good seeds are hard to find! Maybe because we used only 3% as much data as Yarowsky (1995), & fewer kinds of features.
- Drug, duty, sentence, land, language, position

Statistical classifier bootstrapped from hand-picked seeds
- Accuracy: 12 of 12 times
- Baseline: 5 of 12 times
- Chance: 6 of 6 times

Accuracy: 76-90%

Reducing supervision for decision-list WSD

Eisner & Karakos (2005) unsupervised strapping
Yarowsky (1995) minimally supervised bootstrapping
Gale et al. (1992) supervised classifiers

How about no supervision at all?

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.

Automatic construction of pseudowords

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

Use this to train the meta-classifier

**Does pseudoword training work as well?**

1. Average correlation w/ predicted fertility stays at 85%

2. Duty sentence land drug language position

   - Our top pick is still the very best seed
   - Our top pick is the 2nd best seed
   - Top pick works okay, but the very best seed is our 2nd or 3rd pick

3. Statistical significance diagram is unchanged:

   - Strapped classifier (top pick):
     - 12 of 12 times
   - Classifiers bootstrapped from hand-picked seeds:
     - 5 of 12 times
   - Chance:
     - 6 of 6 times

---

**Opens up lots of future work**

- Compare to other unsupervised methods (Schütze 1998)
- Other tasks (discussed in the paper!)
  - 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 \(\rightarrow\) new seeds)
  - Better meta-classifier (e.g., polynomial SVM)
  - Additional clues: Variant ways to measure confidence, etc.
  - Task-specific clues

---

**Future work: Task-specific clues**

- My classification obeys "one sense per discourse!"
- My classification is not stable within document or within topic.
- True senses have these properties.
  - We didn’t happen to use them while bootstrapping.
  - So we can use them instead to validate the result.

---

**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 away, anyway.
- Works great for WSD! (Other unsup. learning too?)