Bootstrapping Coreference Classifiers with Multiple Machine Learning Algorithms

Vincent Ng and Claire Cardie
Department of Computer Science
Cornell University
Plan for the Talk

- Noun phrase coreference resolution
- Standard machine learning framework
- Weakly supervised approaches
  - related work
  - our bootstrapping algorithm
- Evaluation
- An example ranking method for bootstrapping
Noun Phrase Coreference

Identify all noun phrases that refer to the same entity

Queen Elizabeth set about transforming her husband, King George VI, into a viable monarch. Logue, a renowned speech therapist, was summoned to help the King overcome his speech impediment...
Noun Phrase Coreference

Identify all noun phrases that refer to the same entity

Queen Elizabeth set about transforming her husband, King George VI, into a viable monarch. Logue, a renowned speech therapist, was summoned to help the King overcome his speech impediment...
Noun Phrase Coreference

Identify all noun phrases that refer to the same entity

Queen Elizabeth set about transforming her husband, King George VI, into a viable monarch. Logue, a renowned speech therapist, was summoned to help the King overcome his speech impediment...
Noun Phrase Coreference

Identify all noun phrases that refer to the same entity

Queen Elizabeth set about transforming her husband, King George VI, into a viable monarch. Logue, a renowned speech therapist, was summoned to help the King overcome his speech impediment...
Noun Phrase Coreference

Identify all noun phrases that refer to the same entity

Queen Elizabeth set about transforming her husband, King George VI, into a viable monarch. Logue, a renowned speech therapist, was summoned to help the King overcome his speech impediment...
Plan for the Talk

- Noun phrase coreference resolution
- Standard machine learning framework
- Weakly supervised approaches
  - related work
  - our bootstrapping algorithm
- Evaluation
- An example ranking method for bootstrapping
Standard Machine Learning Framework

Classification

- given a description of two noun phrases, $NP_i$ and $NP_j$, classify the pair as coreferent or not coreferent

[Queen Elizabeth] set about transforming [her] [husband], ...

Aone & Bennett [1995]; Connolly et al. [1994]; McCarthy & Lehnert [1995]; Ng & Cardie [2002]; Soon, Ng & Lim [2001]
Standard Machine Learning Framework

- Clustering
  - coordinates pairwise coreference decisions

\[ \text{Queen Elizabeth}, \quad \text{her} \]
\[ \text{King George VI}, \quad \text{husband} \]
\[ \text{Logue}, \quad \text{a renowned speech therapist} \]
Supervised vs. Weakly Supervised Approaches

- Differ only in the amount of labeled data used to train the coreference classifier

- The clustering mechanism is the same in both cases
Plan for the Talk

- Noun phrase coreference resolution
- Standard machine learning framework
- Weakly supervised approaches
  - related work
  - our bootstrapping algorithm
- Evaluation
- An example ranking method for bootstrapping
Related Work (Harabagiu et al., 2001)

- Bootstrap *knowledge sources* for coreference resolution of common nouns using WordNet
Related Work (Müller et al., 2002)

- Use co-training to bootstrap classifiers for resolution of German anaphors.

- Co-training shows no performance improvements for any type of anaphor except pronouns over a baseline classifier trained on a small set of labeled data.

- Suggest that view factorization is non-trivial for reference resolution for which no natural feature split has been found.
  - do not investigate different methods for feature splitting.
Related Work (Ng and Cardie, HLT-NAACL 2003)

- Investigate bootstrapping methods for coreference resolution
  - different methods for view factorization for co-training
  - single-view bootstrapping methods
    - self-training with bagging (Banko and Brill, 2001)
    - weakly supervised EM (Nigam et al., 2000)

- Co-training is sensitive to the choice of views

- Single-view weakly supervised learners are a viable alternative to co-training for bootstrapping coreference classifiers
Goal of the Study

- Further investigate methods for bootstrapping coreference classifiers that do not require explicit view factorization
  
  - use different learning algorithms in lieu of different views (Steedman et al., 2003; Goldman and Zhou, 2000)
  - propose a general method for ranking unlabeled instances to be fed back into the bootstrapping loop
Plan for the Talk

- Noun phrase coreference resolution
- Standard machine learning framework
- Weakly supervised approaches
  - related work
  - our bootstrapping algorithm
- Evaluation
- An example ranking method for bootstrapping
A Bootstrapping Algorithm for Coreference

- Does not require explicit view factorization
- Combines ideas of two existing co-training algorithms
  - Steedman et al. (EACL, 2003)
  - Goldman and Zhou (ICML, 2000)
The Blum and Mitchell Co-Training Algorithm

Given: $L$ (labeled data), $U$ (unlabeled data), $V_1$, $V_2$ (views)
The Blum and Mitchell Co-Training Algorithm

Given: \( L \) (labeled data), \( U \) (unlabeled data), \( V_1, V_2 \) (views)

repeat
The Blum and Mitchell Co-Training Algorithm

Given: $L$ (labeled data), $U$ (unlabeled data), $V_1$, $V_2$ (views)

repeat
  - train a classifier $h_1$ on $V_1$ of $L$
  - train a classifier $h_2$ on $V_2$ of $L$
The Blum and Mitchell Co-Training Algorithm

Given: $L$ (labeled data), $U$ (unlabeled data), $V_1$, $V_2$ (views)

repeat

- train a classifier $h_1$ on $V_1$ of $L$
- train a classifier $h_2$ on $V_2$ of $L$
- form a data pool $D$ by randomly selecting $d$ instances from $U$
The Blum and Mitchell Co-Training Algorithm

- Given: \( L \) (labeled data), \( U \) (unlabeled data), \( V_1, V_2 \) (views)

- repeat
  - train a classifier \( h_1 \) on \( V_1 \) of \( L \)
  - train a classifier \( h_2 \) on \( V_2 \) of \( L \)
  - form a data pool \( D \) by randomly selecting \( d \) instances from \( U \)
  - use \( h_1 \) to label instances in \( D \)
  - use \( h_2 \) to label instances in \( D \)
The Blum and Mitchell Co-Training Algorithm

Given: \( L \) (labeled data), \( U \) (unlabeled data), \( V_1, V_2 \) (views)

repeat
- train a classifier \( h_1 \) on \( V_1 \) of \( L \)
- train a classifier \( h_2 \) on \( V_2 \) of \( L \)
- form a data pool \( D \) by randomly selecting \( d \) instances from \( U \)
- use \( h_1 \) to label instances in \( D \)
- use \( h_2 \) to label instances in \( D \)
- add the \( g \) most confidently labeled instances by \( h_1 \) to \( L \)
- add the \( g \) most confidently labeled instances by \( h_2 \) to \( L \)
The Blum and Mitchell Co-Training Algorithm

Given: \( L \) (labeled data), \( U \) (unlabeled data), \( V_1, V_2 \) (views)

repeat

- train a classifier \( h_1 \) on \( V_1 \) of \( L \)
- train a classifier \( h_2 \) on \( V_2 \) of \( L \)
- form a data pool \( D \) by randomly selecting \( d \) instances from \( U \)
- use \( h_1 \) to label instances in \( D \)
- use \( h_2 \) to label instances in \( D \)
- add the \( g \) most confidently labeled instances by \( h_1 \) to \( L \)
- add the \( g \) most confidently labeled instances by \( h_2 \) to \( L \)
- replenish \( D \) by \( 2g \) instances
The Steedman et al. Co-Training Algorithm

- A variation of the Blum and Mitchell algorithm applied to statistical parsing

- Differs from Blum and Mitchell in three respects
  - use **two diverse parsers** to substitute for the two views
    - the two parsers correspond to coarsely **different features**
  - data pool is **flushed** after each iteration
  - each parser labels unlabeled sentences for **the other parser**
Our Bootstrapping Algorithm

- A variation of the Steedman et al. algorithm

- Use two different learning algorithms that have access to the same feature set (cf. Goldman and Zhou (2000))

- The learners should be chosen so that the classifiers are
  - accurate
  - complementary
Our Bootstrapping Algorithm

- A variation of the Steedman et al. algorithm

- Use two different learning algorithms that have access to the same feature set (cf. Goldman and Zhou (2000))

- The learners should be chosen so that the classifiers are
  - accurate
  - complementary

- Learning algorithms
  - naïve Bayes
  - decision list learner (Collins and Singer, 1999)
Plan for the Talk

- Noun phrase coreference resolution
- Standard machine learning framework
- Weakly supervised approaches
  - related work
  - our bootstrapping algorithm
- Evaluation
- An example ranking method for bootstrapping
Evaluation

- Evaluate the performance of our single-view, multi-learner bootstrapping algorithm (SVML) on coreference resolution

- Compare SVML against three baselines
  - No bootstrapping
  - Co-training
  - Self-training
## Bootstrapping Experiments

<table>
<thead>
<tr>
<th></th>
<th>Bootstrapping?</th>
<th>Multiple Views?</th>
<th>Multiple Learners?</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Bootstrapping</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>Co-Training</td>
<td>✓</td>
<td>✓</td>
<td>✗</td>
</tr>
<tr>
<td>SVML</td>
<td>✓</td>
<td>✗</td>
<td>✓</td>
</tr>
<tr>
<td>Self-Training</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
</tr>
</tbody>
</table>
Data Sets

- MUC-6 and MUC-7 coreference data sets
  - documents annotated with coreference information
  - MUC-6: 30 dryrun texts + 30 evaluation texts
  - MUC-7: 30 dryrun texts + 20 evaluation texts

- Evaluation texts are reserved for testing

- From the dryrun texts
  - 1000 randomly selected instances as labeled data (L)
  - remaining instances as unlabeled data (U)

- Results averaged across five independent runs
### Bootstrapping Experiments

<table>
<thead>
<tr>
<th></th>
<th>Bootstrapping?</th>
<th>Multiple Views?</th>
<th>Multiple Learners?</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Bootstrapping</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Co-Training</td>
<td>✓</td>
<td>✓</td>
<td>X</td>
</tr>
<tr>
<td>SVML</td>
<td>✓</td>
<td>X</td>
<td>✓</td>
</tr>
<tr>
<td>Self-Training</td>
<td>✓</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>
### Results: No Bootstrapping

Train a classifier on 1000 instances using all of the features

<table>
<thead>
<tr>
<th></th>
<th>MUC-6</th>
<th>MUC-7</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Naive Bayes</td>
<td>Decision List</td>
</tr>
<tr>
<td></td>
<td>R  P  F</td>
<td>R  P  F</td>
</tr>
<tr>
<td>No Bootstrapping</td>
<td>50.7 52.6 <strong>51.6</strong></td>
<td>17.9 72.0 <strong>28.7</strong></td>
</tr>
</tbody>
</table>
## Bootstrapping Experiments

<table>
<thead>
<tr>
<th></th>
<th>Bootstrapping?</th>
<th>Multiple Views?</th>
<th>Multiple Learners?</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Bootstrapping</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>Co-Training</td>
<td>✓</td>
<td>✓</td>
<td>✗</td>
</tr>
<tr>
<td>SVML</td>
<td>✓</td>
<td>✗</td>
<td>✓</td>
</tr>
<tr>
<td>Self-Training</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
</tr>
</tbody>
</table>
Experiments: Co-Training

Training

- bootstrap two view classifiers using L and U under different combinations of views, pool sizes and growth sizes
- input parameters
  - views (3 heuristic methods for view factorization): Mueller et al.’s (2002) greedy method, random splitting, splitting according to the feature type
  - data pool size: 500, 1000, 5000
  - growth size: 10, 50, 100, 200

Testing

- each classifier makes an independent decision
- final prediction: decision associated the higher confidence
Results: Co-Training

<table>
<thead>
<tr>
<th></th>
<th>MUC-6</th>
<th>MUC-7</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Naive Bayes</td>
<td>Decision List</td>
</tr>
<tr>
<td></td>
<td>R</td>
<td>P</td>
</tr>
<tr>
<td>No Bootstrapping</td>
<td>50.7</td>
<td>52.6</td>
</tr>
<tr>
<td>Co-Training</td>
<td>33.3</td>
<td>90.7</td>
</tr>
</tbody>
</table>

Co-training produces improvements over the baseline in only two of the four classifier/data set combinations.
## Bootstrapping Experiments

<table>
<thead>
<tr>
<th></th>
<th>Bootstrapping?</th>
<th>Multiple Views?</th>
<th>Multiple Learners?</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Bootstrapping</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>Co-Training</td>
<td>✓</td>
<td>✓</td>
<td>✗</td>
</tr>
<tr>
<td>SVML</td>
<td>✓</td>
<td>✗</td>
<td>✓</td>
</tr>
<tr>
<td>Self-Training</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
</tr>
</tbody>
</table>
Experiments: SVML

- **Training**
  - bootstrap two classifiers with the same view using L and U under different combinations of pool sizes and growth sizes
  - input parameters
    - data pool size: 500, 1000, 5000
    - growth size: 10, 50, 100, 200

- **Testing**
  - one of the classifiers is chosen to make predictions
Results: SVML

<table>
<thead>
<tr>
<th></th>
<th>MUC-6</th>
<th></th>
<th>MUC-7</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Naive Bayes</td>
<td>Decision List</td>
<td>Naive Bayes</td>
<td>Decision List</td>
</tr>
<tr>
<td></td>
<td>R</td>
<td>P</td>
<td>F</td>
<td>R</td>
</tr>
<tr>
<td>No Bootstrapping</td>
<td>50.7</td>
<td>52.6</td>
<td><strong>51.6</strong></td>
<td>17.9</td>
</tr>
<tr>
<td>Co-Training</td>
<td>33.3</td>
<td>90.7</td>
<td><strong>48.7</strong></td>
<td>19.5</td>
</tr>
<tr>
<td>SVML</td>
<td>53.6</td>
<td>79.0</td>
<td><strong>63.9</strong></td>
<td>40.1</td>
</tr>
</tbody>
</table>

- SVML outperforms co-training in all cases
  - simultaneous rise in recall and precision
## Bootstrapping Experiments

<table>
<thead>
<tr>
<th></th>
<th>Bootstrapping?</th>
<th>Multiple Views?</th>
<th>Multiple Learners?</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Bootstrapping</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>Co-Training</td>
<td>✓</td>
<td>✓</td>
<td>✗</td>
</tr>
<tr>
<td>SVML</td>
<td>✓</td>
<td>✗</td>
<td>✓</td>
</tr>
<tr>
<td>Self-Training</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
</tr>
</tbody>
</table>
Experiments: Self-Training

- Additional check that the decision lists and naïve Bayes classifiers are benefiting from each other

- At each self-training iteration, the classifier
  - labels all 5000 instances in the data pool
  - adds the most confidently labeled 50 instances to the labeled data
# Results: Self-Training

<table>
<thead>
<tr>
<th></th>
<th>MUC-6</th>
<th></th>
<th>MUC-7</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Naive Bayes</td>
<td>Decision List</td>
<td>Naive Bayes</td>
<td>Decision List</td>
</tr>
<tr>
<td></td>
<td>R</td>
<td>P</td>
<td>F</td>
<td>R</td>
</tr>
<tr>
<td><strong>No Bootstrapping</strong></td>
<td>50.7</td>
<td>52.6</td>
<td><strong>51.6</strong></td>
<td>17.9</td>
</tr>
<tr>
<td><strong>Co-Training</strong></td>
<td>33.3</td>
<td>90.7</td>
<td><strong>48.7</strong></td>
<td>19.5</td>
</tr>
<tr>
<td><strong>SVML</strong></td>
<td>53.6</td>
<td>79.0</td>
<td><strong>63.9</strong></td>
<td>40.1</td>
</tr>
<tr>
<td><strong>Self-Training</strong></td>
<td>48.3</td>
<td>63.5</td>
<td><strong>54.9</strong></td>
<td>18.7</td>
</tr>
</tbody>
</table>

- Self-training only yields marginal gains over the baseline.
Plan for the Talk

- Noun phrase coreference resolution
- Standard machine learning framework
- Weakly supervised approaches
  - related work
  - our bootstrapping algorithm
- Evaluation
- An example ranking method for bootstrapping
F-measure Learning Curves (MUC-6)
An Alternative Ranking Method

u Goal
  ▪ alleviate the problem of performance deterioration

u Hypothesis
  ▪ the drop is caused by the degradation in the quality of the bootstrapped data (cf. Pierce and Cardie, 1999)
  ▪ a more “conservative” example ranking method can help

u Motivated by Steedman et al. (HLT-NAACL 2003)
  ▪ use example selection methods to explore the trade-off between maximizing coverage and maximizing accuracy
The Ranking Method

- Ranks instances based on three preferences

- Preference 1: favors instances whose label is agreed upon by both classifiers

- Preference 2: favors instances that are confidently labeled by one classifier but not both

- Preference 3: ranks according to Blum and Mitchell’s rank-by-confidence method
Effects of the Ranking Methods (MUC-6)
Summary

- Proposed a single-view, multi-learner bootstrapping algorithm for coreference resolution and showed that the algorithm is a better alternative to co-training for this task.

- Investigated an example ranking method for bootstrapping that can potentially alleviate the problem of performance deterioration in the course of bootstrapping.