Data Selection in Semi-supervised Learning for Name Tagging

Heng Ji and Ralph Grishman
(hengji, grishman)@cs.nyu.edu
July 2006
Outline

- Problem in Using Very Very Large Unlabeled Data
- Prior Work
  - Two Semi-supervised Learning Methods for Name Tagging
    - Bootstrapping
    - Self-Training
  - Unlabeled Data Selection
    - Document Selection
    - Sentence Selection
- Experimental Results
- Conclusions and Future work
Goal: To Boost a Good-Performance Name Tagger

We want a simple, **effective** and **low-cost** scheme

Labeled Data

HMM Name Tagger with 88% F-Measure

Name Tagger with > 90% F-Measure
Solution 1: Hire Human Annotator?

Labeled Data

HMM Name Tagger
With 88% F-Measure

Unlabeled Data

High-cost mental work

John came from New York University in USA. He is now in Sydney to join ACL conference.
Solution 2: Add VERY VERY Large Unlabeled Data Directly?

Labeled Data

<PER>John</PER>
came from <ORG>New York University</ORG> in <GPE>USA</GPE>.

HMM Name Tagger with 88% F-Measure

Unlabeled Data

Name Tagger with 88.1% F-Measure

Non-effective
Prior Work

- **Very large data always helps**
  - Banko and Brill (2001): a logarithmic trend in performance as corpus size increases without performance reaching an upper bound
  - Thesaurus extraction (Curran and Moens, 2002)
  - Is-a relation extraction (Ravichandran et al., 2004)

- **Very large data does not always help**
  - (Curran, 2002) and (Curran and Osborne, 2002): the choice of statistical model is more important
  - Our work: Data selection is more important
Our General Solution

- Where to get the unlabeled data
  - Extra large unlabeled data → Bootstrapping
  - Test set itself → Self-Training

- How to select uniformly ‘useful’ data
  - Clean and reliable
  - Relevant to test set
  - Automatically select
Semi-supervised Learning for Name Tagging

Motivation

The text in which the tagger has low confidence may get ‘support’ from those texts previously labeled with high confidence

- Learn a previously-unseen token tagged with high confidence as a predictive token
- Learn a previously-unseen context that appears consistently as a predictive context

Iterative Procedure

- Take the unlabeled text about which the tagger has greatest confidence in its decisions, tag it
- Add the tagged text to the training set
- Re-train the tagger
Bootstrapping for Name Tagging

Unlabeled Data → Document Selection

\[ C_1 \rightarrow \ldots \rightarrow C_i \rightarrow \ldots \rightarrow C_n \]

NameM ← baseline tagger

OldNameM ← NameM

\[ C_i' \leftarrow C_i \text{ tagged with NameM} \]

\[ C_i'' \leftarrow \text{sentences selected from } C_i' \]

Retrain NameM

NameM = OldNameM

i = i + 1

NameM performs better on dev set?

No

Yes
Self-Training for Name Tagging

Test Set

Document Clustering

T_1

\[ \ldots \]

T_i

\[ \ldots \]

T_n

NameM \leftarrow \text{baseline tagger}

T'_i \leftarrow T_i \text{tagged with NameM}

\text{Adjust Selection Measure Threshold}

T_"i" \leftarrow 5\% \text{sentences selected from } T'_i

\text{Retrain NameM}

\text{Add } T_"i" \text{to training corpus}

\text{Save } T_i' \text{as System output}

i = i + 1

T_"i" \text{ Empty?}

Yes

No
Select Relevant Documents

Test Set

Query Set Q

Useful ('Relevant') Unlabeled Data

Baseline Name Tagger

Name Candidates from N-Best lists

Unlabeled Data
Cross-entropy to Count Document Relevancy

Definition

\[ H(\text{TestSet}, d) = - \sum_{x \in Q} \text{prob}(x \mid \text{TestSet}) \times \log_{2} \text{prob}(x \mid d) \]

- \( Q \): the name candidates in the top N best lists of the sentences in the test set
- \( x \): a name candidate in the \( Q \)
- \( d \): a document in the unlabeled corpus

Usage

- In bootstrapping: to select documents in the unlabeled corpus
- In self-training: to cluster documents in the test set
Select Reliable and Informative Sentences

**Margin**
- Definition: the difference in log probabilities between the first hypothesis and the second hypothesis
- The sentences with high margins are a mix of interesting and uninteresting samples

**Name Coreference**
- k names tagged in sentence S: \{N1, N2, ..., Nk\}, which are coreferred to by \{CorefNum1, CorefNum2, ..., CorefNumk\} other names, count:

\[
\text{AveCoref}_S = \left( \sum_{i=1}^{k} \text{CorefNum}_i \right) / k
\]

- Select S with large \text{AveCoref}_S

**Name count and sentence length**
- Don’t use a sentence if it includes fewer than six words, or doesn’t include any names
Experiments: Baseline System

- A multiple-hypotheses HMM name tagger to process English and Chinese documents
- Name types: Person (PER), Organization (ORG) and Geo-political (GPE)
## Experiments: Data

<table>
<thead>
<tr>
<th>Data</th>
<th>English</th>
<th>Chinese</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Baseline Training data</strong></td>
<td>ACE02,03,04 989,003 words</td>
<td>Beijing Corpus +ACE03,04,05 1,460,648 words</td>
</tr>
<tr>
<td><strong>Unlabeled Data</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>196,494 docs in Mar-Jun of 2003 (69M words) from ACE05 unlabeled data</td>
<td>41061 docs in Nov,Dec of 2000, and Jan of 2001 (25M words) from ACE05 and TDT4 transcripts</td>
</tr>
<tr>
<td>Selected Docs</td>
<td>62584 docs (1,314,148 Sentences)</td>
<td>14,537 docs (222,359 sentences)</td>
</tr>
<tr>
<td>Selected Sentences</td>
<td>290,973 sentences (6,049,378 words)</td>
<td>55,385 sentences (1,128,505 words)</td>
</tr>
<tr>
<td><strong>Dev Set</strong></td>
<td>20 ACE04 texts in Oct of 2000</td>
<td>90 ACE05 texts in Oct of 2000</td>
</tr>
<tr>
<td><strong>Test Set</strong></td>
<td>20 ACE04 texts in Oct of 2000 and 80 ACE05 texts in Mar-May of 2003 (3093 names, 1205 PERs, 1021GPEs, 867 ORGs)</td>
<td>90 ACE05 texts in Oct of 2000 (3093 names, 1013 PERs, 695 GPEs, 769 ORGs)</td>
</tr>
</tbody>
</table>
# Experiment: Overall Performance

<table>
<thead>
<tr>
<th>Learner</th>
<th>English</th>
<th></th>
<th></th>
<th>Chinese</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
<td>F</td>
<td>P</td>
<td>R</td>
<td>F</td>
</tr>
<tr>
<td>Baseline</td>
<td>87.3</td>
<td>87.6</td>
<td>87.4</td>
<td>88.2</td>
<td>87.6</td>
<td>87.9</td>
</tr>
<tr>
<td>Bootstrapping with data selection</td>
<td>88.2</td>
<td>88.6</td>
<td>88.4</td>
<td>89.8</td>
<td>89.5</td>
<td>89.6</td>
</tr>
<tr>
<td>Self-training</td>
<td>88.1</td>
<td>88.4</td>
<td>88.2</td>
<td>89.5</td>
<td>88.3</td>
<td>88.9</td>
</tr>
<tr>
<td>Bootstrapping with data selection + Self-training</td>
<td>89.0</td>
<td>89.2</td>
<td>89.1</td>
<td>90.2</td>
<td>89.7</td>
<td>90.0</td>
</tr>
</tbody>
</table>

- Passed significance testing
Experiments: Impact of Data Size in Bootstrapping

- **English**
  - Dev Set
  - Test Set

- **Chinese**
  - Dev Set
  - Test Set

Data Size

F-Measure
Experiments: Impact of Data Selection in Bootstrapping

Chinese

Performance (%)
Experiments: Impact of Sentence Selection in Self-Training
Experiments: Adjust Margin Threshold in Self-Training

<table>
<thead>
<tr>
<th>No. of Iterations</th>
<th>Number of Sentences added</th>
<th>Number of Tags added</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>91.4</td>
</tr>
<tr>
<td>1</td>
<td>37</td>
<td>28</td>
<td>91.9</td>
</tr>
<tr>
<td>2</td>
<td>69</td>
<td>22</td>
<td>92.1</td>
</tr>
<tr>
<td>3</td>
<td>107</td>
<td>21</td>
<td>92.4</td>
</tr>
<tr>
<td>4</td>
<td>128</td>
<td>11</td>
<td>92.6</td>
</tr>
<tr>
<td>5</td>
<td>146</td>
<td>9</td>
<td>92.7</td>
</tr>
<tr>
<td>6</td>
<td>163</td>
<td>8</td>
<td>92.8</td>
</tr>
<tr>
<td>7</td>
<td>178</td>
<td>6</td>
<td>92.8</td>
</tr>
<tr>
<td>8</td>
<td>190</td>
<td>0</td>
<td>92.8</td>
</tr>
</tbody>
</table>

- Test on one English cluster including 7 documents and 190 sentences
- Lower the margin threshold so that about 5% of the sentences (with the largest margin) are added to the training corpus
Conclusions and Future Work

- Demonstrated the effectiveness of two semi-supervised learning methods for name tagging
- Investigated the importance of data selection
  Exploiting a very large corpus did not by itself produce the greatest performance gain
- We select “easy” sentences to add automatically to the training set, which can be combined with active learning approach (select “difficult” sentences + label manually)
- Future Work
  - Try topic identification techniques to select relevant unlabeled documents
  - Use the downstream IE components to measure tagging confidence
  - Apply clustering as a pre-processing step for bootstrapping
Thanks!