Scaling to Very Very Large Corpora for Natural Language Disambiguation

Michele Banko and Eric Brill
Microsoft Research

Background of Issue

- Compare algorithms
  - Conclusions need common test corpora
    • Fixes the size of training/test sets
- Meanwhile available data growing
- Large cost of annotating data hinders development of new corpora

Confusion Set Disambiguation

- Example sets:
  - {principle, principal}
  - {to, two, too}
  - {weather, whether}
- Key Property #1: disambiguate from a small set of potential values
- Key Property #2: labeled data is available and free

Better Results at Low Cost?

- How to get better performance?
  - Get a Ph.D. and invent a new algorithm
  - Tune parameters and optimize old ways
    • it’s easy to fix a bad implementation
  - Why not just train on more data, esp. if it’s available?
Details of the Paper

• Learning Methods
  – Perceptron
  – Winnow
  – naïve Bayes
  – Memory (remembers previous and next words)
• Corpus Size
  – 1 Million → 1 Billions words

Learning Curves

Cost of Larger Corpus

Optimizations for Less Data

Voting

• What is Voting?
  – Train a set of classifiers on the same corpus, then for a test classification use democracy
• Complementarity (how often they agree)
  – Direct relationship with training corpus size
Efficacy of Voting

Not So Fast…

- Although this supports a conclusion to use more data, how realistic would that be?
- Remember the “Key Properties” from earlier?
- It is only for a few problems that access to large amounts of labeled data exists.
- Manual annotation is seemingly impractical
- Let’s try to take advantage of it anyway…

Active Learning

- “involves intelligently selecting a portion of samples for annotation from a pool of as-yet unannotated training samples.”
- Essentially, maximizing the utility of any fixed amount of manual effort

Active Learning Examples

- Run a seed learner over the test data, and use confidence ratings as indicators of usefulness
- Alternatively, run a set of seed learners and use their agreement as an indicator
Bagging

- Generates many classifiers
- To measure uncertainty of a classification
- Select, with replacement, random sentences from the original corpus
- Generate N training sets this way, all of size equal to the original corpus

Active Learning

- Co-training and Bootstrapping
- Start with a training set of high confidence examples (perhaps manually annotated)
- Iterate:
  - Train and run your classifier over the test set
  - Add those samples of highest confidence from the test set into the training set

Weakly Supervised Learning

<table>
<thead>
<tr>
<th>Classifiers</th>
<th>Test Accuracy</th>
<th>In Agreement</th>
<th>Test Accuracy</th>
<th>In Agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.8734</td>
<td></td>
<td>0.8183</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>0.6892</td>
<td></td>
<td>0.6313</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0.6286</td>
<td></td>
<td>0.5937</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0.6027</td>
<td></td>
<td>0.5720</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.5497</td>
<td></td>
<td>0.5400</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.5000</td>
<td></td>
<td>0.5000</td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Committee Agreement vs. Accuracy

<table>
<thead>
<tr>
<th>Test Accuracy</th>
<th>% Total Training Data</th>
<th>Test Accuracy</th>
<th>% Total Training Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.9624</td>
<td>0.1</td>
<td>0.8183</td>
<td>0.1</td>
</tr>
<tr>
<td>0.9588</td>
<td>0.6</td>
<td>0.8333</td>
<td>0.5</td>
</tr>
<tr>
<td>0.9465</td>
<td>1.2</td>
<td>0.8333</td>
<td>1.0</td>
</tr>
<tr>
<td>0.9719</td>
<td>2.2</td>
<td>0.8270</td>
<td>9.2</td>
</tr>
<tr>
<td>0.9588</td>
<td>6.1</td>
<td>0.8248</td>
<td>42.9</td>
</tr>
<tr>
<td>0.9878</td>
<td>100</td>
<td>0.9021</td>
<td>100</td>
</tr>
</tbody>
</table>

Table 3. Committee-Based Unsupervised Learning
Weakly Supervised Learning

Summary

- Often more data is available than researchers are using for experimentation
- This data helps to varying degrees
  - If it’s labeled, can make a big difference without requiring extra work (ex. confusion sets)
  - If it’s available and some annotation can occur, active learning can help
  - If it’s available but no extra work is possible, benefit can still be found (ex. bootstrapping)
- Authors suggest moving “towards increasing the size of annotated training collections”