Bootstrapping Parallel Corpora

Chris Callison-Burch
and Miles Osborne

May 31, 2003
The Game Plan

• Introduction to weakly supervised learning: Co-Training

• Parallel Corpora: How much data for MT? How much will it cost?

• Fitting translation into Co-Training framework

• Results, and even better results
Weakly Supervised Learning

• Goal: Reduce cost of creating corpora by (semi-)automating the process

• Methodology: Train learners on labeled data, use them to label more data, and retrain.

• Types:
  – Self-Training
  – Co-Training
  – Active Learning
Co-Training Algorithm

$L = \text{small set of annotated data,}$
$U = \text{a larger set of unannotated data,}$
$V_1, V_2 = \text{two views of the data.}$

\[
\begin{align*}
\text{repeat until done} \\
\text{train classifier } h_1 \text{ on view } V_1 \text{ of } L \\
\text{train classifier } h_2 \text{ on view } V_2 \text{ of } L \\
\text{allow } h_1 \text{ to posit labels for examples in } U \\
\text{allow } h_1 \text{ to posit labels for examples in } U \\
\text{add } h_1 \text{'s most confidently labeled examples to } L \\
\text{add } h_2 \text{'s most confidently labeled examples to } L
\end{align*}
\]
Preconditions for co-training

- Must have multiple views on the data being labeled

- Each view must be
  1. sufficient for the task
  2. independent of each other view
Machine translation is supervised learning

- Parallel corpora are annotated training data
- Source sentence is “labeled” with translation
- Word-level alignments are estimated from parallel corpora
- Increasing data leads to improved accuracy...
Training corpus size v accuracy

![Graph showing the relationship between training corpus size and accuracy.](image)
Example system

- IBM’s “Candide” system for translating between French and English
- Trained on ten years’ worth of the proceedings of Canadian Parliament
- ... 3,000,000 parallel sentences
- Such large resources aren’t generally available!
- So using statistical MT for a lot of languages won’t work.
Cost for a new language pair

• Germann (2001) estimates cost for new language pair

• Creates Tamil-English parallel corpus from scratch

• Cost for professional translation = $0.36 per Tamil word

• Cost for corpus size of Canadian Hansard =
  2,870,000 sentences * 20 words/sent * 0.36 $/word = $20 million.
Data cost is a problem

- Need to reduce cost of creating corpus
- Better exploit small resources
- Perfect for weakly supervised learning!
Fitting translation into co-training framework

• In parallel corpora source sentences are “labeled” with their translations

• Source sentence is a “view” onto that translation

• Existing translations of the source sentence can be used as addition views on new translations
# Views and Labels

<table>
<thead>
<tr>
<th>View 1</th>
<th>View 2</th>
<th>Labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gewässer der Europäischen Gemeinschaft</td>
<td>Eaux de la Communauté européenne</td>
<td>European Community waters</td>
</tr>
<tr>
<td>Die Ausgaben je Schüler</td>
<td>Les dépenses par élève</td>
<td>Expenditure per pupil</td>
</tr>
<tr>
<td>Vorläufige Zahlen</td>
<td>Données provisoires</td>
<td>Provisional figures</td>
</tr>
<tr>
<td>die Klage wird als offensichtlich unzulässig abgewiesen</td>
<td>Le recours est rejeté comme manifestement irrecevable</td>
<td>The action is dismissed as manifestly inadmissible</td>
</tr>
<tr>
<td>die Französische Republik trägt ihre eigenen Kosten</td>
<td>La République française supportera ses propres dépens</td>
<td>France was ordered to bear its own costs</td>
</tr>
<tr>
<td>Binnenproduktion, ausgedrückt in % der Binnenverwendung</td>
<td>Production domestique exprimée en pourcentage de l'utilisation domestique</td>
<td>Domestic output as a % of domestic use</td>
</tr>
<tr>
<td>Nur Industrie</td>
<td>Seulement l'industrie</td>
<td>Industry only</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Co-training for statistical machine translation

1. **French**
   - maison bleue
   - bleue Haus
   - casa azul

2. **German**
   - Maison bleu
   - blaues Haus
   - Casa azul
   - English target
   - Blue maison
   - blaues House
   - Blue house

3. **Spanish**
   - Blue maison
   - blaues Haus
   - Blue house

4. **English**
   - Blue house

Callison-Burch and Osborne

Bootstrapping Parallel Corpora

May 31, 2003
Experimental setup

- Labeled data: Parallel corpora with 20,000 sentences
- Views: Translation models of five languages into English
- Pool of data w/o English labels: Parallel corpus 60,000 sentences in the five languages
- Each re-training round adds 10,000 machine translated sentences, and removes them from the pool.
Results of Co-training

![Bar chart showing results of Co-training for different languages and rounds.]

- **FR**: Round 0 and Round 1 show similar performance, with Round 0 slightly higher.
- **ES**: Round 0 and Round 1 perform better than Round 2 and Round 3.
- **DE**: Round 2 and Round 3 show lower performance compared to Round 0 and Round 1.
- **IT**: Round 0 and Round 1 perform better than Round 2 and Round 3.
- **PT**: Round 0 and Round 1 show similar performance, with Round 0 slightly higher.
Variant Experiment

- Previous co-training experiment used uniformly sized seed corpora

- That’s unrealistic; varying amount of training data depending on languages

- Simulate for a language pair for which no data is available

- Use a number of translation models to create a corpus

- This is a special case of co-training
Results of “coaching”

[Graph showing the relationship between training corpus size and accuracy.]
Discussion and Future Work

• Co-training gives positive gains with equal views
  – Different parameters may yield better results

• Co-training gives dramatic gains with unequal views
  – Useful for porting to new languages

• Apply to new languages of countries joining the EU

• Adapt method so that it can be used for Active Learning
Thanks!
Selection methods

- For first co-training experiment used oracle
- For second, just used everything since views are unequal

Number of possibilities:
  - Maximize performance on a held out set
  - Probability as measure of confidence
  - Number of unseen words
  - Sentence length
  - External measure such as language model