Using Large Monolingual and Bilingual Corpora to Improve Coordination Disambiguation

Shane Bergsma
Department of Computer Science and HLTCOE, Johns Hopkins University

David Yarowsky

Kenneth Church

Resolving Coordination Ambiguity

Problem: Which words are being linked by a conjunction?
Specifically: Coordination in complex Noun Phrases (NPs)

[ dairy and meat ] production asbestos and [ polyvinyl chloride ]

Interpretation: yes: [ dairy production ] and [ meat production ]
no: [ asbestos chloride ] and [ polyvinyl chloride ]

Hard problem! Treebank-trained parsers fail on it
Why? Could help Information Retrieval
Could also help syntactic Machine Translation

Like PP-attachment, etc., the specific lexical items matter
• Both sequences above have part-of-speech tags: "NN and NN NN"
• We need to look beyond labeled data for the key information
e.g. paraphrase (on the web) diagnostic (Nakov & Hearst, 2005)

Size Matters for Resolving Coordination

WSJ portion of Penn Treebank
• 1 MILLION words (labeled data)

Marcus et al. (1993)
We use: Vadas & Curran (2007) annotations, giving syntax of all NPs in WSJ portion of Penn Treebank

Collison-Burch et al. (2010): WMT 2010
Koehn (2005): Europarl

Bitexts
• 1 BILLION words (bilingual data)

We use: English-to-[Czech, Danish, German, Greek, Spanish, Finnish, French, Italian, Dutch, Portuguese, Swedish] bilingual data

Web text (N-grams)
• 1 TRILLION words (unlabeled data)

We use Google N-grams V2 (Lin et al., 2010)

Better Together: Mono and Bilingual Data

<table>
<thead>
<tr>
<th>Phrase</th>
<th>Corpus Evidence (English and Foreign)</th>
<th>Pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>dairy</td>
<td>English: production of dairy and meat</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>English: dairy production and meat production</td>
<td>1 2</td>
</tr>
<tr>
<td>2</td>
<td>Spanish: meat and dairy production</td>
<td>1 3</td>
</tr>
<tr>
<td>3</td>
<td>Finnish: maaidon- ja lihanuntuantoon</td>
<td>1 - 3</td>
</tr>
<tr>
<td>asbestos</td>
<td>English: polyvinyl chloride and asbestos</td>
<td>2 5</td>
</tr>
<tr>
<td>1</td>
<td>English: asbestos and polyvinyl chloride</td>
<td>1 3</td>
</tr>
<tr>
<td>2</td>
<td>English: asbestos and chloride</td>
<td>1 3</td>
</tr>
<tr>
<td>3</td>
<td>Portuguese: o amianto e clorote de polivinilo</td>
<td>1 - 3</td>
</tr>
<tr>
<td>4</td>
<td>Italian: F’asbesto e il polivinilchlore</td>
<td>1 - 3</td>
</tr>
</tbody>
</table>

Exploiting Unlabeled Text

Given some labeled examples of the two types of coordination:

- Create features for counts of patterns in:
  a) Bitexts (Foreign Patterns)
  b) Web text (English Patterns)

- Also create binary features for words/tags
- Train a logistic regression classifier to classify the coordination type
- Iteratively co-train using the Monoclassifier to label new examples for the bilingual classifier, and vice versa

Results on Europarl Examples

Error rate (%) of Co-trained Classifiers on Europarl test data

Results on WSJ Examples

Error rate (%) on WSJ Corpus

Co-trained classifier using only monolingual features advances the state-of-the-art on WSJ data

Check out our new annotated data and evaluation scripts at:
http://www.clsp.jhu.edu/sbergsma/coordNP_ACL11.zip