Breaking the Resource Bottleneck for Multilingual Processing

University of Edinburgh
IGK Summer School
September 6, 2004

Rebecca Hwa
University of Pittsburgh
hwa@cs.pitt.edu
Supervised Learning

• Training examples are pairs of problems and answers

• For part-of-speech (POS) tagging: word in context, POS tag pairs

• For parsing: sentence, parse tree pairs

• For text categorization: article, category pairs
Manual Annotation Issues

• Guideline development
• Laborious, time-consuming
• Costly
• Inconsistencies
• Inter-annotator (dis)agreements

Few widely-accepted annotated resources
• Data are typically news stories in English
# Multilingual Processing

<table>
<thead>
<tr>
<th>Language</th>
<th>Treebank</th>
<th>Total Dev. Time</th>
<th>Corpus Size</th>
<th>Parser Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>Penn Treebank</td>
<td>5 years</td>
<td>1M words 40k sentences</td>
<td>90%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chinese</td>
<td>Chinese Treebank v2</td>
<td>2 years</td>
<td>100K words 4k sentences</td>
<td>75%</td>
</tr>
<tr>
<td></td>
<td>Chinese Treebank v4</td>
<td>4 years</td>
<td>400K words 15k sentences</td>
<td>~80%</td>
</tr>
<tr>
<td>Others</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(e.g., Hindi, Farsi)</td>
<td>?</td>
<td></td>
<td>?</td>
<td>?</td>
</tr>
</tbody>
</table>
Research Questions

• How can annotated resources for non-English languages be acquired efficiently?
  – How to make use of resources we already have?

• How good are the resulting resources at training non-English systems?
  – Are the trained systems directly or indirectly useful?
Roadmap

• Motivation

• Annotation Projection
  – Overview
  – Theoretical Challenges
  – Practical Challenges

• Empirical studies

• Future Work
Projection Framework

bilingual corpus

English Target lang.

Analyze the English sentences
Align the words in the sentences

Projected training data in target lang.

Trained system in target lang.

Train

Analyses of sentences

New sentences In target lan.

Projection
Transformation
Filtering
Possible Applications of Projected Resources

- Morphological analyzer
- Base noun phrase chunker
- Name entity tagger
- POS tagger
- Syntactic dependency parser
- Semantic parser

...
Step 0. Acquire sentence aligned parallel text

The Chinese side expressed satisfaction regarding this subject

中国方面对此表示满意
Step 1. Align the Words

The Chinese side expressed satisfaction regarding this subject.

中国 方面 对 此 表示 满意
Step 2. Analyze the English Data

Example: POS Tagging

The Chinese side expressed satisfaction regarding this subject
Step 2. Analyze the English Data

Example: Parsing

The Chinese side expressed satisfaction regarding this subject.
Step 3: Project Annotations across Alignments

Example: Parsing

The Chinese side expressed satisfaction regarding this subject.

中国方面对此表示满意
Step 3: Project Annotations across Alignments

Example: Parsing

The Chinese side expressed satisfaction regarding this subject.
Challenges

• Theoretical challenge
  – Divergences: different languages express the same thing differently
  – Word alignments are not one-to-one

• Practical challenge
  – Framework relies on several components
  – Individual component errors can propagate
Unaligned English

regarding this subject

对 此
regarding this subject

對此

*e*
 regard, this subject
Many-to-1

regarding this subject

对 此
Many-to-1
The Chinese expressed.

中国方面表示
1-to-Many

The Chinese expressed *e*.

中国 方面 表示
The Chinese方面表示

1-to-Many

The DT
Chinese PN
expressed VBD

中国 PN
方面 PN
表示 VBD

mod

subj
Unaligned Chinese

The Chinese expressed

中国 方面 表示
The Chinese expressed
Addressing Word Alignment Mismatches

• Use known linguistic facts about the target language to transform the projected annotations
• For one-to-many and many-to-one cases, select annotation based on grammatical categories
  – In Chinese, the head of a noun phrase is the last word
• Can incorporate some unaligned words back into the projected annotations
  – Functional words (e.g., aspectual, measure words)
  – Easily enumerable lexical categories (e.g., $, RMB, yen)
Challenges: Component Errors

• English parser not perfect
  – Optimistically, 90% accurate (news text)
  – But only 86% on mixed genres like Brown
  – Still lower for imperfect English (e.g., blogs)

• Word alignments not perfect
  – For languages similar to English, low alignment error rate (10-15%)
  – For Chinese, alignment error rate is 40-50%
Improving Robustness against Component Errors

• **Filter** data that have been badly processed (e.g., suspected poor word alignments)
  – Too many words not aligned
  – Too many (non-consecutive) words of one lang. map to one word in the other lang.

• **Aggressive re-weighting** [Yarowsky & Ngai, 2001]
  – Re-normalize probability distributions to remove annotations with low likelihoods
Roadmap

• Motivation
• Projected Annotation
• Empirical studies
  – Complexity of the learning tasks
    • Training a POS tagger and dependency parser
  – Similarities of the language pairs
    • English-French/Spanish
    • English-Chinese
  – Evaluation methods
• Future Work
Evaluation Methods

- **Direct evaluation**: how accurate are the projected resources?
  - Compare projected annotations against manually created gold standard directly

- **Indirect evaluation**: how accurate are the trained systems?
  - Use the projected resources to train a new system, then test it on new sentences and compare output against manually created gold standard

- Relationship between **performance degradation** and **component errors**

- Do **post projection transformation and filtering and re-weighting** help?
Projection Framework

bilingual corpus

English

Target lang.

Analyze the English sentences

Align the words in the sentences

Projected training data in target lang.

Train

Projected training data

Trained system in target lang.

Analyses of sentences

New sentences in target lang.

Direct Evaluation

Indirect Evaluation

Projection

Transformation

Filtering
Projected Tagging

• English-French
  – As reported by Yarowsky and Ngai (HLT-2001)

• English-Chinese
  – Data: 240K parallel sentences from FBIS
  – English Tagger: Ratnaparkhi’s MaxEnt tagger
  – Alignment: IBM MT model 4 (GIZA++)
  – Test data: 1000 sentences from the ChTB (v4)
  – Chinese Tagger: Trigram Model
Direct Evaluation: Accuracy of Projected POS Tags

<table>
<thead>
<tr>
<th></th>
<th>English-French*</th>
<th>English-Chinese</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manually word aligned</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Projection only</td>
<td>85%</td>
<td>63%</td>
</tr>
<tr>
<td>+ Transformation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Automatically word aligned</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Projection only</td>
<td>76%</td>
<td>50%</td>
</tr>
<tr>
<td>+ Transformation</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* English-French experiments are taken from Yarowsky and Ngai, *HLT 2001*
## Indirect Evaluation: Accuracy of a Trained POS Tagger

<table>
<thead>
<tr>
<th></th>
<th>English-French*</th>
<th>English-Chinese</th>
</tr>
</thead>
<tbody>
<tr>
<td>POS Tagger with supervised training</td>
<td>97%</td>
<td>93%</td>
</tr>
<tr>
<td>Baseline</td>
<td>53%</td>
<td>53%</td>
</tr>
<tr>
<td>Projection only</td>
<td>86%</td>
<td>48%</td>
</tr>
<tr>
<td>+ Transformation</td>
<td>96%</td>
<td>64%</td>
</tr>
<tr>
<td>+ Filtering and Re-weighting</td>
<td></td>
<td>71%</td>
</tr>
</tbody>
</table>

* English-French experiments are taken from Yarowsky and Ngai, *HLT 2001*
Projected Tagging Findings

• More difficult to bootstrap a POS Tagger for Chinese than for French
  – English and French share similar tag sets
  – Fewer categorical divergences between English and French
  – Chinese has higher average number of tags per word than French
    • Re-weighting does not help as much for Chinese

• How can we close the gap?
Projected Parsing

• English Parser: Collins converted to dependency
• Alignment: IBM MT model 4 (GIZA++)
• English-Spanish
  – Parallel Data: 98K sentences from FBIS/Bible/UN
  – Test data: 200 unseen sentences from FBIS/Bible/UN
  – Upper bound parser: an off-the-shelf constraint grammar parser
• English-Chinese
  – Parallel Data: 240K sentences from FBIS
  – Test data: 2800 sentences from the ChTB (v4)
Direct Evaluation: Accuracy of Projected Parse Trees

<table>
<thead>
<tr>
<th></th>
<th>English-Spanish</th>
<th>English-Chinese</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Manually word aligned</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Projection only</td>
<td>37%</td>
<td>38%</td>
</tr>
<tr>
<td>+ Transformation</td>
<td>70%</td>
<td>67%</td>
</tr>
<tr>
<td><strong>Automatically word aligned</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Projection only</td>
<td>34%</td>
<td>26%</td>
</tr>
<tr>
<td>+ Transformation</td>
<td>66%</td>
<td>52%</td>
</tr>
</tbody>
</table>

Rebecca Hwa  
University of Pittsburgh
# Accuracy of a Trained Parser

<table>
<thead>
<tr>
<th></th>
<th>English-Spanish</th>
<th>English-Chinese</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resource-expensive parsers</td>
<td>69%</td>
<td>Chinese Treebank v4 64%**</td>
</tr>
<tr>
<td>Baseline (mod next)</td>
<td>34%</td>
<td>35%</td>
</tr>
<tr>
<td>+ Transformations</td>
<td>39%</td>
<td>44%</td>
</tr>
<tr>
<td>Projection</td>
<td></td>
<td></td>
</tr>
<tr>
<td>+ Transformation</td>
<td>67%</td>
<td></td>
</tr>
<tr>
<td>+ Filtering</td>
<td>72%</td>
<td>54%</td>
</tr>
</tbody>
</table>

Rebecca Hwa

University of Pittsburgh
Learning Curve Comparisons

- Human Annotated Treebank
- Projected Treebank
- Baseline 2 (mod-next + rules)
- Baseline 1 (mod-next)
Conclusion

• Use projection to acquire annotated resources for Chinese by bootstrapping from English resources

• The projected resources have an accuracy rate of nearly 70% in principle for both tagging and parsing.

• Reducing noise caused by word-alignment errors is still a major challenge.

• Systems trained on the induced resources outperform similarly inexpensive options.
Future Directions

• Reduce error rates of the word-alignment models
• Improve the projection algorithm to address more language divergences
• Develop more sophisticated techniques to filter out errors from the induced resources
# Acknowledgements

<table>
<thead>
<tr>
<th>Univ. of Maryland</th>
<th>Univ. of Pittsburgh</th>
</tr>
</thead>
<tbody>
<tr>
<td>Philip Resnik</td>
<td>Chenhai Xi</td>
</tr>
<tr>
<td>Amy Weinberg</td>
<td>Karina Ivanetich</td>
</tr>
<tr>
<td>Clara Cabezas</td>
<td>Behrang Mohit</td>
</tr>
<tr>
<td>Okan Kolak</td>
<td>Carol Nichols</td>
</tr>
<tr>
<td>Adam Lopez</td>
<td></td>
</tr>
</tbody>
</table>
Reserve slides
## English-French (Y&N table)

<table>
<thead>
<tr>
<th>Model</th>
<th>Evaluate on E-F Aligned French</th>
<th>Evaluate on Unseen Monolingual French</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Core Tagset</td>
<td>Eng Eqv Tagset</td>
</tr>
<tr>
<td>(a) Direct transfer (auto-aligned, auto-project)</td>
<td>.76</td>
<td>.69</td>
</tr>
<tr>
<td>(b) Direct transfer (hand-aligned, auto-project)</td>
<td>.85</td>
<td>.78</td>
</tr>
<tr>
<td>(c) Standard bigram model (auto-aligned, auto-project)</td>
<td>.86</td>
<td>.82</td>
</tr>
<tr>
<td>(d) Noise-robust bigram induction (auto-aligned, auto-project)</td>
<td>.96</td>
<td>.93</td>
</tr>
<tr>
<td>(e) Standard bigram model (trained on heldout goldstandard)</td>
<td>.97</td>
<td>.96</td>
</tr>
</tbody>
</table>

Table 4: Evaluation of 5 POS tagger induction models on 2 French datasets and 2 tagsets
Next Step: Filtering:

Explore what training data leads to better accuracy

<table>
<thead>
<tr>
<th></th>
<th>English, no Chinese in each sentence</th>
<th>Chinese, no English in each sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 To None:</td>
<td>20.1%</td>
<td>30.9%</td>
</tr>
<tr>
<td>Chinese, no English</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(97% of OTHER)</td>
<td></td>
</tr>
<tr>
<td>1 To None:</td>
<td>14.4%</td>
<td>34.6%</td>
</tr>
<tr>
<td>English, no Chinese</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 Chinese to Many English</td>
<td></td>
<td>Why Filtering May Improve Accuracy:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Percentages of Correspondences over All Words</td>
</tr>
<tr>
<td>1 Chinese to 1 English</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Initial Accuracy Results

<table>
<thead>
<tr>
<th></th>
<th>Test on ChTB (1165)</th>
<th>Test on ChTB Core (1165)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train on ChTB</td>
<td>92.7%</td>
<td></td>
</tr>
<tr>
<td>(14,000)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Train on ChTB Core</td>
<td></td>
<td>92.9%</td>
</tr>
<tr>
<td>(14,000)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Train on FBIS Core</td>
<td></td>
<td>48.2%</td>
</tr>
<tr>
<td>(240,000)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Accuracy after Filtering

<table>
<thead>
<tr>
<th>E no C</th>
<th>C no E</th>
<th>.4</th>
<th>.3</th>
<th>.2</th>
<th>.1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>48.0</td>
<td>49.6</td>
<td>51.5</td>
<td>53.9</td>
</tr>
<tr>
<td>.4</td>
<td>183,714</td>
<td>139,722</td>
<td>76,605</td>
<td>17768</td>
<td></td>
</tr>
<tr>
<td>.3</td>
<td>48.0</td>
<td>49.4</td>
<td>51.06</td>
<td>53.6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>162,831</td>
<td>122,302</td>
<td>64,310</td>
<td>13265</td>
<td></td>
</tr>
<tr>
<td>.2</td>
<td>47.1</td>
<td>48.7</td>
<td>51.1</td>
<td>52.8</td>
<td></td>
</tr>
<tr>
<td></td>
<td>81,981</td>
<td>55,672</td>
<td>24,952</td>
<td>4,312</td>
<td></td>
</tr>
<tr>
<td>.1</td>
<td>45.2</td>
<td>46.4</td>
<td>48.7</td>
<td>46.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>7417</td>
<td>4683</td>
<td>2,180</td>
<td>729</td>
<td></td>
</tr>
</tbody>
</table>

- **blue** = % accuracy
- **green** = number of sentences
## Accuracy after Filtering

<table>
<thead>
<tr>
<th>English -no-Chinese</th>
<th>.4</th>
<th>.3</th>
<th>.2</th>
<th>.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>.4</td>
<td>64.0 183,714</td>
<td>66.1 139,722</td>
<td>67.7 76,605</td>
<td>69.9 17,768</td>
</tr>
<tr>
<td>.3</td>
<td>64.0 162,831</td>
<td>66.2 122,302</td>
<td>67.7 64,310</td>
<td>70.0 13,265</td>
</tr>
<tr>
<td>.2</td>
<td>63.5 81,981</td>
<td>65.6 55,672</td>
<td>67.7 24,952</td>
<td>70.6 4,312</td>
</tr>
<tr>
<td>.1</td>
<td>63.1 7417</td>
<td>64.4 4683</td>
<td>66.7 2,180</td>
<td>65.7 729</td>
</tr>
</tbody>
</table>
Next Step: Will Re-Weighting POS tags Improve Accuracy?

Average Number of Tags $n$ by Word Frequency

<table>
<thead>
<tr>
<th></th>
<th>$n \geq 100$</th>
<th>$5 &lt; n &lt; 100$</th>
<th>$n \leq 5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>3.8</td>
<td>6.7</td>
<td>4.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2.0</td>
</tr>
</tbody>
</table>

% of Occurrence of Most Frequent Tags by Word Frequency

<table>
<thead>
<tr>
<th></th>
<th>$n \geq 100$</th>
<th>$5 &lt; n &lt; 100$</th>
<th>$n \leq 5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>59%</td>
<td>56%</td>
<td>55%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>66%</td>
</tr>
</tbody>
</table>