Leveraging transliterations from multiple languages

Aditya Bhargava, Bradley Hauer, and Grzegorz Kondrak
University of Alberta
NEWS 2011
Introduction

- Most previous work on transliteration has focused on a single language
  - English to Hindi, English to Japanese, Arabic to English, etc.
- But data from other languages can be helpful
- Improve existing model's results using supplemental data
Introduction

• Also experiment with:
  • Incorporating other models
  • Hindi romanization
  • English-to-Chinese alignment lengths
Previous work

- Previous NEWS
  - DirecTL/DirecTL+ (Jiampojamarn et al., 2009/2010)
    - Discriminative, online, max-margin
  - Sequitur + SMT combination (Finch and Sumita, 2010)
    - Sequitur is generative, joint n-gram
- Applying supplemental transliterations to G2P (Bhargava and Kondrak, 2011)
  - We apply this method verbatim
  - Based on SVM re-ranking
# Test data overlap

<table>
<thead>
<tr>
<th>Language</th>
<th>Test set size</th>
<th>Test set overlap</th>
</tr>
</thead>
<tbody>
<tr>
<td>EnBa</td>
<td>1,000</td>
<td>498</td>
</tr>
<tr>
<td>EnCh</td>
<td>2,000</td>
<td>2,000</td>
</tr>
<tr>
<td>EnHe</td>
<td>1,000</td>
<td>525</td>
</tr>
<tr>
<td>EnHi</td>
<td>1,000</td>
<td>889</td>
</tr>
<tr>
<td>EnJa</td>
<td>1,815</td>
<td>734</td>
</tr>
<tr>
<td>EnKa</td>
<td>1,000</td>
<td>883</td>
</tr>
<tr>
<td>EnKo</td>
<td>609</td>
<td>608</td>
</tr>
<tr>
<td>EnPe</td>
<td>2,000</td>
<td>1,049</td>
</tr>
<tr>
<td>EnTa</td>
<td>1,000</td>
<td>884</td>
</tr>
<tr>
<td>EnTh</td>
<td>2,000</td>
<td>1,564</td>
</tr>
</tbody>
</table>
Re-ranking

DOS

ドス

ДОС

Доос

דוס

ตอส

...?
Re-ranking

- SVM re-ranking using all other languages
- Features:
  - N-gram features based on character alignments
  - Similarity features based on alignment scores
- Transliteration data are noisy; handled by:
  - Granular n-gram features
  - Multiple languages
- DirecTL+ baseline
EnHi transliteration re-ranking

Word accuracy

- DirecTL+
- DirecTL+ w/ supp. TLs
- Best other
EnHi transliteration re-ranking

Word accuracy

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>DirecTL+</td>
<td>45</td>
</tr>
<tr>
<td>DirecTL+ w/ supp. TLs</td>
<td>50</td>
</tr>
<tr>
<td>Best other</td>
<td></td>
</tr>
</tbody>
</table>
EnHi transliteration re-ranking

![Bar chart showing word accuracy for different methods: DirectTL+, DirectTL+ with support TLs, and Best other.](Image)
Re-ranking with Sequitur

- Use Sequitur's output for re-ranking
- Exact same features
EnHi Sequitur re-ranking

Word accuracy

DirectTL+   +Sequitur   +TLs   Best other
EnHi Sequitur re-ranking

<table>
<thead>
<tr>
<th>Method</th>
<th>Word Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>DirectTL+</td>
<td>45</td>
</tr>
<tr>
<td>+Sequitur</td>
<td>50</td>
</tr>
<tr>
<td>+TLs</td>
<td>55</td>
</tr>
<tr>
<td>Best other</td>
<td>45</td>
</tr>
</tbody>
</table>
Hindi romanization

- Devanagari alphabet has combined consonants & vowels
- We experiment with romanizing Hindi
  - Gives DirecTL+ direct individual control
  - Context-sensitive rule-based romanization
  - Use romanized Hindi for training DirecTL+, do testing, then convert outputs to Devanagari
Chinese alignment length

- DirecTL+ relies on many-to-many alignments (M2M-Aligner)
- We experiment with maximum alignment length
  - 3-1 vs. 7-1
EnCh alignment length

- DirecTL+ 3-1
- DirecTL+ 7-1

Word accuracy

0 5 10 15 20 25 30 35 40

DirecTL+ 3-1 DirecTL+ 7-1
Conclusion

- SVM re-ranking for transliteration
  - Great improvements with supplemental transliterations
  - Also see improvements for system combination
- Romanization
  - Didn't work for EnHi (unlike EnJa in 2010)
- EnCh alignment lengths
  - Must be careful to choose a good value!