Towards Cross-language WSD for Quechua

Alex Rudnick (alexr@cs.indiana.edu)
School of Informatics and Computing, Indiana University

Contributions
- Can do cross-language WSD for Spanish-Quechua adjectives. Slightly beat a high baseline.
- Cataloging some available Spanish-Quechua tools and resources
- Reusable open source software

About Cross-Language WSD
- Classification task: predict correct translation using source language text
- “word senses” constrained by distinctions in target language
- Useful for translation, increasingly used in MT

Quechua
- Some 10 million speakers in the Andes mountain range
- Under-resourced, no available high-quality MT
- Morphologically rich, except for adjectives
- HLTDI is developing morphological analyzer and MT system

Wordnet for Spanish
- Free version is fairly sparse
- Reusable software to query it from Python

KNN with Wordnet Similarity
- Find words in the local context with similar wordnet entries.
- Count hops in the hypernymy graph.

Cross-Language Simplified Lesk
- Look up Quechua dictionary entries for target adjectives
- Count matching Spanish words in source sentence

General-purpose Classifiers
- Word context features
- Dependency parses from FreeLing help slightly
- Decision trees, Naïve Bayes, KNN
- Can only beat baseline by disagreeing with it

Results

<table>
<thead>
<tr>
<th>classifier</th>
<th>features</th>
<th>accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFS baseline</td>
<td>training instances</td>
<td>76.1</td>
</tr>
<tr>
<td></td>
<td>corpus</td>
<td>69.1</td>
</tr>
<tr>
<td></td>
<td>other stories</td>
<td>61.7</td>
</tr>
<tr>
<td></td>
<td>uniform guess</td>
<td>38.9</td>
</tr>
<tr>
<td>Simplified Lesk</td>
<td>words</td>
<td>65.9</td>
</tr>
<tr>
<td></td>
<td>decision trees</td>
<td>76.6</td>
</tr>
<tr>
<td></td>
<td>words, wn</td>
<td>76.0</td>
</tr>
<tr>
<td></td>
<td>words, parse</td>
<td>77.2</td>
</tr>
<tr>
<td>KNN</td>
<td>words</td>
<td>77.6</td>
</tr>
<tr>
<td></td>
<td>wn</td>
<td>75.3</td>
</tr>
<tr>
<td></td>
<td>words, wn</td>
<td>77.2</td>
</tr>
<tr>
<td></td>
<td>words, parse</td>
<td>77.6</td>
</tr>
</tbody>
</table>

Future Work
- Generalize to other parts of speech
- Integrate with our MT system
- Disambiguate whole phrases
- Next languages: English, Guaraní, Amharic?

Resources
- Spanish-Quechua dictionaries
- Catholic Bible
- NLTK
- FreeLing
- Stories and elicitation corpus from AVENUE