

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



Resources

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

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

classifier	features	accuracy	
MFS baseline	training instances	76.1	
	corpus	69.1	
	other stories	61.7	
	uniform guess	38.9	
Simplified Lesk		65.9	
	decision trees	words	76.6
		words, wn	76.0
	words, parse	77.2	
KNN	words	77.6	
	wn	75.3	
	words, wn	77.2	
	words, parse	77.6	

Future Work

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