

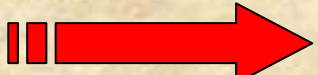
LING 573 || 30 May 2007

CHALLENGE LANGUAGE

*Cross-linguistic POS Tagger for
Hindi, Spanish, and **Icelandic***

Sabrina Burleigh
Ankit Srivastava

OUR SYSTEM

- Obtain and Clean IGT
- Tag translation lines, project onto source lines
- Pre-processing
- Generate a seed lexicon and morpheme learner from IGT
- Induce LEX (lexicon over raw text from seed lexicon)
- Construct a HMM and generate a sequence of tags for test data using Viterbi smoothing
- Post-processing
- Evaluate against Gold standard
- Experiments and Improvements 

PERFORMANCE

DESC	HND	SPN	ICE
# IGT	562	1,092	754
Training data	3.7 M words	3.2 M words	720K words
Testing data	2 K words	2 K words	22 K words
Baseline tagger	45.69 %	30.26 %	20.04 %
Tagger so far	62.53 %	47.28 %	53.89 %

ICELANDIC

DESC	VITERBI	POST
All NN	20.04 %	30.01 %
No Morph	43.35 %	53.89 %
With Morph	40.08 %	50.43 %

REMAINING

- Optimize Viterbi
- Perform other experiments
- Increase Testing and Training size
- Run EM (esthmm) and testvit
- Revise tag mappings