

LING 573 | May 16, 2007

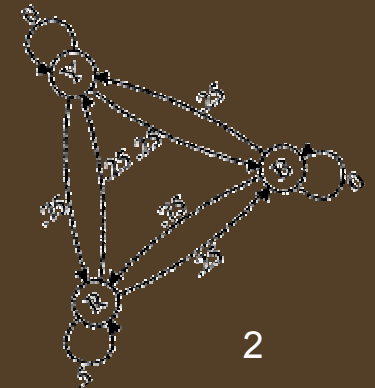
CROSS-LINGUISTIC ANNOTATOR

Lexicon and HMM Induction

Sabrina Burleigh and Ankit Srivastava

PROGRESS CHECK

- Hindi is HD and Spanish is LD
- IGT with projected tags on source line
- Seed lexicon extracted from IGT
- Lexicon induced over raw text
- Compute probabilities from frequency data
- Plug in initial model parameters in the HMM
- Run “esthmm” and “testvit”
- Evaluate results

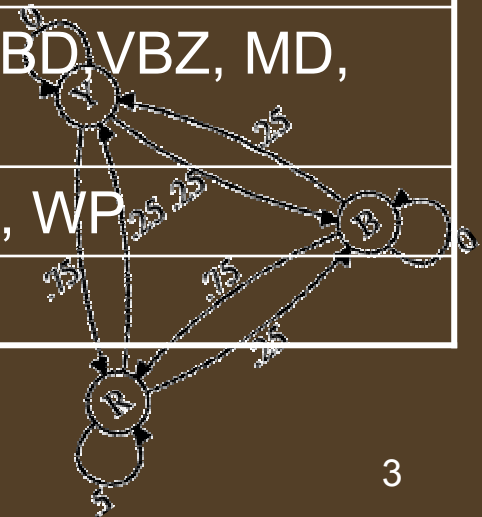


LEXICON INDUCTION

- Convert tag to the universal tag set (13 tags)
- Filter tags

PUNC	. , ? !
NN	NN, NNP, NNS, etc
SCC	NEG, UH, EX
JJ	JJ, JJR
RB	RB, RBR
IN	IN, TO, MM, PREP
CD	CD

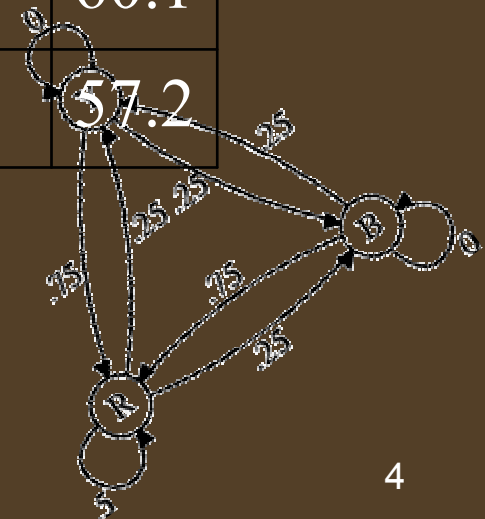
CC	REL, CC
IND	PRP, PRP\$, PRO, DEM, ..
DT	DT, PDT, DEF, ..
VB	VB, VBD, VBZ, MD, VBG
QW	WDT, WP
MISC	MM



LEXICON INDUCTION

- Words tagged in seed lexicon

	# of Words	# of Words Tagged	%
Hindi	2855	1716	60.1
Spanish	6564	3756	57.2

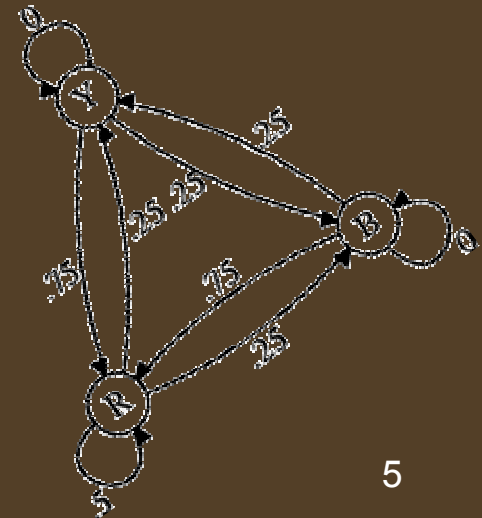


LEXICON INDUCTION

- Tokenize Raw Text

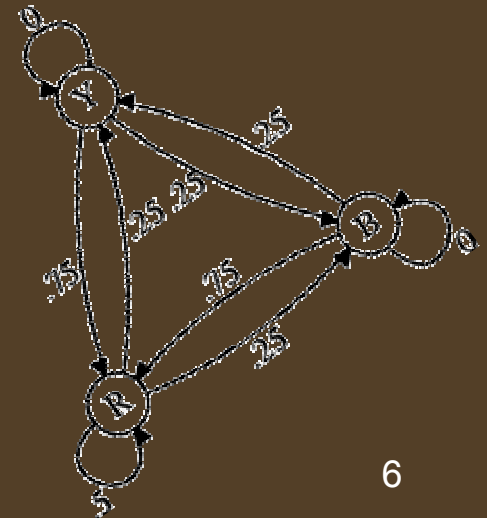
HINDI	AnnCorr:: 160K lines;
SPANISH	Spanish Newswire:: 28M lines;

90% train || 10% test



LEXICON INDUCTION

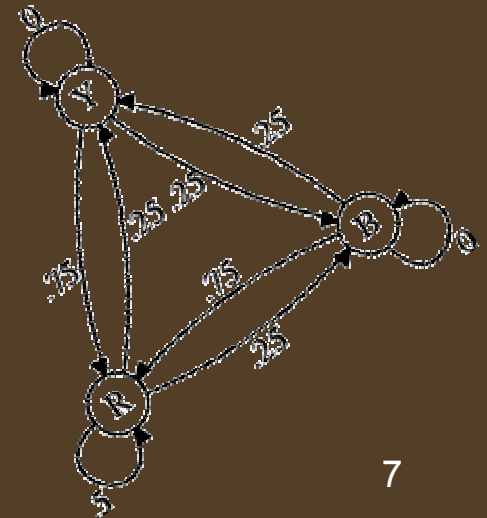
- Record frequencies of words, morphemes, tags
- Knowledge induced from morphemes (attached to base words) and suffixes (last 3 characters)



LEXICON INDUCTION

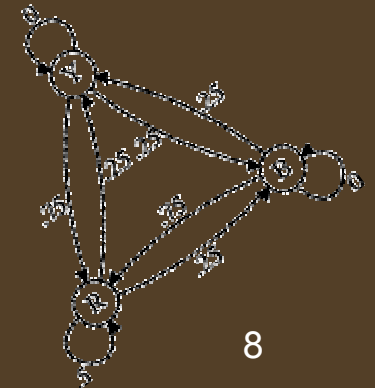
<word> <tag1>=<freq1> <tag2>=<freq2> ...

- sanskriti CC=34 CD=34 DT=34 IN=34 IND=34
 JJ=34 NN=34 PUNC=34 QW=34 RB=34 SCC=34
 VB=34
- tresres CD=4 IN=1 IND=1 VB=1



INITIAL HMM MODEL

- EMISSION: $P(\text{word} \mid \text{tag}) = \#(\text{word_tag}) / \# \text{tag}$
- TRANSITION: Equiprobable / learning transitions
- PI : Equiprobable



HMM INDUCTION

- Issues with setting the initial model
- Esthmm and testvit



ISSUES TO BE ADDRESSED

- Sum of Probabilities in HMM model < 1
- Tag to Tag transitions
- Baseline tag mapping
- Evaluation

