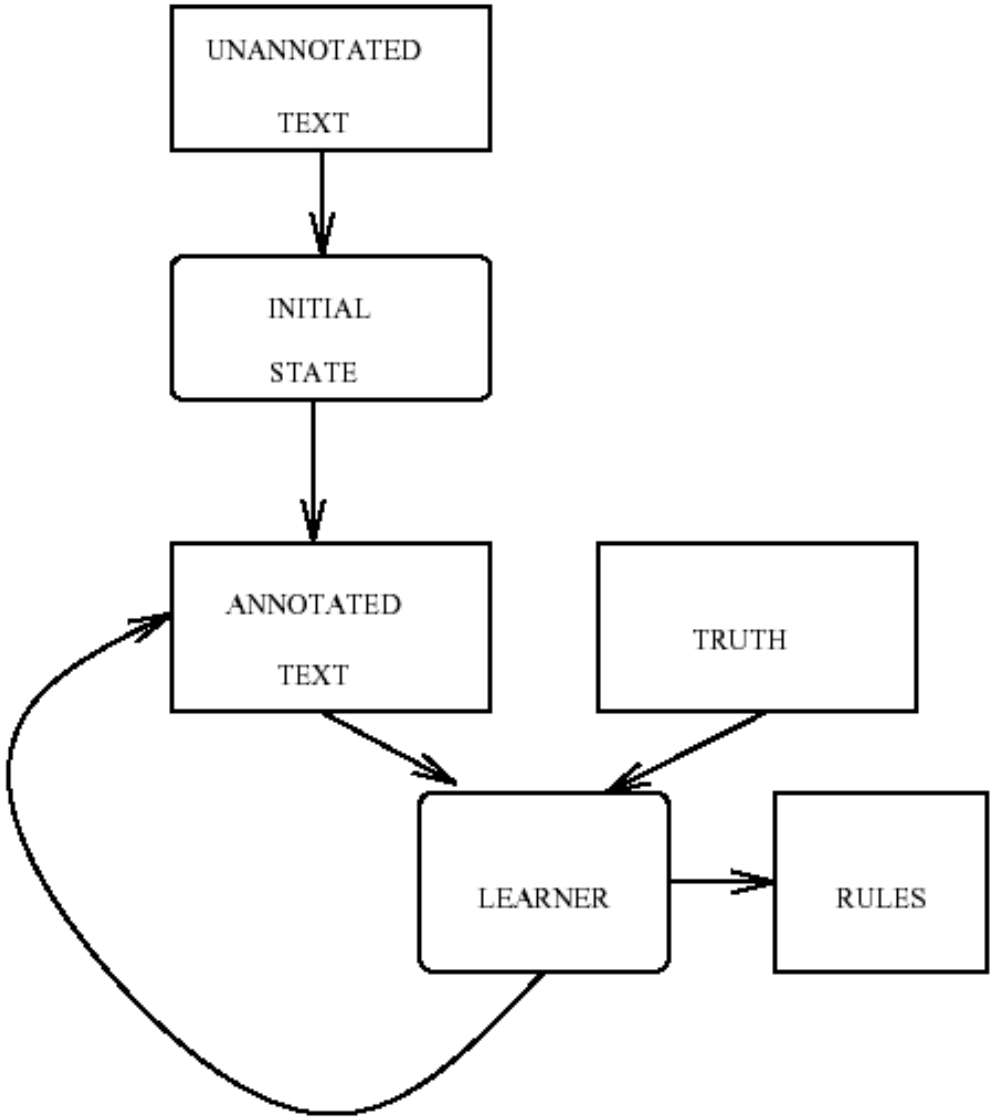
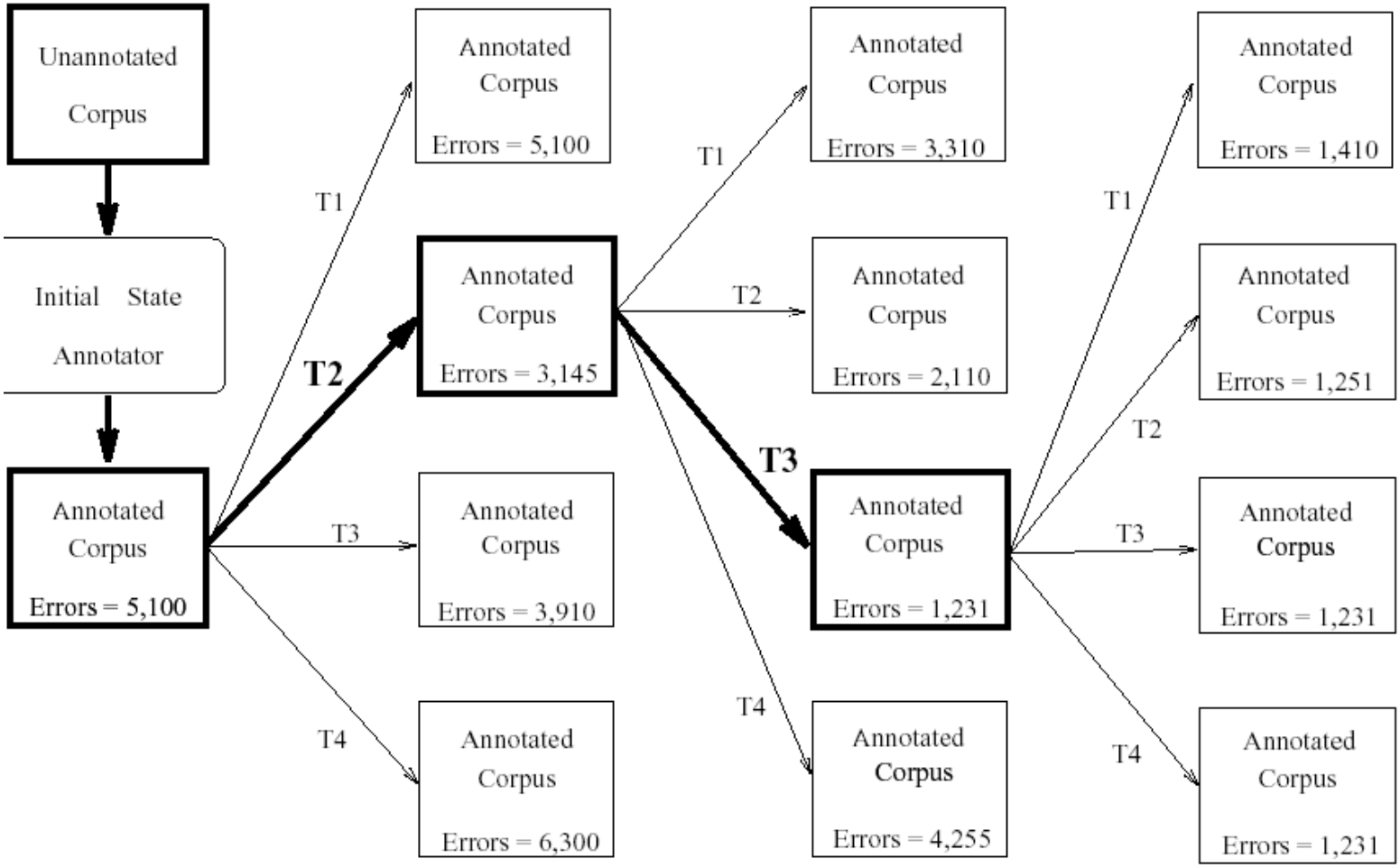
CMPT-882: Statistical Learning of Natural Language

Lecture #7

Anoop Sarkar
anoop@cs.sfu.ca
http://www.sfu.ca/~anoop

- Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging. Eric Brill. *Computational Linguistics* 1995.


UNANNOTATED
TEXT
INITIAL
STATE
ANNOTATED
TEXT
TRUTH
LEARNER
RULES


Unannotated Corpus
Initial State Annotator
Annotated Corpus Errors = 5,100
T1
T2
T3
T4
Annotated Corpus Errors = 5,100
Annotated Corpus Errors = 3,145
Annotated Corpus Errors = 3,910
Annotated Corpus Errors = 6,300
T1
T2
T3
T4
Annotated Corpus Errors = 3,310
Annotated Corpus Errors = 2,110
Annotated Corpus Errors = 1,231
Annotated Corpus Errors = 4,255
T1
T2
T3
T4
Annotated Corpus Errors = 1,410
Annotated Corpus Errors = 1,251
Annotated Corpus Errors = 1,231
Annotated Corpus Errors = 1,231

1. apply initial state annotator to corpus

2. **while** transformations can still be found **do**

3. **for** from_tag = tag_1 **to** tag_n

4. **for** to_tag = tag_1 **to** tag_n

5. **for** corpus_position = 1 **to** corpus_size

6. **if** (correct_tag(corpus_position) == to_tag

&& current_tag(corpus_position) == from_tag)

7. num_good_transformations(tag(corpus_position -1))++

8. **else if** (correct_tag(corpus_position) == from_tag

&& current_tag(corpus_position) == from_tag)

9. num_bad_transformations(tag(corpus_position-1))++

10. find max_T (num_good_transformations(T) - num_bad_transformations(T))

11. if this is the best scoring rule found yet then store as best rule:

Change tag from from_tag to to_tag if previous tag is T

12. apply best rule to training corpus

13. append best rule to ordered list of transformations

	Change Tag		
#	From	To	Condition
1	NN	VB	Previous tag is *TO*
2	VBP	VB	One of the previous three tags is *MD*
3	NN	VB	One of the previous two tags is *MD*
4	VB	NN	One of the previous two tags is *DT*
5	VBD	VBN	One of the previous three tags is *VBZ*
6	VBN	VBD	Previous tag is *PRP*
7	VBN	VBD	Previous tag is *NNP*
8	VBD	VBN	Previous tag is *VBD*
9	VBP	VB	Previous tag is *TO*
10	POS	VBZ	Previous tag is *PRP*
11	VB	VBP	Previous tag is *NNS*
12	VBD	VBN	One of previous three tags is *VBP*
13	IN	WDT	One of next two tags is *VB*
14	VBD	VBN	One of previous two tags is *VB*
15	VB	VBP	Previous tag is *PRP*
16	IN	WDT	Next tag is *VBZ*
17	IN	DT	Next tag is *NN*
18	JJ	NNP	Next tag is *NNP*
19	IN	WDT	Next tag is *VBD*
20	JJR	RBR	Next tag is *JJ*

Change tag **a** to tag **b** when:

1. The preceding (following) word is w.
2. The word two before (after) is w.
3. One of the two preceding (following) words is w.
4. The current word is w and the preceding (following) word is x.
5. The current word is w and the preceding (following) word is tagged z.
6. The current word is w.
7. The preceding (following) word is w and the preceding (following) tag is t.
8. The current word is w, the preceding (following) word is w_2 and the preceding (following) tag is t.

Method	Training Corpus Size (Words)	# of Rules or Context. Probs.	Acc. (%)
Stochastic	64 K	6,170	96.3
Stochastic	1 Million	10,000	96.7
Rule-Based With Lex. Rules	64 K	215	96.7
Rule-Based With Lex. Rules	600 K	447	97.2
Rule-Based w/o Lex. Rules	600 K	378	97.0