Tagging: An Overview

Rule-based Disambiguation

- Example after-morphology data (using Penn tagset):

I	watch	a	fly	.
NN	NN	DT	NN	.
PRP	VB	NN	VB	
	VBP		VBP	

- Rules using
  - word forms, from context & current position
  - tags, from context and current position
  - tag sets, from context and current position
  - combinations thereof

Example Rules

I	watch	a	fly
NN	NN	DT	NN
PRP	VB	NN	VB
	VBP		VBP

- If-then style:
  - $DT_{eq,-1,Tag} \Rightarrow NN$

    (implies $NN_{in,0,Set}$ as a condition)
  - $PRP_{eq,-1,Tag}$ *and* $DT_{eq,+1,Tag} \Rightarrow VBP$
  - $\{DT,NN\}_{sub,0,Set} \Rightarrow DT$
  - $\{VB,VBZ,VBP,VBD,VBG\}_{inc,+1,Tag} \Rightarrow$ *not* DT
- Regular expressions:
  - *not*(<*,*,DT> <*,*,*not*NN>))
  - *not*(<*,*,PRP>,<*,*,*not*VBP>,<*,*,DT>)
  - *not*(<*,$\{DT,NN\}_{sub,}$*not*DT>)
  - *not*(<*,*,DT>,<*,*,{VB,VBZ,VBP,VBD,VBG}>)

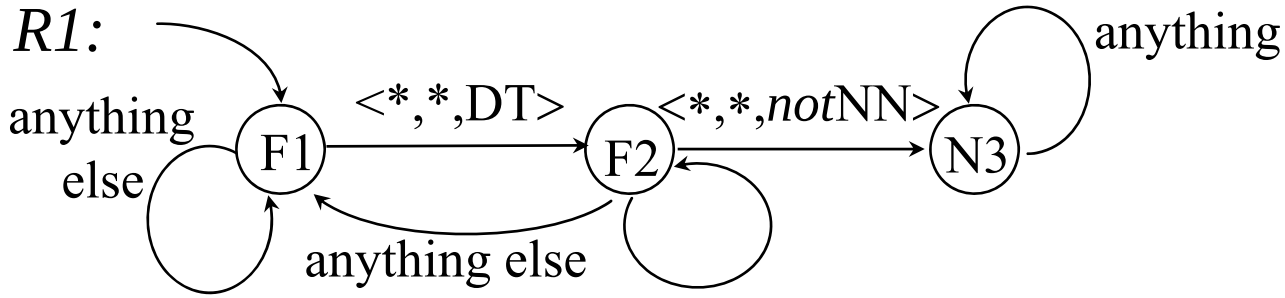
Implementation

- Finite State Automata
  - parallel (each rule ~ automaton);
    - algorithm: keep all paths which cause all automata say *yes*
  - compile into single FSA (intersection)
- Algorithm:
  - a version of Viterbi search, but:
    - no probabilities ("categorical" rules)
    - multiple input:
      - keep track of all possible paths

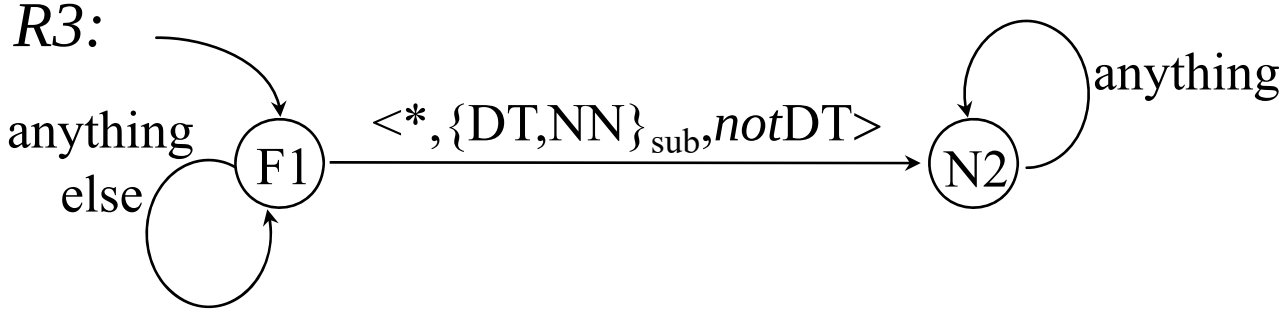
Example: the FSA

- *R1: not*(<*,*,DT> <*,*,*not*NN>))
- *R2: not*(<*,*,PRP>,<*,*,*not*VBP>,<*,*,DT>)
- *R3: not*(<*,{DT,NN}$_{sub,}$DT>)
- *R4: not*(<*,*,DT>,<*,*,{VB,VBZ,VBP,VBD,VBG}>)

- *R1:*



- *R3:*



Applying the FSA

I	watch	a	f
NN	NN	DT	N
PRP	VB	NN	V
	VBP		V

- *R1: not*(<*,*,DT> <*,*,*not*NN>))
- *R2: not*(<*,*,PRP>,<*,*,*not*VBP>,<*,*,DT>)
- *R3: not*(<*,{DT,NN}$_{sub}$,DT>)
- *R4: not*(<*,*,DT>,<*,*,{VB,VBZ,VBP,VBD,VBG}>)

- R1 blocks:

a	fly
DT	
	VB
	VBP

remains:

a	fly
DT	NN

or

a	fly
	NN
NN	VB
	VBP

- R2 blocks:

I	watch	a
	NN	DT
PRP	VB	

remains e.g.:

I	watch	a
		DT
PRP		
	VBP	

and more

- R3 blocks:

a
NN

remains only:

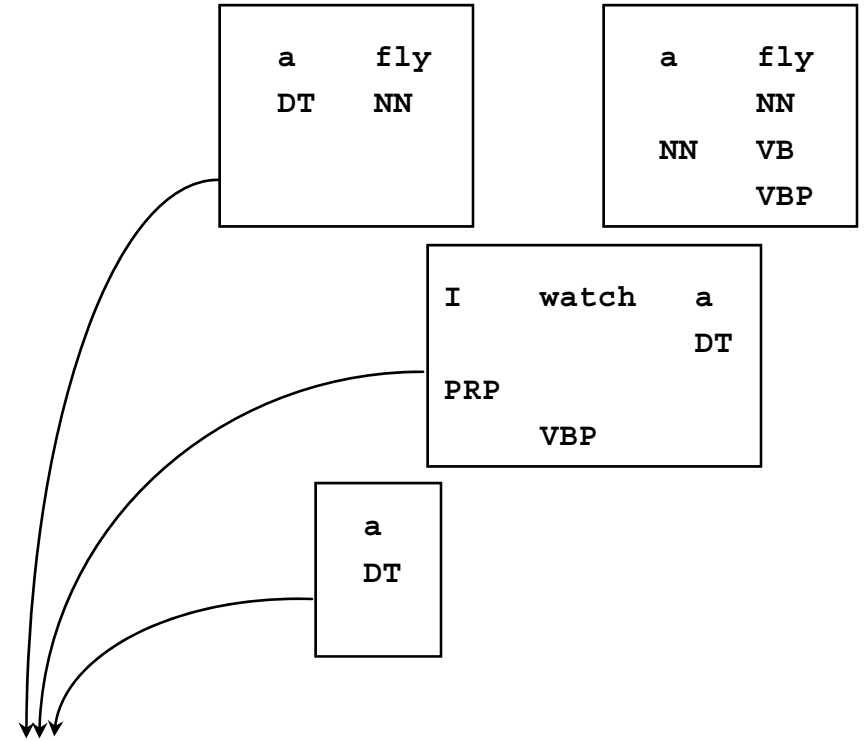
a
DT

- R4 $\subset$ R1!

Applying the FSA (Cont.)

I	watch	a	fly
NN	NN	DT	NN
PRP	VB	NN	VB
	VBP		VBP

- Combine:

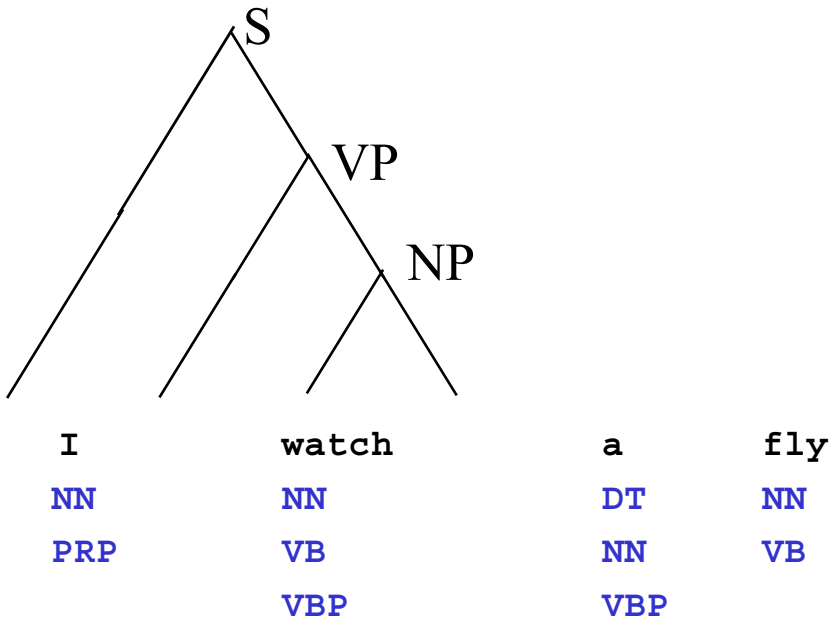


- Result:

I	watch	a	fly	.
PRP	VBP	DT	NN	.

Tagging by Parsing

- Build a parse tree from the multiple input:



- Track down rules: e.g., NP → DT NN: extract (a/DT fly/NN)
- More difficult than tagging itself; results mixed

Statistical Methods (Overview)

- "Probabilistic":
  - HMM
    - Merialdo and many more (XLT)
  - Maximum Entropy
    - DellaPietra et al., Ratnaparkhi, and others
- Rule-based:
  - TBEDL (Transformation Based, Error Driven Learning)
    - Brill's tagger
  - Example-based
    - Daelemans, Zavrel, others
- Feature-based (inflective languages)
- Classifier Combination (Brill's ideas)