Part-of-Speech (POS)
Tagging Revisited

Mark Sharp
CS-536 Machine Learning
Term Project
Fall 2003
The False Drop Problem

blind venetian
vs
venetian blind

dog bites man
vs
man bites dog
POS Tagging

blind/adj venetian/propernoun vs venetian/adj blind/noun
dog/noun bites/verb man/noun vs man/noun bites/verb dog/noun
Syntactic Parsing

blind/adj venetian/propernoun
vs
venetian/adj blind/noun

dog/noun/subj bites/verb/pred man/noun/DO
vs
man/noun/subj bites/verb/pred dog/noun/DO
Semantic Tagging

blind/adj venetian/propernoun <person>

vs

venetian/adj blind/noun <houseware>

dog/noun/subj bites/verb/pred man/noun/DO <not news>

vs

man/noun/subj bites/verb/pred dog/noun/DO <news>
Wanted: automatic POS tagger

• ~95% of words are unambiguous
• Many hand-tagged corpora for training
• Rule-based: $tag_n \leftarrow feature_1, feature_2, \ldots$
• Stochastic: $P(tag_n | tag_{n-1}, tag_{n-2}, \ldots)$
• Brill: search “rule space” for most effective “transformations” of dominant/majority tags and paint them onto tagged text dynamically.
Problem: rule space is infinite
Brill’s solution: narrow window + rule templates

...?...
Shotgun Approach

- Weka J48 (C4.5 pruned decision tree)
- Susanne corpus: 571 unique token/words with >1 POS tag out of 48257 tokens.
- Attributes:
  1. $\text{token}_n$
  2. $\text{tag}_n$ (class to predict)
  3-22. $\text{tag}_{n-10}$, $\text{tag}_{n-9}$, ..., $\text{tag}_{n+9}$, $\text{tag}_{n+10}$
  23-42. $\text{token}_{n-10}$, $\text{token}_{n-9}$, ..., $\text{token}_{n+10}$
- Result: 4.1% $\Rightarrow$ 2.2% error rate (curve C)
TBL Approach (true)

- For each of 571 unique token/words with >1 POS tag, compute and apply transformation rule in order of error impact.
- Attributes:
  1. $\text{token}_n$
  2. $\text{tag}_n$ (class to predict)
  3-22. $\text{tag}_{n-10}, \text{tag}_{n-9}, \ldots \text{tag}_{n+9}, \text{tag}_{n+10}$
  23-42. $\text{token}_{n-10}, \text{token}_{n-9}, \ldots \text{token}_{n+10}$
- Big Java project!
Shotgun Approach

........?........
TBL Approach (true)

........?........

.................
TBL Approach (simulated)

• Do a shotgun 22-attribute J48 for 544 low-impact tokens (curve A).

• For each of 17 high-impact tokens, do a single-token 42-attribute J48 and apply transformation rule in order of error impact (curve B).

• Result: 4.1% ➞ 2.1% error rate

• For single tokens, J48 rules look a lot like Brill’s templates !!!
TBL Approach (simulated)
A J48 transformation rule

if (token[n] EQUALS "to")
    then set tag[n] to "TO"

if (tag[n+1] MEMBER_OF {CD COL DOL DT JJ JJR NN NNP NNS PDT PP PPS RB VBG VBN WDT WRB})  [6 new cases]
    then change tag[n] to "IN"

if (tag[n+1] EQUALS "RB")
    if (tag[n+2] EQUALS "VB")  [3 new cases]
        then change tag[n] back to "TO"
Details, paper, data, scripts, etc.

http://scils.rutgers.edu/~msharp/CS536_ML/