POS Tagger

Gang Li
5/8 2012
• POS tagging is the fundamental part of many NLP applications, like Machine translation.

• Ambiguity: a word may have many part of speech.
• Example: book (noun, verb)

• Method: Winnow machine learning algorithm

• Corpus used: Wall Street Journal Corpus (Upeen Treebank)
Preprocessing

- Tree to flat sentence
Preprocessing

- Tree to flat sentence

regular expression to match pattern
POS Network

- The network consists of nodes for each POS

I have a book shelf.
For word "book", extract context

PRP V DT book N.
POS before book: DT, V
POS after book N
The context is given to the node
each node returns a score, highest win
Training

- 35254 sentences used for training
- Costs 3.5 hours

Map module

\[ \sum w_i > \theta \]

Positive or negative? Update weights on edges
Demo
Future work

• Adjustment of the winnow algorithm
  – this kind of algorithm has a optimized parameters

• Baseline?
  – noise is unavoidable?

• Improve data structure and algorithm implementation
  – I use "set" in some functions, it is not effecient for updating

• Unknow words
  – some other methods to deal with unknown words
Thank you!