CSC 594 Topics in AI – Applied Natural Language Processing

Fall 2009/2010

8. Shallow Parsing

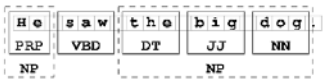
Shallow Parsing

- Break text up into non-overlapping contiguous subsets of tokens.
  - Also called chunking, partial parsing, light parsing.
- What is it useful for?
  - Named entity recognition
    - people, locations, organizations
  - Studying linguistic patterns
    - gave NP
    - gave up NP in NP
    - gave NP NP
    - gave NP to NP
  - Can ignore complex structure when not relevant

Source: Marti Hearst, i256, at UC Berkeley

Segmenting vs. Labeling

- Tokenization *segments* the text
- Tagging *labels* the text
- Shallow parsing does both simultaneously.



[$_{NP}$ He] saw [$_{NP}$ the big dog]

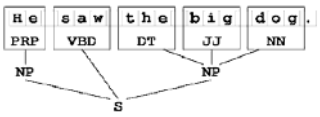
Finite-state Rule-based Chunking

- Chunking identifies basic phrases in a sentence
  - Usually NP, VP, AdjP, PP
  - Chunks are flat-structured and non-recursive → can be identified by **regular expressions**
  - Example rules: (where each rule makes a *finite state transducer*)
    - NP -> (Det) NN* NN
    - NP -> NNP
    - VP -> VB
    - VP -> Aux VB
  - Since non-recursive, **no PP-attachment** ambiguities

4

Partial Parsing by Finite-State Cascades

- By combining the finite-state transducers hierarchically (both chunks and non-chunks) as **cascades**, we can have a tree that spans the entire sentence → **partial parsing**
- Partial parsing can approximate full parsing (by CFG)
  - e.g. S -> PP* NP PP* VBD NP PP*



5

CoNLL Collection (1)

- From the CoNLL Competition from 2000
- Goal: Create machine learning methods to improve on the chunking task

[NP He] [VP reckons] [NP the current account deficit] [VP will narrow] [PP to] [NP only # 1.8 billion] [PP in] [NP September] .

- Data in IOB format:
  - Word POS-tag IOB-tag
  - Training set: 8936 sentences
  - Test set: 2012 sentences
- POS tags from the Brill tagger
  - Penn Treebank Tags

```
He          PRP  B-NP
reckons     VBZ  B-VP
the         DT   B-NP
current     JJ   I-NP
account     NN   I-NP
deficit     NN   I-NP
will        MD   B-VP
narrow      VB   I-VP
to          TO   B-PP
only        RB   B-NP
#           #    I-NP
1.8         CD   I-NP
billion     CD   I-NP
in          IN   B-PP
September   NNP  B-NP
.           .    O
```

Source: Marti Hearst, i256, at UC Berkeley

6

CoNLL Collection (2)

- ML methods used:
  - Various methods including those for sequence labeling
- Evaluation measure: F-score
  - 2*precision*recall / (recall+precision)
  - Baseline was: select the chunk tag that is most frequently associated with the POS tag, F =77.07
  - Best score in the contest was F=94.13

7

Machine Learning-based Chunking

- Supervised ML techniques to train a chunker → **sequential classification**
- Words in the data are annotated with IOB tags
  - B- -- beginning of a chunk/NE
  - I- -- internal of a chunk/NE
  - O – outside of any chunk/NE

Word	POS	Chunk	EntityType
U.N.	NNP	I-NP	I-ORG
official	NN	I-NP	O
Ekeus	NNP	I-NP	I-PER
heads	VBZ	I-VP	O
for	IN	I-PP	O
Baghdad	NNP	I-NP	I-LOC
.	.	O	O

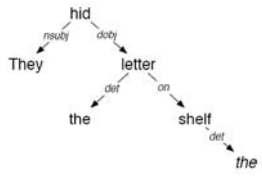
- Features for a word are typically:
  - Window of 2 words before & after the word
  - Their parts-of-speech and chunk tags

8

Dependency Parsing

- Dependency parsing is NOT partial parsing
- However the output is a list of dependency relations between chunks, extracted from the full tree

"They hid the letter on the shelf"



Dependency Relations
nsubj(they, hide)
dobj(hide, letter)
det(letter, the)
det(shelf, the)
prep_on(letter, shelf)

9