# Chunking

LING 570

Fei Xia

# What is chunking?

- Also called *partial or shallow parsing*.
- Task: to assign some additional structure to tagged input.
  - The structure is often not nested: “dividing input text into non-overlapping segments”
  - Some material in the input can be skipped over.

Ex: [The cow] in [the barn] ate …

# Why chunking?

- Used when full parsing is not feasible or not desirable.
- Often application-specific
- An example: find subcategorization frames for verbs:

  give NP to NP

  give NP NP

  give NP up

- Another example: Information Extraction (IE)

# General process

- Tokenization: The student bought two books
- POS tagging: DT N V CD N
- Chunking: NP NP
- Extraction: NP V NP
- …

# Evaluation

- System output: the set of chunks returned by the chunk parser
- Gold system: the set of chunks in the gold standard
- Correct: the correct set of chunks
- Prec = Correct/Guessed
- Recall = Correct/Gold
- F-score = 2 Prec * Recall / (Prec + Recall)

# Rule-based approach

- Longest match (Abney 1995):
  - One FSA for each phrasal category
    - NP → D? (Adj | N)* N
  - Process the input sentence from left to right
    - Find the winner for the position (i.e., the longest match)
    - If no match for a given word, skipped it (i.e., didn't chunk it)
  - Ex: NP → D? (Adj | N)* N
  - Input: "Time flies like an arrow"
  - Results: Precision 0.92, Recall 0.88

# Treating the chunking task into a sequence labeling problem

- Tagset:
  - IOB scheme:
    - B-X: first word of a chunk of type X
    - I-X: non-initial word of a chunk of type X
    - O: outside chunks
  - Other schemes: IOBE, etc.
    - B-X
    - I-X
    - O
    - E-X: the last word of a chunk of type X

# An example

| | | | | | |
|---|---|---|---|---|---|
| | We | saw | the | yellow | dog |
| | PRP | VBD | DT | JJ | NN |
| IOB: | B-NP | O | B-NP | I-NP | I-NP |
| IOBE: | B-NP | O | B-NP | I-NP | E-NP |

# Algorithms

- Any classification algorithm
  - MaxEnt
  - SVM
  - Boosting
  - …
- Any sequence labeling algorithm
  - HMM
  - CRF
  - …