# PARSING
## Shallow parsing (chunking)

*Natural Language Processing*

*Ing. R. Tedesco. PhD, AA 20-21*

# Chunking (aka shallow parsing)

- **Identifying and classifying the flat, non-overlapping segments of a sentence**
  - This set typically includes noun phrases, verb phrases, adjective phrases, and prepositional phrases
  - [$_{NP}$ The morning flight] [$_{PP}$ from] [$_{NP}$ Denver] [$_{VP}$ has arrived.]
- **Leverages POS tagging**
- **Two approaches:**
  - Finite-state rules able to catch phrase segments (FST)
  - Machine learning. We present this approach

# Shallow Parsing

| | |
|---|---|
| Company Name: | Bridgestone Sports Co |
| Verb Group: | said |
| Noun Group: | Friday |
| Noun Group: | it |
| Verb Group | had set up |
| Noun Group: | a joint venture |
| Preposition: | in |
| Location: | Taiwan |
| Preposition: | with |
| Noun Group: | a local concern |
| Conjunction: | and |
| Noun Group | a Japanese trading house |
| Verb Group: | to produce |
| Noun Group: | golf clubs |

# Tags (Penn treebank corpus)

| TAG | DESCRIPTION | WORD POS'S APPEARING | EXAMPLE | % |
|-----|-------------|---------------------|---------|---|
| NP | noun phrase | DT+RB+JJ+NN + PR | *the strange bird* | 51 |
| PP | prepositional phrase | TO+IN | *in between* | 19 |
| VP | verb phrase | RB+MD+VB | *was looking* | 9 |
| ADVP | adverb phrase | RB | *also* | 6 |
| ADJP | adjective phrase | CC+RB+JJ | *warm and cosy* | 3 |
| SBAR | subordinating conjunction | IN | *whether or not* | 3 |
| PRT | particle | RP | *up the stairs* | 1 |
| INTJ | interjection | UH | *hello* | 0 |

# CoNLL corpus

- To train stochastic chunkers

- token, POS, and chunk type

- IBO tagging, for chunk types:

  B_   begin of a chunk
  I_    inside the chunk
  O     not part of a chunk

| 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 |

# Machine learning based chunking

- The chunker slides a context window over the sentence classifying words as it proceeds
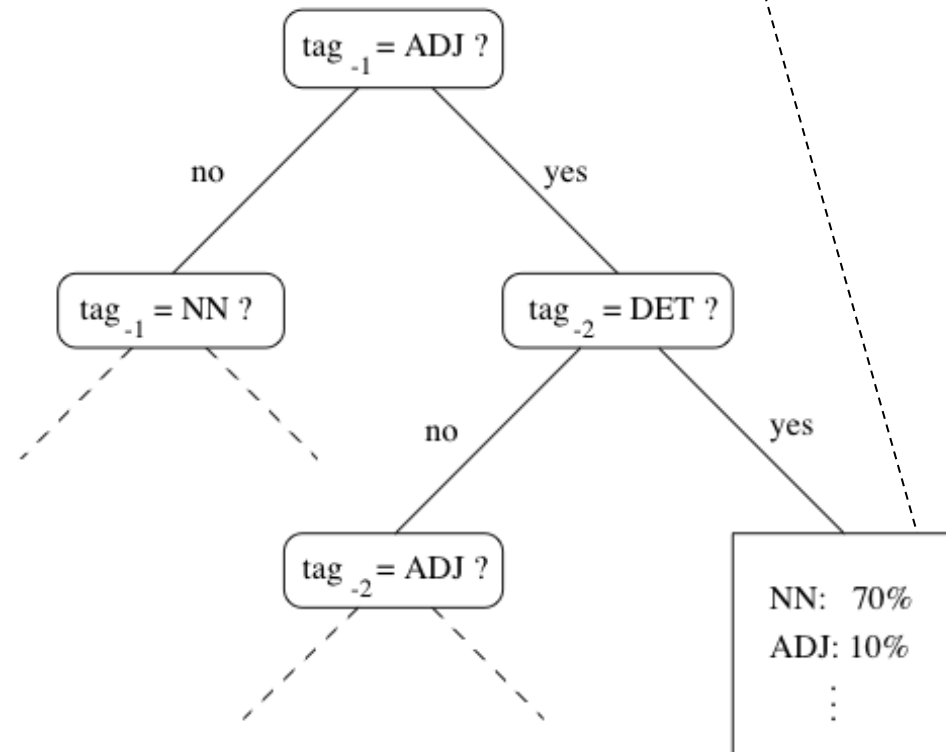- At this point the classifier is attempting to label *flights*
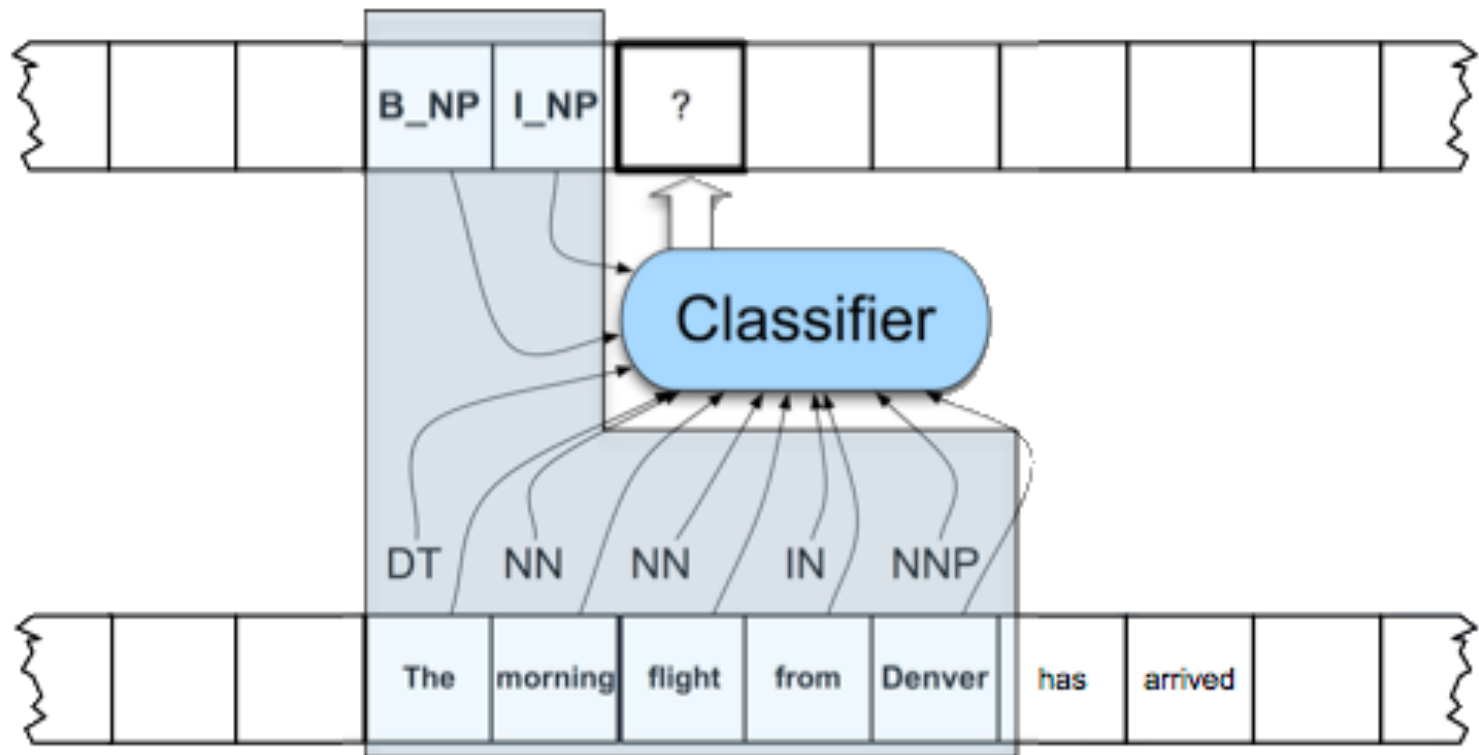
# TreeTagger

- POS tagging & chunking

- Uses a 2nd order HMM; estimates transition probability by means of a model (not an *N*-gram)

- Uses a binary decision tree

  - Built from a training corpus of trigrams with POS's

$$p(T^{(t)} \mid T^{(t-1)}, T^{(t-2)})$$



tag$_{-1}$ = ADJ ?

no / yes

tag$_{-1}$ = NN ?    tag$_{-2}$ = DET ?

no / yes

tag$_{-2}$ = ADJ ?

NN:  70%
ADJ: 10%
⋮

demo

# Illinois Chunker: POS + chunking

- HMM-based chunking
- Extends the HMM model: transition probability depends on observation

$$P(T^{(t)} \mid T^{(t-1)}, W^{(t)})$$

demo

# Shallow parsers

- CHAOS
  - http://ai-nlp.info.uniroma2.it/external/chaosproject/
- TreeTagger:
  - http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html
- Illinois
  - http://cogcomp.cs.illinois.edu/page/software_view/13