

---

# PARSING

## Shallow parsing (chunking)

---

*Natural Language Processing*

*Ing. R. Tedesco. PhD, AA 19-20*

# 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	%
<b>NP</b>	noun phrase	<b>DT+RB+JJ+NN + PR</b>	<i>the strange bird</i>	51
<b>PP</b>	prepositional phrase	<b>TO+IN</b>	<i>in between</i>	19
<b>VP</b>	verb phrase	<b>RB+MD+VB</b>	<i>was looking</i>	9
<b>ADVP</b>	adverb phrase	<b>RB</b>	<i>also</i>	6
<b>ADJP</b>	adjective phrase	<b>CC+RB+JJ</b>	<i>warm and cosy</i>	3
<b>SBAR</b>	subordinating conjunction	<b>IN</b>	<i><u>whether</u> or not</i>	3
<b>PRT</b>	particle	<b>RP</b>	<i><u>up</u> the stairs</i>	1
<b>INTJ</b>	interjection	<b>UH</b>	<i>hello</i>	0

# CoNLL corpus

- To train stochastic chunkers
- token, POS, and chunk type
- IBO tagging, for chunk types:

B\_ begin of a chunk

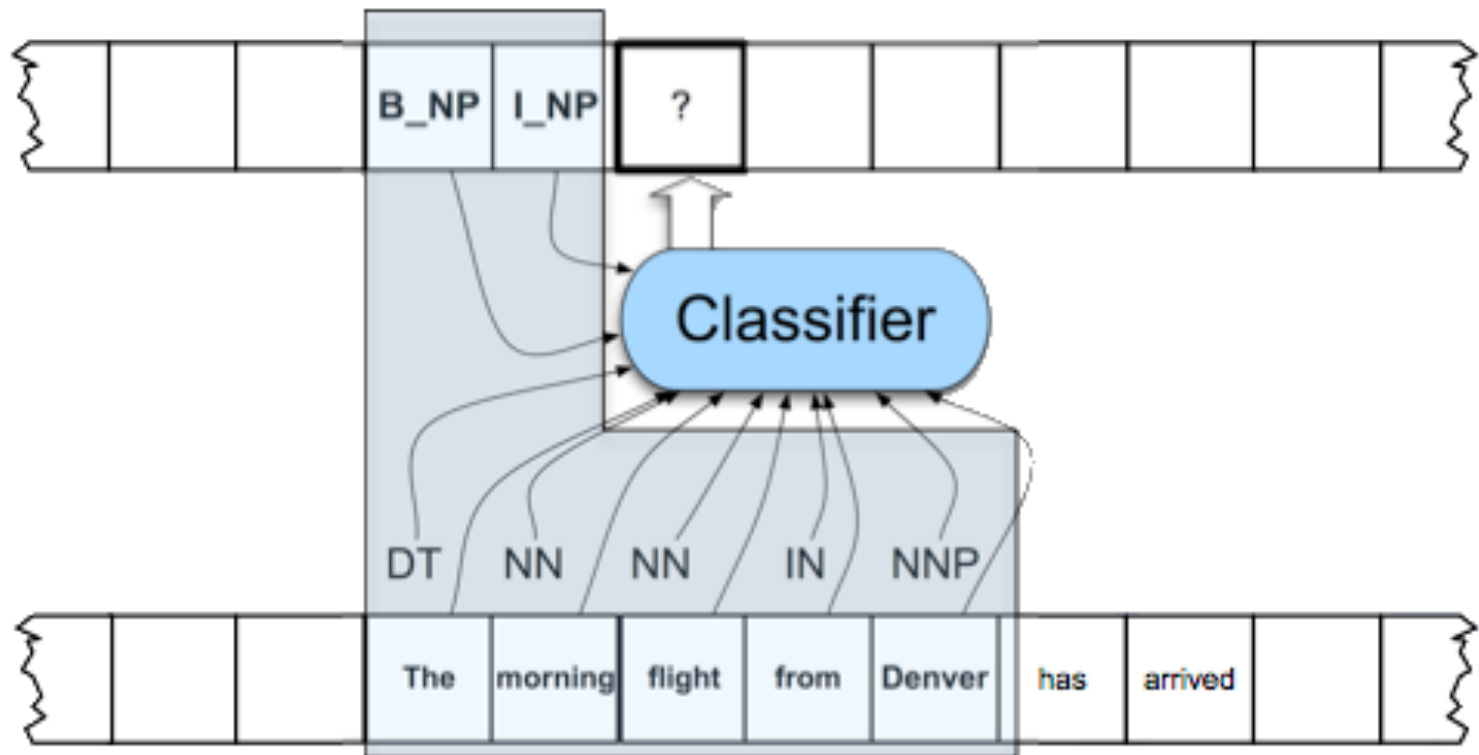
I\_ inside the chunk

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*



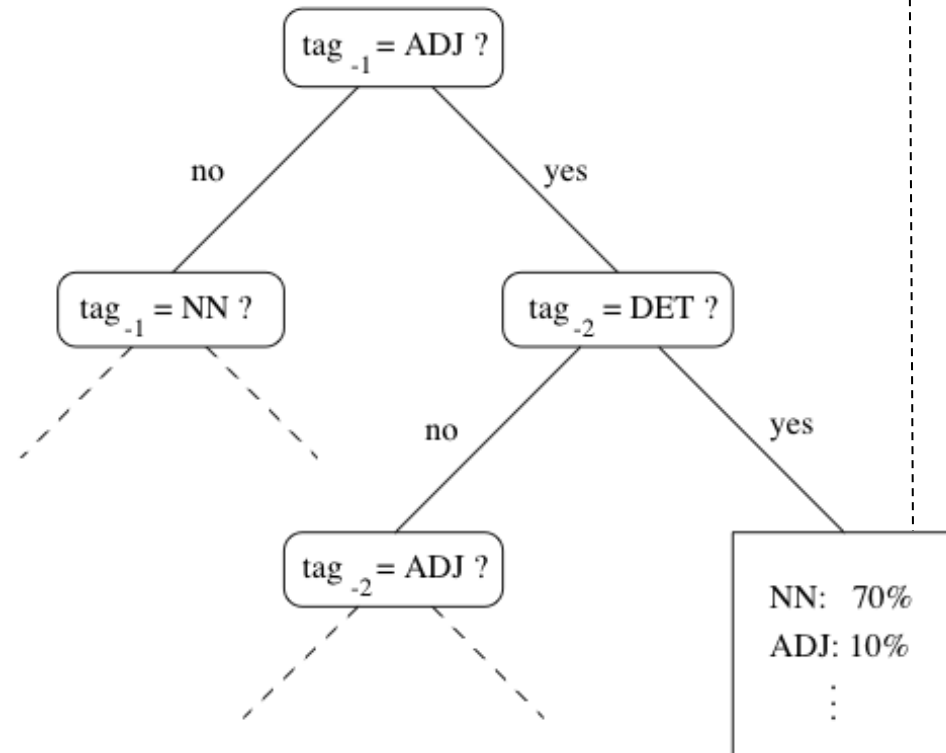
# TreeTagger

$$p(T^{(t)}=NN \mid T^{(t-1)}=ADJ, T^{(t-2)}=DET)=0.7$$
$$p(T^{(t)}=ADJ \mid T^{(t-1)}=ADJ, T^{(t-2)}=DET)=0.1$$

...

- POS tagging & chunking
- Uses a 2<sup>nd</sup> 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

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



demo

# Illinois Chunker: POS + chunking

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

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

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](http://cogcomp.cs.illinois.edu/page/software_view/13)
-