Syntax: tools
(POS tagging, chunking)

Ing. Roberto Tedesco, PhD
roberto.tedesco@polimi.it

arcslab
adaptable, relational and cognitive software environments

NLP – AA 17-18
Prof. L. Sbattella

POS tagging

- Morphological analysis considers one word at a time
- Often, this is not enough to disambiguate part-of-speech
  - E.g. *Talk*: verb or name?
- The context is *needed*
- POS taggers do just that

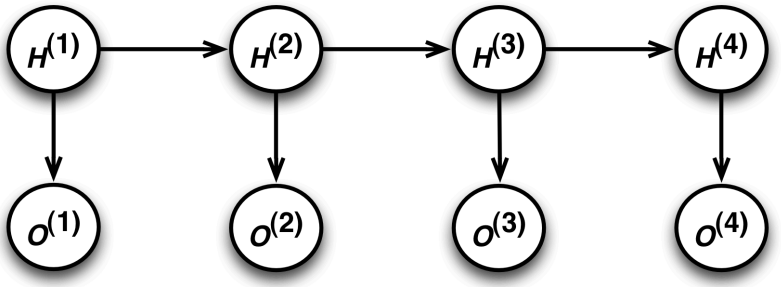
Methodologies for POS tagging

- A *language model* is needed
- Several approaches
  - Stochastic and non stochastic
- Two popular stochastic models:
  - HMM-based
  - Entropy Maximization-based

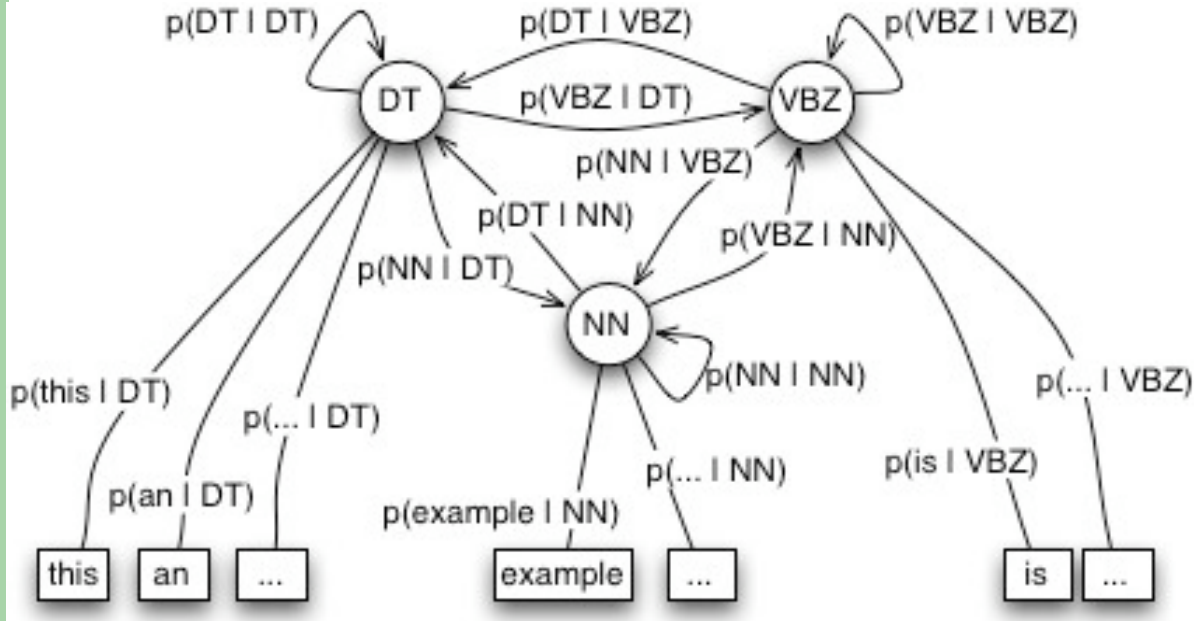
HMM POS tagging

- Works a sequence at a time
- Hidden variable: tag set
  - E,g, stochastic var. H={t_a, t_b, …}: tag set (hidden state)
- Observable variable: the set of the word forms (lexicon)
  - E.g. stochastic var. O={w_a, w_b, w_c, …}: word set (output)
- Given a sequence of words <$w^{(i)}$>, Viterbi calculates the most probable sequence of tags <$t^{(i)}$>

Hidden Markov Models (HMM)



Unrolled view

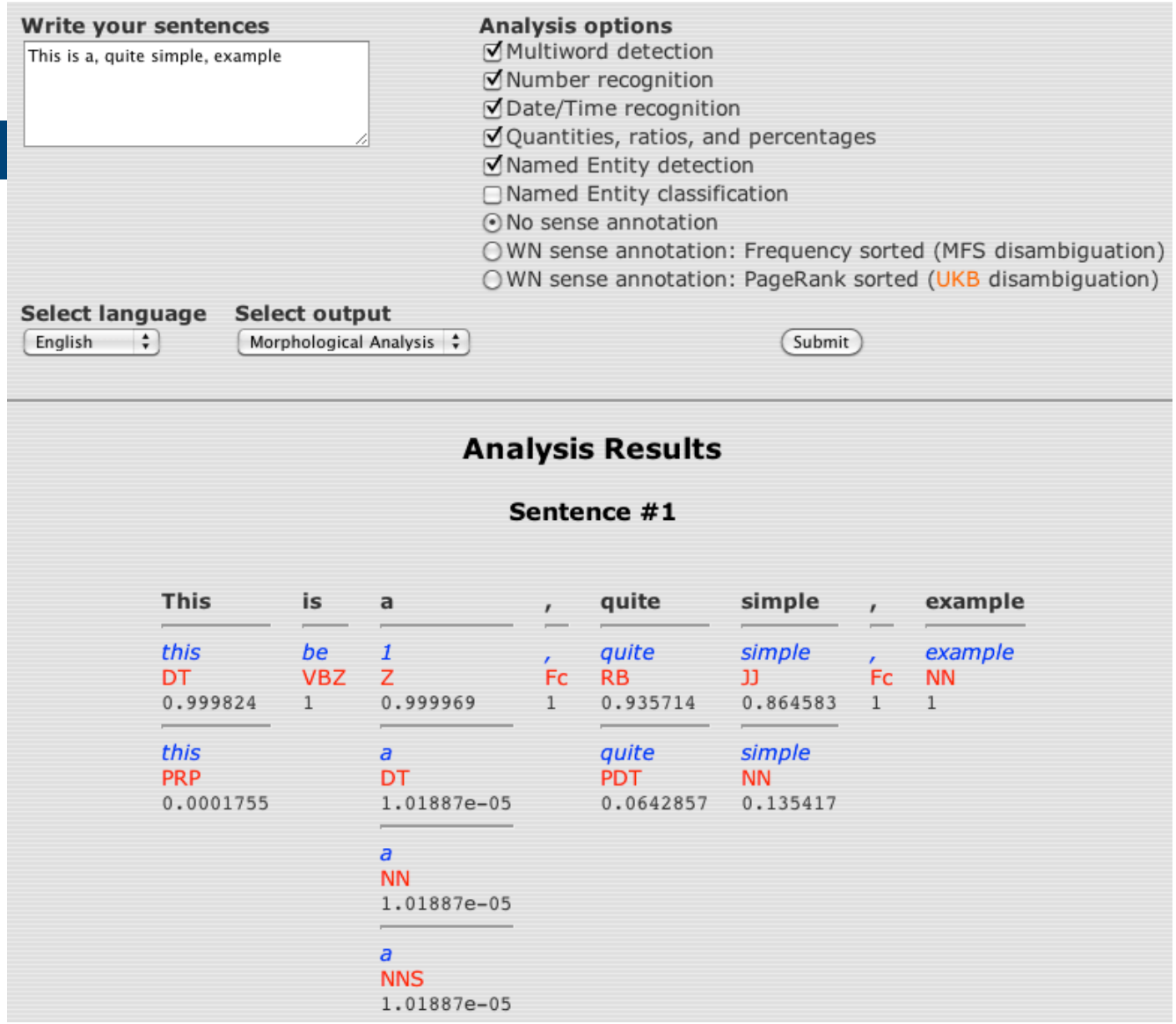


Graph view

- $H^{(t)}$={DT, VBZ, NN, … }
- $O^{(t)}$={…, an, …, example, …, is, …, this, …}

FreeLing

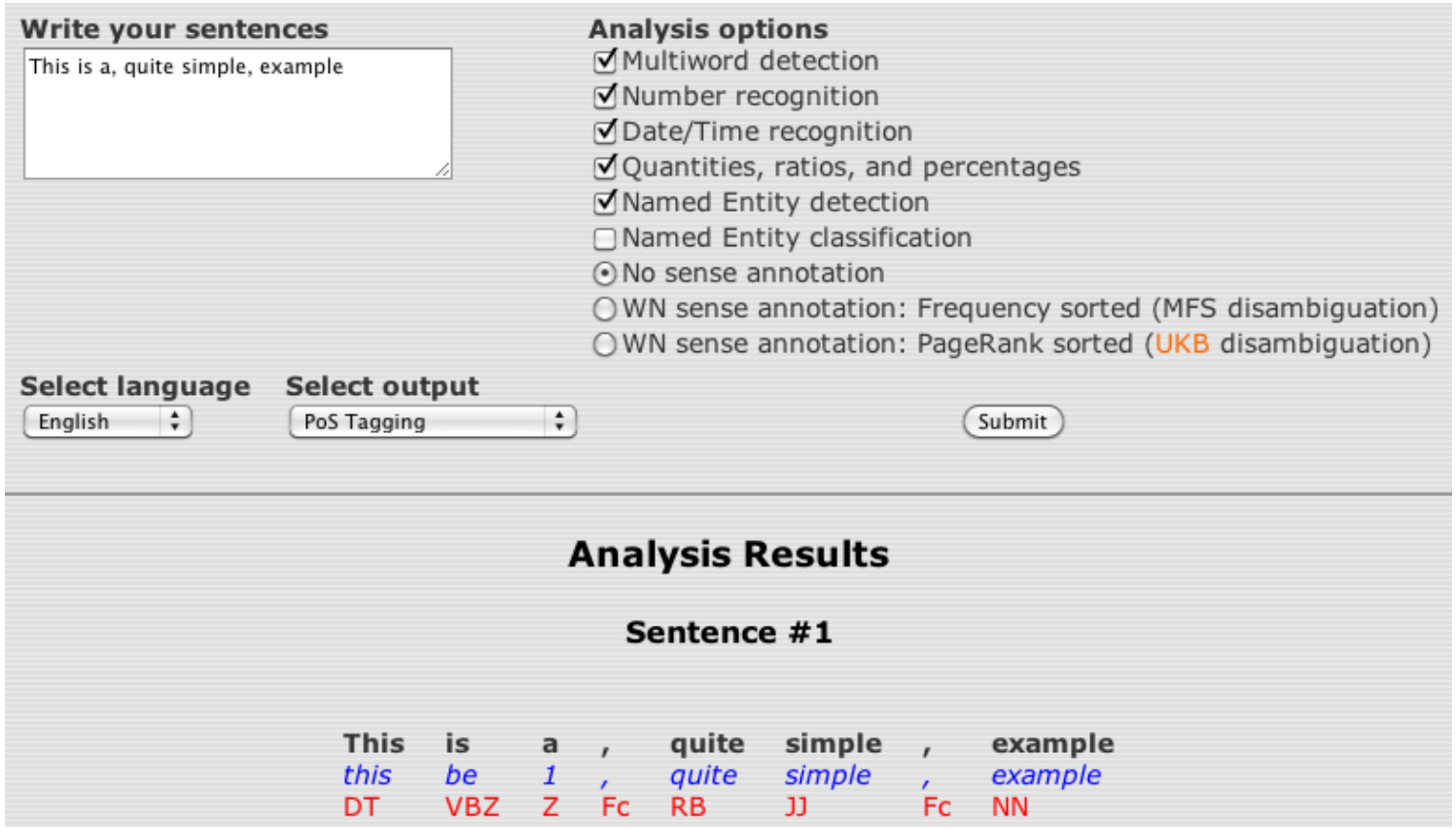
- As a morphologic analyzer



demo

FreeLing

- POS tagging
- HMM



demo

Stanford POS tagger

- Stanford POS Tagger
- Entropy Maximization
  - Uses a CMM, basically a MEMM
- Java based

- demo

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

Tags (Penn treebank corpus)

Tag	Description	Words	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

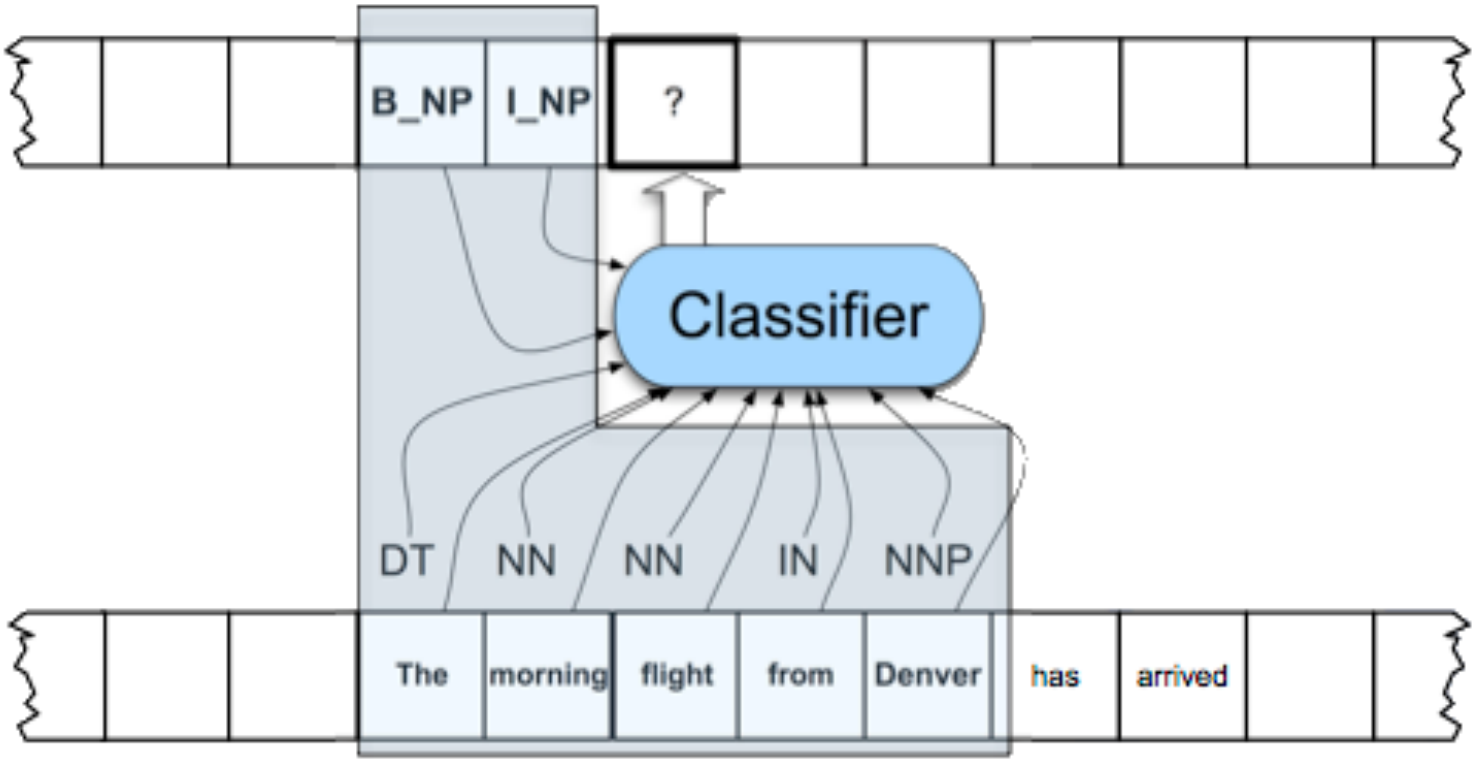
- To train stochastic chunkers
- token, POS, and chunk type
- IBO tagging, for chunk types:

  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*



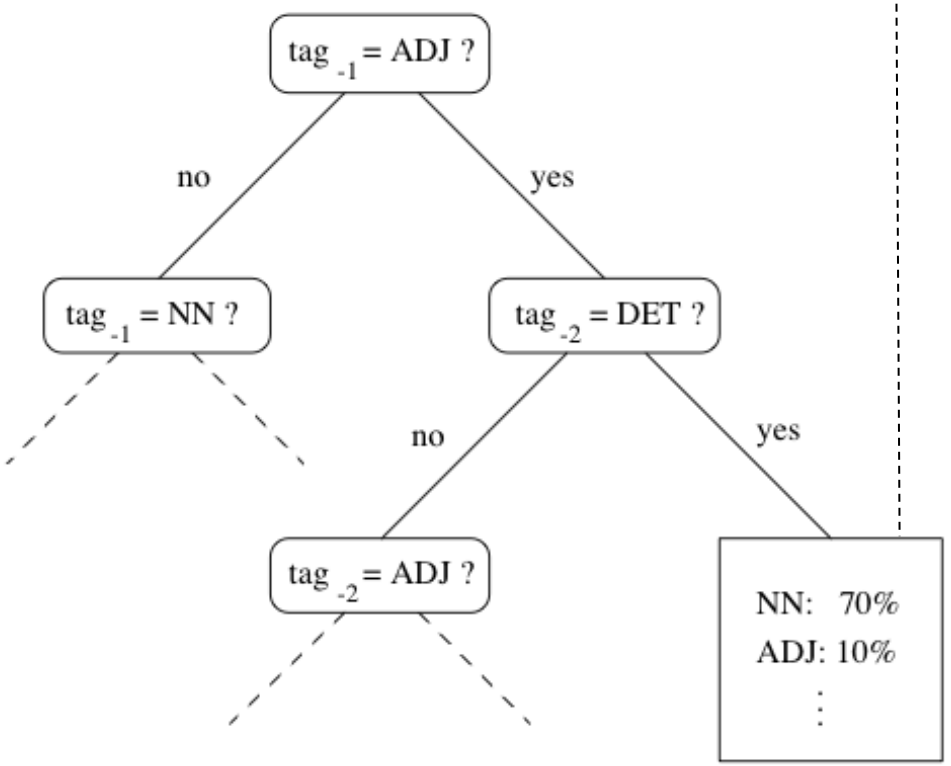
TreeTagger

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

- POS tagging
- 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

demo

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



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

REFERENCES

POS Tagging

- TreeTagger:
  - http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html
- Stanford
  - http://nlp.stanford.edu/software/index.shtml
- FreeLing (and parser, morpho analyzer, …)
  - http://nlp.lsi.upc.edu/freeling/

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