Chunking and Named Entities

MAS.S60

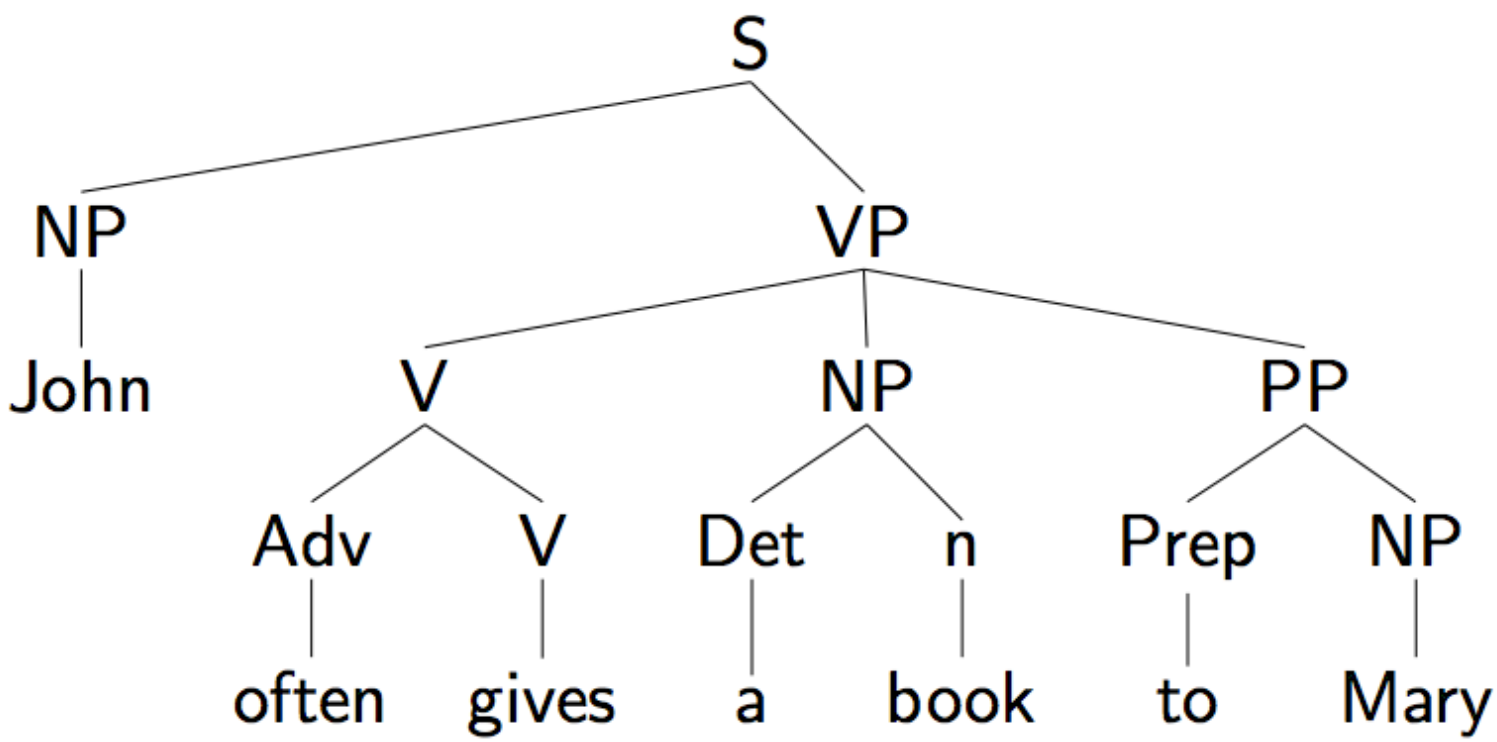
Catherine Havasi

Rob Speer

This is the (start of) the good stuff

"While Republicans point to the country's ills, Barack Obama is presenting a message of optimism, which some say could backfire if the economy declines." ~ NYTimes

Parsing



Why parse?

- A good framework for a larger, more robust, end to end system that can “sit and think”.
- Actually interested in syntax
- Machine translation – need to know what relates to what to learn correlations

Why not parse?

- Parsing is SLOW.
- Parsing is ambiguous (elephants)
- Parsing adapts badly to new domains (Twitter)

Why not parse?

- Parsing is SLOW.
- Parsing is ambiguous (elephants)
- Parsing adapts badly to new domains (Twitter)
- **Parsing doesn't get you enough bang for your buck**

What information do you need?

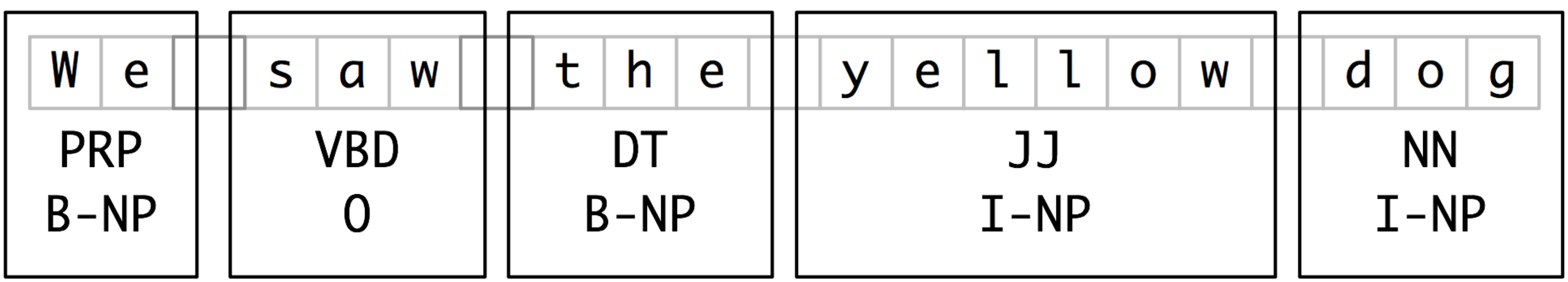
<NLTK SLIDES!>

In-class lab

- In class, we worked in groups to have a regex chunker bakeoff

The IOB representation

- Every token is **I**n a chunk or **O**ut of a chunk.
- Distinguish the **B**eginnings of chunks.
- Now chunks work just like tags



The IOB representation

- Also known as the CONLL representation
- To convert tree -> IOB:
  - nltk.chunk.tree2conlltags(tree)
- To convert IOB -> tree:
  - nltk.chunk.conlltags2tree(iob)

The machine learning approach

- A chunker is basically a tagger
- A tagger is basically a classifier

N-gram chunkers

- A unigram chunker simply assigns one chunk tag to each POS tag
  - DT = B-NP
  - NN = I-NP
  - VB = O
- F-measure = 83.2% on CONLL2000
- A bigram chunker gets f-measure = 84.5%

Parts of speech aren't enough

- Joey/NN sold/VBD the/DT farmer/NN rice/NN ./.
- Nick/NN broke/VBD my/DT computer/NN monitor/NN ./.

Chunking with feature-based classifiers

- You guessed it, Naïve Bayes again
- Can we make this classifier better by choosing the right features?

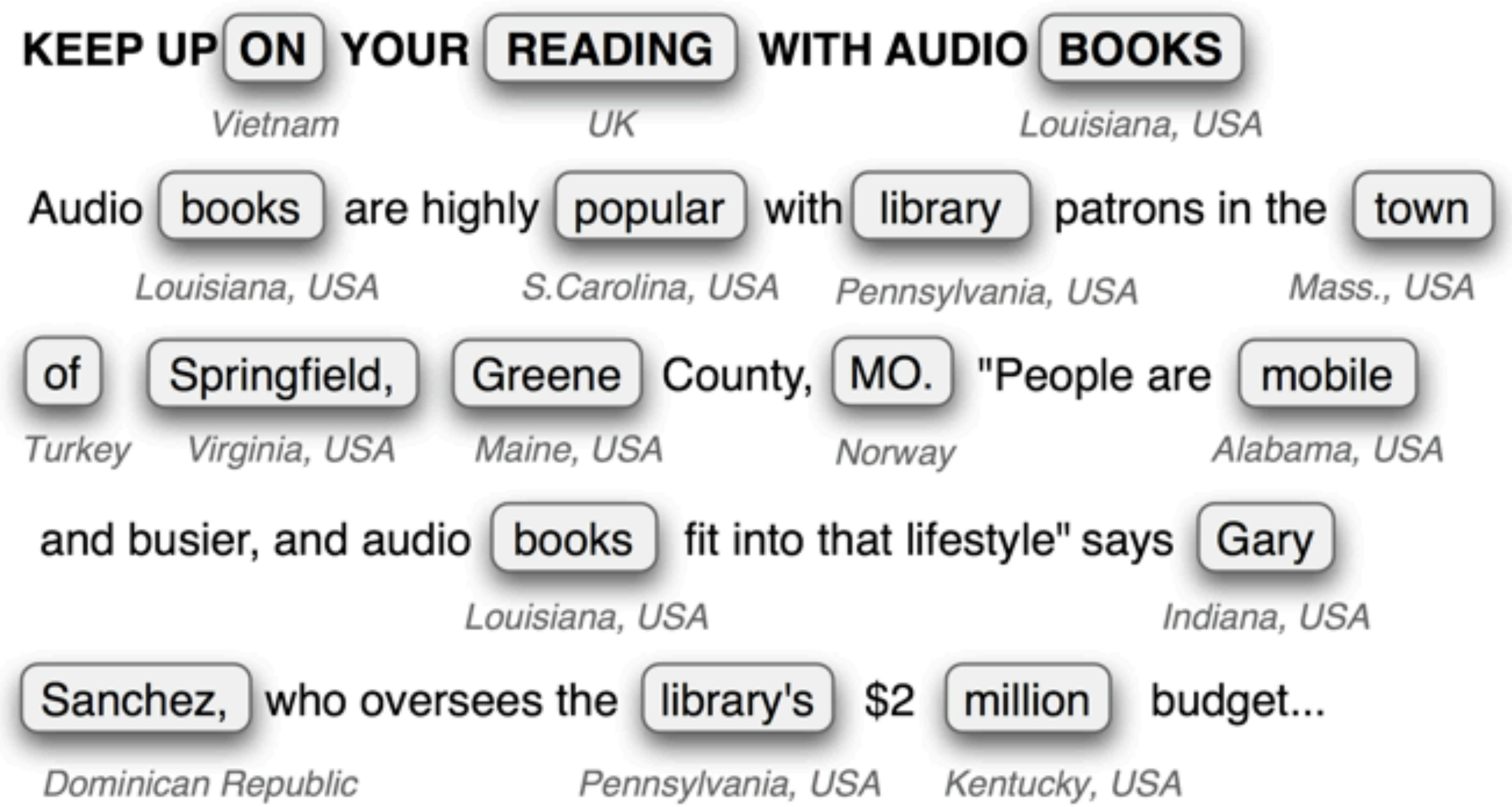
Named Entities

- Barack Obama
- Lady Gaga
- Congress
- Library (the town in PA)
- Library of Congress
- 2008-06-29
- Georgia-Pacific Corp.

Named Entity Recognition

- Sometimes “NER”
- Identify and find all mentions in unstructured text of named entities
  - Identify the boundary of the NE
  - If possible, intuit its type

Looking it up in Wikipedia



Ambiguity

- New companies happen every day
- **May** and **Christian**
- Estee Lauder

Different chunk types are different IOB tags

- Add new kinds of chunks for entity types.

Joi Ito runs the MIT Media Lab.

B-PER	I-PER	O	O	B-ORG	I-ORG	I-ORG

NLTK’s NER

- “Luckily” NLTK provides a NER Classifier nltk.ne_chunk()
  - binary=True means just tag them NE
  - Binary=False give us PERSON, ORGANIZATION, and GPE

In action!

```
>>> sent = nltk.corpus.treebank.tagged_sents()[22]
>>> print nltk.ne_chunk(sent, binary=True) [1]
(S
  The/DT
  (NE U.S./NNP)
  is/VBZ
  one/CD
  ...
  according/VBG
  to/TO
  (NE Brooke/NNP T./NNP Mossman/NNP)
  ...)
```

```
>>> print nltk.ne_chunk(sent)
(S
  The/DT
  (GPE U.S./NNP)
  is/VBZ
  one/CD
  ...
  according/VBG
  to/TO
  (PERSON Brooke/NNP T./NNP Mossman/NNP)
  ...)
```

State of the Art

- **Stanford Named Entity Recognizer (NER)**
- http://nlp.stanford.edu/software/CRF-NER.shtml
- Uses Gibbs Sampling to converge a Conditional Random Field
  - Uses the Markov property
  - You probably don't care how it works
  - And the good thing is, you don't need to!

Features Used

Feature	NER	TF
Current Word	✓	✓
Previous Word	✓	✓
Next Word	✓	✓
Current Word Character n-gram	all	length ≤ 6
Current POS Tag	✓	
Surrounding POS Tag Sequence	✓	
Current Word Shape	✓	✓
Surrounding Word Shape Sequence	✓	✓
Presence of Word in Left Window	size 4	size 9
Presence of Word in Right Window	size 4	size 9

Table 2: Features used by the CRF for the two tasks: named entity recognition (NER) and template filling (TF).

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005),* pp. 363-370

Assignment

- We made some chunkers that were reasonably successful at the CoNLL 2000 chunking task.
- Now, do it again, in Dutch
  - CoNLL 2002: Named entity recognition in Dutch and Spanish
- Baseline: F = 40.8%. You can do better!

Slide Credits

- More than always: Steven Bird, Ewan Klein, Ed Loper & NLTK