Language Processing with Python

Methods in Computational Linguistics I
September 3

Last Time

- Introduction to the Class
- First look at Python
- and NLTK

Today

- Language Processing with Python (Chapter 1)
- Strings
- Using NLTK
- Parsing

Natural Language Toolkit

- Text material
  - Raw text
  - Annotated Text
- Tools
  - Part of speech taggers
  - Semantic analysis
- Resources
  - WordNet, Treebanks

NLTK Demo

- Demo time!

Major CL tasks

- Part of Speech Tagging
- Parsing
- Word Net
- Named Entity Recognition
- Information Retrieval
- Sentiment Analysis
- Document Clustering
- Topic Segmentation
- Authoring
- Machine Translation
- Summarization
- Information Extraction
- Spoken Dialog Systems
- Natural Language Generation
- Word Sense Disambiguation

Part of Speech Tagging

- Task: Given a string of words, identify the parts of speech for each word.

A	man	walks	into	a	bar.
Det.	Noun	Verb	Prep.	Det.	Noun

Part of Speech Tagging

- Surface level syntax.
- Primary operation
  - Parsing
  - Word Sense Disambiguation
  - Semantic Role labeling
  - Segmentation
    - Discourse, Topic, Sentence

How do we do it?

- Learn from Data.
- Annotated Data:

A	man	walks	into	a	bar.
Det.	Noun	Verb	Prep.	Det.	Noun

- Unlabeled Data:

A man walks home.

The pitcher issued four walks.

Part of speech tagging

	Det	Noun	Verb	Prep	Adj
A	0.9	0.1	0.0	0.0	0.0
man	0.0	0.6	0.2	0.0	0.2
walks	0.0	0.2	0.8	0.0	0.0
into	0.0	0.0	0.0	1.0	0.0
bar	0.0	0.7	0.3	0.0	0.0

Limitations

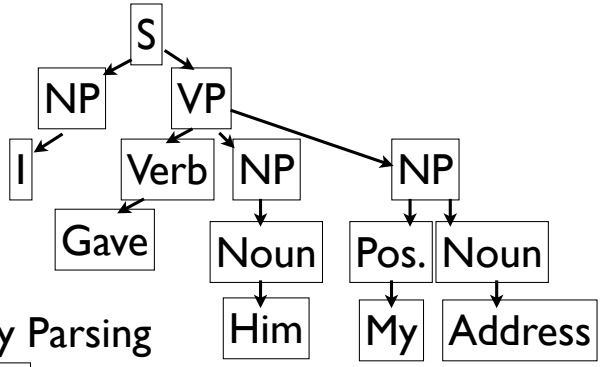
- Unseen tokens
- Uncommon interpretations
- Long term dependencies

Parsing

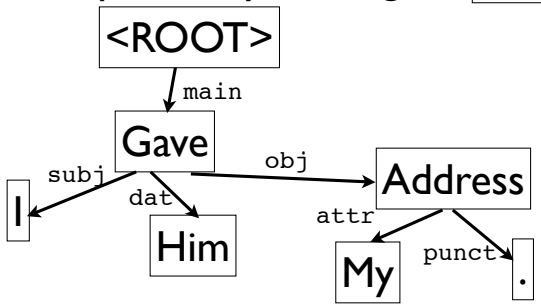
- Generate a Parse Tree from:
  - The surface form (words) of the text
  - Part of Speech Tokens

Parsing Styles

- Parse Trees



- Dependency Parsing



Context Free Grammars for Parsing

- S → VP
- S → NP VP
- NP → Det Nom
- Nom → Noun
- Nom → Adj Noun
- VP → Verb Noun

Using these rules

- Construct a parse that fits the desired text.

Limitations

- The grammar must be built by hand.
- Can't handle ungrammatical sentences.
- Can't resolve ambiguity.

Probabilistic Parsing

- Assign each transition a probability
- Find the parse with the greatest "likelihood"

Next Time

- Text Corpora in NLTK