Corpora in NLTK

Methods in Computational Linguistics I
Lesson 3

Last Time

- Text Processing

Today

- Text Corpora in NLTK
- for loops
- stored python scripts
- WordNet

What is a corpus?

- A (relatively) unprocessed set of data
- Penn Treebank
- WSJ
- WordNet?
- Switchboard

How are corpora used in CL?

- Corpora are the lifeblood of computational linguistics research.
- Collection and Dissemination of Corpora
- Descriptive Statistics
- Statistical Modeling
- Consistent Evaluation

Collection and Dissemination

- Collecting language material is a valuable research outcome.
- Endangered languages
- Hard-to-find data
  - Authentic emotion
  - Proprietary information
  - fMRI
  - Articulatory data

Descriptive Statistics

- Describe some linguistic phenomenon using statistics.
- How likely is the word "tweet" to be a verb?
- Is a sentence more likely to start with the word "the" or "baboon"?
- What language use phenomena are associated with positive product reviews?

Statistical Modeling

- Use statistics to make predictions about unseen language data.
- Train a statistical model based on part of a corpus.
- Evaluate on unseen material.
- Models are task specific.

Consistent Evaluation

- Different approaches to a common task can be evaluated on common material.
- “Shared Tasks”
- This allows for “state-of-the-art” results to be established and verified.

Using Corpora with NLTK and Python

- NLTK includes many corpora
- Books of the bible
- Public domain books (Project Gutenberg)
- News
- Web chat
- Blogs
- Some non-English material as well.

NLTK Corpus Demos

- What corpora are included
- conditional frequency distributions
- for loops
- python scripts
  - calling python from outside the interpreter

NLTK resources

- Corpora often also include data annotations
- NLTK has a variety of methods to access corpus annotations.
  - corpus.sents()
  - Author
  - Part of speech (POS) tags
  - Parse tree information
  - Word net annotations
  - Discourse information
  - etc.
- We'll come back to these as needed.

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

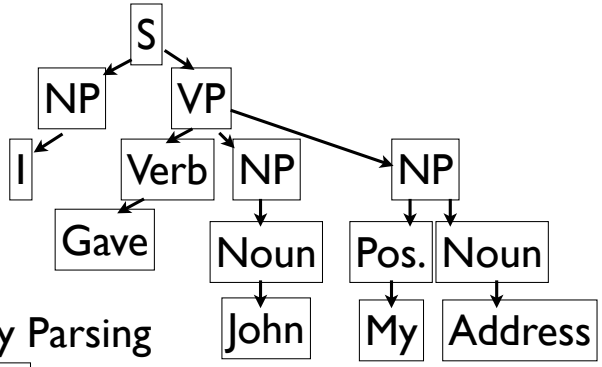
- Unseen tokens
- Uncommon interpretations
- Long term dependencies

Parsing

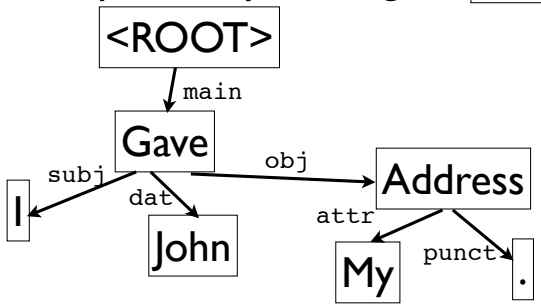
- Generate a Parse Tree from:
  - The surface form (words) of the text
  - Part of Speech Tokens

Parsing Styles

- Parse Trees



- Dependency Parsing



Context Free Grammars for Parsing

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

- **Det** → "A", "The"
- **Noun** → "I", "John", "Address"
- **Verb** → "Gave"
- **Adj** → "My", "Blue"
- **Adv** → "Quickly"

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

- Functions, Lists and Tuples