Advanced NLP Processing

Beyond tokens…

Dan Jurafsky

NLP Text Processing Pipeline

nltk covers POS tagging, phrase chunking

Stanford NLP toolkit provides parsing, coreference, NER

- Document → Sections and Paragraphs
- Paragraphs → Sentences (sentence segmentation / extraction)
- Sentences → Tokens
- Tokens → Lemmas or Morphological Variants / Stems
- **Tokens → Part-of-speech (POS) Tags**
- **Tokens, POS Tags → Phrase Chunks (Named entities and Keyphrases)**
- **Tokens, POS Tags → Parse Trees**
  - **Augment above with coreference, entailment, sentiment, …**

Part-of-speech tagging

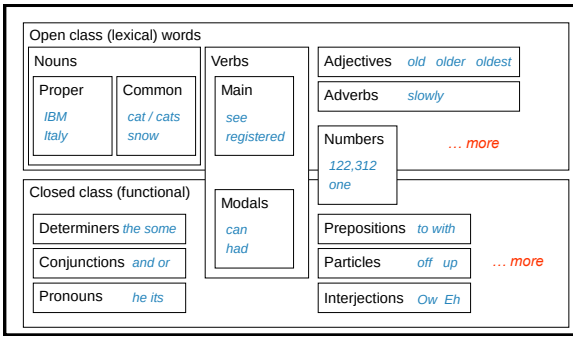
A simple but useful form of linguistic analysis

Christopher Manning

Dan Jurafsky

Parts of Speech

- Perhaps starting with Aristotle in the West (384–322 BCE), there was the idea of having parts of speech
  - a.k.a lexical categories, word classes, "tags", POS
- It comes from Dionysius Thrax of Alexandria (c. 100 BCE) the idea that is still with us that there are 8 parts of speech
  - But actually his 8 aren't exactly the ones we are taught today
    - Thrax: noun, verb, article, adverb, preposition, conjunction, participle, pronoun
    - School grammar: noun, verb, adjective, adverb, preposition, conjunction, pronoun, interjection



Dan Jurafsky

Open vs. Closed classes

- Open vs. Closed classes
  - Closed:
    - determiners: ***a, an, the***
    - pronouns: ***she, he, I***
    - prepositions: ***on, under, over, near, by, …***
    - Why "closed"?
  - Open:
    - Nouns, Verbs, Adjectives, Adverbs.

Dan Jurafsky

POS Tagging

- Words often have more than one POS: *back*
  - The *back* door = JJ
  - On my *back* = NN
  - Win the voters *back* = RB
  - Promised to *back* the bill = VB
- The POS tagging problem is to determine the POS tag for a particular instance of a word.

Dan Jurafsky

POS Tagging

- Input: Plays well with others
- Ambiguity: NNS/VBZ UH/JJ/NN/RB IN NNS
- Output: Plays/VBZ well/RB with/IN others/NNS
- Uses:
  - Text-to-speech (how do we pronounce "lead"?)
  - Can write regexps like (Det) Adj* N+ over the output for phrases, etc.
  - As input to or to speed up a full parser
  - If you know the tag, you can back off to it in other tasks

Penn Treebank POS tags

Dan Jurafsky

POS tagging performance

- How many tags are correct? (Tag accuracy)
  - About 97% currently
  - But baseline is already 90%
    - Baseline is performance of stupidest possible method
      - Tag every word with its most frequent tag
      - Tag unknown words as nouns
  - Partly easy because
    - Many words are unambiguous
    - You get points for them (*the*, *a*, etc.) and for punctuation marks!

Dan Jurafsky

Deciding on the correct part of speech can be difficult even for people

- Mrs/NNP Shaefer/NNP never/RB got/VBD around/RP to/TO joining/VBG
- All/DT we/PRP gotta/VBN do/VB is/VBZ go/VB around/IN the/DT corner/NN
- Chateau/NNP Petrus/NNP costs/VBZ around/RB 250/CD

Dan Jurafsky

How difficult is POS tagging?

- About 11% of the word types in the Brown corpus are ambiguous with regard to part of speech
- But they tend to be very common words. E.g., *that*
  - I know *that* he is honest = IN
  - Yes, *that* play was nice = DT
  - You can't go *that* far = RB
- 40% of the word tokens are ambiguous

Phrase Chunking and Special Noun Phrases

Dan Jurafsky

Phrase Chunking

- Find all non-recursive noun phrases (NPs) and verb phrases (VPs) in a sentence.
  - [NP I] [VP ate] [NP the spaghetti] [PP with] [NP meatballs].
  - [NP He] [VP reckons] [NP the current account deficit] [VP will narrow] [PP to] [NP only # 1.8 billion] [PP in] [NP September]

13

Dan Jurafsky

Phrase Chunking as Sequence Labeling

- Tag individual words with one of 3 tags
  - B (Begin) word starts new target phrase
  - I (Inside) word is part of target phrase but not the first word
  - O (Other) word is not part of target phrase
- Sample for NP chunking
  - He reckons the current account deficit will narrow to only # 1.8 billion in September.

Begin **Inside** **Other**

14

Dan Jurafsky

Named Entity Recognition (NER)

- A special class of **Proper Noun Phrases**
- **People:** Scott Sanner, President Obama, Madonna
- **Places:** New York, Madison Square Garden, Millenium Park
- **Organizations:** New York Times, University of Toronto

15

Dan Jurafsky

Keyphrases

- Useful noun phrases, but not necessarily Proper Nouns, e.g.,
  - "machine learning"
  - "support vector machines"
  - "genetically modified organisms"
- A subset of frequent noun phrases (harder to extract than NEs)
  - This paper has the best method I've found so far:
    "Automatic Recognition of Multi-Word Terms: the C-value/NC-value Method
    Katerina Frantziy, Sophia Ananiadouy, Hideki Mima" IJODL 2000.
    http://personalpages.manchester.ac.uk/staff/sophia.ananiadou/ijodl2000.pdf

16

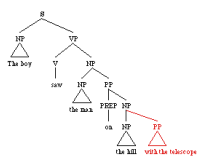
Statistical Natural Language Parsing

Parsing: Two views of syntactic structure

17

Dan Jurafsky

Why parsing?



- "The boy saw the man on the hill with the telescope."
  - Who had the telescope?
- Depends on whether you attach "with the telescope" to "I" or "man on the hill"
- How do you determine attachments? Parsing.
  - Some sentences are inherently ambiguous: attachment ambiguity.

18

Dan Jurafsky

For fun

- Who polices the police?
- Police police police police.
- Who polices the police police? ☺
- Point: we need more than word order / POS to interpret sentences… we need structure.

19

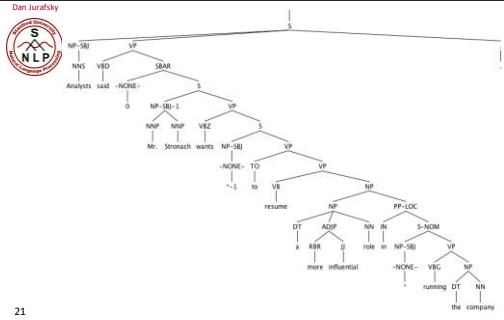
Dan Jurafsky

Two views of linguistic structure: 1. Constituency (phrase structure)

- Phrase structure organizes words into nested constituents.
- What is a constituent?
- Constituent behaves as unit that can appear in different places:
  - John talked [to the children] [about drugs].
  - John talked [about drugs] [to the children].
  - Substitution/expansion/pro-forms:
    - I sat [on the box/right on top of the box/there].
  - Coordination, regular internal structure, no intrusion, fragments, semantics, …

20

Dan Jurafsky



21

Dan Jurafsky

Grammars for Parse Tree Production

- Parent → Child1 Child2 | Child3 Child4 … | …
- S → NP VP | …
- NP → … NN* …
- VP → … VB* …
- ADJP → … JJ* …
- ADVP → … RB* …
- …

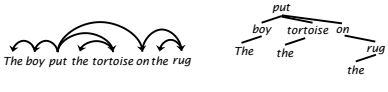
22

Dan Jurafsky

Two views of linguistic structure: 2. Dependency structure

- Dependency structure shows which words depend on (modify or are arguments of) which other words.

The boy put the tortoise on the rug



23

Dan Jurafsky

Two views of linguistic structure: 2. Dependency structure

- Dependency structure shows which words depend on (modify or are arguments of) which other words.
- Can derive dependency tree from parse tree
- What about reverse?

24

Semantic Language Analysis

Coreference and entailment

Dan Jurafsky

Coreference

- Discourse (multiple sentences) use coreferring phrases.
- Example:
  "John saw a beautiful Acura Integra in the dealership.
  He showed it to Bob.
  He bought it."
- What do "He" and "it" refer to in the 2nd sentence?

Dan Jurafsky

Coreference Resolution

- "John saw a beautiful Acura Integra in the dealership.
  He1 showed it1 to Bob.
  He2 bought it2."

Referent	Phrases
John	{John, He1, He2}
Integra	{a beautiful Acura Integra, it1, it2}
Bob	{Bob}
dealership	{the dealership}

- Important in processing reviews: "I liked it!"

Dan Jurafsky

Entailment

- **Question:** When did the Berlin wall open?
- **Text contains:** The Berlin wall fell on November 9, 1989.
- Simple entailment? Does "fall" → "open"?
  - A wall falling is a wall opening
  - A person falling is not a person opening
- Entailment can be highly contextual. But WordNet (in nltk) contains basic entailments, e.g., "snoring" → "sleeping".