Part of Speech Tagging with NLTK

Methods in Computational Linguistics I

Last Time

- A lot of regular expressions

Automatic Part of Speech Tagging

- Input: A string of text
  - Typically a sentence
  - Sometimes with punctuation
- Output: Part of speech tags

POS Tag inventories

- Any POS Tagger will be trained on a specific Tag set.
- These represent the only tags that can be generated.

Penn Treebank tag set

Simple Tag Set - 19 tags (down from ~45)

Tag	Meaning	Examples
ADJ	adjective	new, good, high, special, big, local
ADV	adverb	really, already, still, early, now
CNJ	conjunction	and, or, but, if, while, although
DET	determiner	the, a, some, most, every, no
EX	existential	there, there's
FW	foreign word	dolce, ersatz, esprit, quo, maitre
MOD	modal verb	will, can, would, may, must, should
N	noun	year, home, costs, time, education
NP	proper noun	Alison, Africa, April, Washington

Tag	Meaning	Examples
NUM	number	twenty-four, fourth, 1991, 14:24
PRO	pronoun	he, their, her, its, my, I, us
P	preposition	on, of, at, with, by, into, under
TO	the word to	to
UH	interjection	ah, bang, ha, whee, hmpf, oops
V	verb	is, has, get, do, make, see, run
VD	past tense	said, took, told, made, asked
VG	present participle	making, going, playing, working
VN	past participle	given, taken, begun, sung
WH	wh determiner	who, which, when, what, where, how

nltk.help.upenn_tagset('RB')

nltk.help.upenn_tagset()

Limitations of automatic tagging

- Legitimate ambiguity
  - "Teacher strikes idle children"
  - "Flying planes can be dangerous"
  - "Time flies like an arrow" vs. "Time flies like bananas"
- Garden path sentences
  - "The horse raced past the barn fell"
  - "The man whistling tunes pianos"
  - "The government plans to raise taxes were defeated"
- Out of vocabulary items (OOV)
- Out of domain material

A simple tagger

- Identify the most common tag for every word.
- Tag every OOV word as a Proper Noun
- If in the same domain, with high coverage, this tagger can reach ~90% accuracy.

How does it work?

- Label a large corpus.
- "Model" the statistics of p(tag | word)
- "Model" the statistics of p(tag | tag-1)

- Using NLTK you can make your own tagger with simple regular expressions.

Upper-bound on Tagger performance

- POS Taggers typically use only surface information.
- Any decisions that require long-range syntactic dependency, semantic or pragmatic understanding cannot be addressed.

Using the NLTK Tagger

- NLTK built in tagger
- Brown Corpus
- Processing tag data
- Building a new tagger

Next Time

- Parsing with NLTK