

POS Tagging

Sudeshna Sarkar

31 July 2019

Parts of speech

Parts of speech are constructed by grouping words that function similarly:

- with respect to the words that can occur nearby
- and by their morphological properties

The man all the way home.

- Aristotle (384–322 BCE): the idea of having parts of speech a.k.a lexical categories, word classes, “tags”, POS
- Dionysius Thrax of Alexandria (c. 100 BCE) : 8 parts of speech
 - noun, verb, article, adverb, preposition, conjunction, participle, pronoun

English parts of speech

- 8 parts of speech?
 - Noun (person, place or thing)
 - Verb (actions and processes)
 - Adjective (modify nouns)
 - Adverb (modify verbs)
 - Preposition (on, in, by, to, with)
 - Determiners (a, an, the, what, which, that)
 - Conjunctions (and, but, or)
 - Particle (off, up)

Brown corpus: 87 POS tags

Penn Treebank: ~45 POS tags

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.

Closed vs. Open Class

- ***Closed class*** categories are composed of a small, fixed set of grammatical function words for a given language.
 - Pronouns, Prepositions, Modals, Determiners, Particles, Conjunctions
- Open class categories have large number of words and new ones are easily invented.
 - Nouns (Googler, textlish), Verbs (Google), Adjectives (geeky), Adverb (chompingly)



Open class (lexical) words

Nouns

Proper

IBM

Italy

Mohan

Common

cat / cats

snow

Kitaaba,
kalama,

Verbs

Main

see

registered

giraa, gayaa

Adjectives

old older oldest

sundara, acchaa,

Adverbs

slowly

jaldii, teza

Numbers

122,312

one

... more

Closed class (functional)

Determiners *the some*

Conjunctions *and or*
aur, agar

Pronouns *he its*
Vaha, main

Modals

can

had

Prepositions *to with*

Particles *off up*
to, bhii, hii

Interjections *Ow Eh*

... more

Part Of Speech Tagging

- Annotate each word in a sentence with a part-of-speech marker.

John saw the saw and decided to take it to the table.
NNP VBD DT NN CC VBD TO VB PRP IN DT NN

UDEP POS tags

Open class words	Closed class words	Other
<u>ADJ</u>	<u>ADP</u>	<u>PUNCT</u>
<u>ADV</u>	<u>AUX</u>	<u>SYM</u>
<u>INTJ</u>	<u>CCONJ</u>	<u>X</u>
<u>NOUN</u>	<u>DET</u>	
<u>PROPN</u>	<u>NUM</u>	
<u>VERB</u>	<u>PART</u>	
	<u>PRON</u>	
	<u>SCONJ</u>	

Ambiguity in POS Tagging

- “Like” can be a verb or a preposition
 - I **like**/VBP candy.
 - Time flies **like**/IN an arrow.
- “Around” can be a preposition, particle, or adverb
 - I bought it at the shop **around**/IN the corner.
 - I never got **around**/RP to getting a car.
 - A new Prius costs **around**/RB \$25K.
- What is the POS for “back”?
 - The back door
 - On my back
 - Win the voters back
 - Promised to back the bill

POS Tagging task

- Input: **the lead paint is unsafe**
- Output: **the/Det lead/N paint/N is/V unsafe/Adj**
- Uses:
 - text-to-speech (how do we pronounce “lead”?)
 - can differentiate word senses that involve part of speech differences (what is the meaning of “interest”)
 - can write regexps like `Det Adj* N*` over the output (for filtering collocations)
 - preprocessing for parser

Ambiguity in POS tagging

Like most language components, the challenge with POS tagging is ambiguity

Brown corpus analysis

- 11.5% of word types are ambiguous (this sounds promising!), **but...**
- 40% of word appearances are ambiguous
- Unfortunately, **the ambiguous words tend to be the more frequently used words**

Constituency

Parts of speech can be thought of as the lowest level of syntactic information

Groups *words* together into categories

 likes to eat candy.

What can/can't go here?

Constituency

_____ likes to eat candy.

nouns

Dinesh

Dr Roy

Professor Das

determiner nouns

The man

The boy

The cat

pronouns

He

She

determiner nouns +

The man that I saw

The boy with the blue pants

The cat in the hat

Constituency

Words in languages tend to form into functional groups (parts of speech)

Groups of words (aka phrases) can also be grouped into functional groups

- often some relation to parts of speech
- though, more complex interactions

These phrase groups are called constituents

POS Tagging Approaches

- **Rule-Based**: Human crafted rules based on lexical and other linguistic knowledge.
- **Learning-Based**: Trained on human annotated corpora like the Penn Treebank.
 - **Statistical models**: Hidden Markov Model (HMM), Maximum Entropy Markov Model (MEMM), Conditional Random Field (CRF)
 - **Rule learning**: Transformation Based Learning (TBL)
 - **Neural networks**: Recurrent networks like Long Short Term Memory (LSTMs)