More on POS Tagging

27 November 2024 18:36

Recap on parts of speech:

N	noun	chair, bandwidth, pacing
V	verb	study, debate, munch
ADJ	adjective	purple, tall, ridiculous
ADV	adverb	unfortunately, slowly,
P	preposition	of, by, to
PRO	pronoun	I, me, mine
DET	determiner	the, a, that, those

Closed class of words
- don't accept new members
- small and stable.
- core words → pronouns, prepositions, conjunctions, determiners

Open class.
- accept new members
- Nouns, Verbs, Adjectives, Adverbs

3 ways of POS Tagging → Rule based — SNGTWOL?
→ Transformation based — Brill tagger
→ Stochastic — HMM tagging

A. Rule Based

1. Start w/ a dictionary
2. Use the dict to assign every possible tag.

			NN		
			RB		
	VBN		JJ		VB
PRP	VBD	TO	VB	DT	NN
She	promised	to	back	the	bill

3. Write rules to eliminate tags (like, grammar stuff: x shouldn't follow y and stuff)
↳ Selectively remove tags

			~~NN~~		
			~~RB~~		
	~~VBN~~		~~JJ~~		~~VB~~
PRP	VBD	TO	VB	DT	NN
She	promised	to	back	the	bill

B. Statistical tagging

1. Create dict. w/ each possible tag → $P(V|race) = \frac{\text{count(race is verb)}}{\text{total count(race)}}$
↳ count no. of times each tag occurs for a word.
2. Given a new sentence, pick most freq. tag using (1)

2. Given a new sentence, pick most freq. tag using (1)

C. Transformation based

[Supervised (needs corpus)]

Quick job first, refine using context rules.
(probable)

D. Stochastic

- HMM Taggers
  - Goal: P(tag|word)
    - Maximize: P(word|tag) × P(tag|previous 'n' tags)
- ANN tagger → multiclass classification

* Other sources of disambiguation
  - → words w/ ONE pos tag
  - → certain tags tend to co-occur regularly.