Parts of Speech

Nouns, verbs, adjectives, adverbs, prepositions, ...

- Efficiently assignable with high accuracy
- Valuable in many NLP tasks
- Commonly thought of as mapping to the real world
  - Objects
  - Properties
  - Actions
- In linguistics, understood via
  - Distributional properties (co-occurrences with other words)
  - Morphology, including the affixes they take, e.g., -tion, -ize
  - Intonational
- 45 POS tags defined in the Penn Treebank ($\approx$ 1993)
  - Includes variants such as tense and aspect
  - Includes punctuation

Closed versus Open Class

- Borrowings need special handling
  - a priori
  - schadenfreude
- Distinguishing mass nouns from count nouns
  - No plurals vs. plurals
  - Roughly, Real numbers vs. Natural numbers, but not quite (give examples)
- Syntactic substitutability
- Conjoinable, i.e., with *and*

Closed versus Open Class

- Closed class or function words
  - Change slowly
  - Prepositions, particles, determiners ($\sim$ articles), conjunctions, pronouns, auxiliary verbs, numerals
  - Lend structure to language
- Open classes
  - Nouns
  - Verbs
  - Adjectives
  - Adverbs

Penn Treebank Tagset

From the Wall Street Journal and Brown corpora
Dependency grammars (introduced later) have another tagset

Tag	Description	Example	Tag	Description	Example	Tag	Description	Example
CC	coordinating conjunction	*and, but, or*	PDT	predeterminer	*all, both*	VBP	verb non-3sg present	*eat*
CD	cardinal number	*one, two*	POS	possessive ending	*'s*	VBZ	verb 3sg pres	*eats*
DT	determiner	*a, the*	PRP	personal pronoun	*I, you, he*	WDT	wh-determ.	*which, that*
EX	existential 'there'	*there*	PRP$	possess. pronoun	*your, one's*	WP	wh-pronoun	*what, who*
FW	foreign word	*mea culpa*	RB	adverb	*quickly*	WP$	wh-possess.	*whose*
IN	preposition/ subordin-conj	*of, in, by*	RBR	comparative adverb	*faster*	WRB	wh-adverb	*how, where*
JJ	adjective	*yellow*	RBS	superlatv. adverb	*fastest*	$	dollar sign	$
JJR	comparative adj	*bigger*	RP	particle	*up, off*	#	pound sign	#
JJS	superlative adj	*wildest*	SYM	symbol	+,%, &	“	left quote	‘ or “
LS	list item marker	*1, 2, One*	TO	“to”	*to*	”	right quote	’ or ”
MD	modal	*can, should*	UH	interjection	*ah, oops*	(	left paren	[, (, {, <
NN	sing or mass noun	*llama*	VB	verb base form	*eat*	)	right paren	],), }, >
NNS	noun, plural	*llamas*	VBD	verb past tense	*ate*	,	comma	,
NNP	proper noun, sing.	*IBM*	VBG	verb gerund	*eating*	.	sent-end punc	. ! ?
NNPS	proper noun, plu.	*Carolinas*	VBN	verb past part.	*eaten*	:	sent-mid punc	: ; ... – -

Penn Treebank Markup Example

Large effort on developing NL resources: tagset, labeled datasets, . . .

- Preliminary findings were reported in today's New England Journal of Medicine.

Preliminary/adjective	findings/plural-noun
were/verb-past	reported/verb-part-participle
in/preposition	today/singular-noun
's/possessive	New/proper-noun-singular
England/proper-noun-singular	Journal/proper-noun-singular
of/preposition	Medicine/proper-noun-singular
./sentence-ending	

- Typical way of writing:
  Preliminary/JJ findings/NNS were/VBD reported/VBN in/IN today/NN 's/POS New/NNP England/NNP Journal/NNP of/IN Medicine/NNP ./.
- Does today possess the New England Journal of Medicine?
- Why isn't all of New England Journal of Medicine one noun?

Part of Speech Tagging Challenge

- Many words can take multiple tags depending on context
  - $\sim$ 14–15% of the words in the Wall Street Journal and Brown corpora

Adjective	earnings growth took a back/JJ seat
Mass noun	a small building in the back/NN
Verb present tense	a clear majority of senators back/VBP the bill
Verb	Dave began to back/VB toward the door
Particle	enable the country to buy back/RP about debt
Adverb	I was twenty-one back/RB then

- Simple baseline: most frequent class

Part of Speech Tagging as Sequence Tagging

- Markov (first-order): next state depends on current state but not history
  - Suffix closure
  - Fusion closure
  - Limit closure
- Hidden Markov Model (HMM)
  - States, Q: parts of speech or POS tag
  - Transition probability, A: one POS to the next
  - Observations, O: (sequence of) words from vocabulary
  - Observation likelihood, B: probability of word given POS
  - Initial probability distribution, π: of starting with a POS
- HMM assumptions
  - Probability of POS depends only on previous POS
  - Probability of word depends only on POS

HMM Part of Speech Tagging: Estimation

- Estimate transition probabilities via bigram model on corpus

$$P(q_i|q_{i-1}) = \frac{\text{count}(q_{i-1}, q_i)}{\text{count}(q_{i-1})}$$

  - Estimate ith part of speech given the previous part of speech
- Estimate emission probabilities via POS distributions on corpus

$$P(o_i|q_i) = \frac{\text{count}(q_i, o_i)}{\text{count}(q_i)}$$

HMM Part of Speech Tagging: Final Form

- We seek the most probable POS sequence ($Q = q_1^n$) for a given word sequence

$$\hat{Q} = \underset{Q}{\operatorname{argmax}} P(Q|O)$$

- Applying Bayes

$$\hat{Q} = \underset{Q}{\operatorname{argmax}} \frac{P(O|Q)P(Q)}{P(O)}$$

  - The observation sequence is given—thus, $P(O)$ is fixed, so drop it
  - Markov: each POS depends only on its predecessor
  - Each word depends only on the POS

- Combined model

$$\widehat{q_1^n} = \underset{q_1^n}{\operatorname{argmax}} \prod_i^n P(o_i|q_i)P(q_i|q_{i-1})$$

Viterbi Algorithm

Dynamic programming

- Like the minimum edit distance algorithm, but involves
  - Products of probabilities, not sums of edit costs
  - Maximum over a different set of paths
- To compute a Viterbi matrix, V
  - Each column: an observation (word)
  - Each row: a state (POS)
  - $V[s,t]$: (maximum) probability of being at POS s after seeing the first t words
  - Three probabilities: π, entry; A, transition; B, emission
- Initialize first column to product of probability of beginning from the respective state and the probability of emitting the first word from it

$$V_{s,1} = \pi_s B_s(o_1)$$

- Iteratively, compute

$$V_{s,t} = \max_{i}^{N} V_{[i,t-1]} A_i B_s(o_t)$$

Extensions

Languages like Turkish with complex morphology remain difficult

- Use trigrams instead of bigrams
  - Gain in accuracy $\sim$0.5%
  - Need to extend Viterbi to look at a history of two
- Usual need for addressing sparsity: smoothing and interpolation
- Beam search: limit search to beam width $\beta \ll N$
- Insert end of sentence marker to facilitate search
- Unknown words
  - Base POS probabilities on affixes, e.g., *-tion* indicates nouns, *-ize* verbs, *-ly* adverbs, and *-able* adjectives