INTRODUCTION TO PART-OF-SPEECH (POS) TAGGING

LU XIAO
LXIAO04@SYR.EDU
213 HINDS HALL

ADOPTED SOME MATERIALS DEVELOPED IN PREVIOUS COURSES BY NANCY MCCRACKEN, LIZ LIDDY AND OTHERS; AND SOME INSTRUCTOR RESOURCES FOR THE BOOK "SPEECH AND LANGUAGE PROCESSING" BY DANIEL JURAFSKY AND JAMES H. MARTIN

EMAIL POLICY

- If you can come to the office hours, please do so instead of emailing
- In the email, please mention whether you are in morning session or afternoon session
- Please try **not** to email TA/instructor outside of normal business hours; if you do, please note that we may only be able to reply in the next business time

HOMEWORK POLICY

- Please do not show us your homework and ask if it is correct or not – that evaluation is to be done in grading ☺

CATEGORIES OF KNOWLEDGE

- Phonology
- Morphology
- Syntax
- Semantics
- Pragmatics
- Discourse

POS tags are assigned to words, but may use adjacent words for information

What Is Part-of-speech Tagging?

- The general purpose of a part-of-speech tagger is to associate each word in a text with its correct lexical-syntactic category (represented by a tag)

03/14/1999 (AFP)… the extremist Harkatul Jihad group, reportedly backed by Saudi dissident Osama bin Laden …

… the|DT extremist|JJ Harkatul|NNP Jihad|NNP group|NN ,|, reportedly|RB backed|VBD by|IN Saudi|NNP dissident|NN Osama|NNP bin|NN Laden|NNP …

Called as: Parts-of-speech (POS), lexical categories, word classes, morphological classes, lexical tags...
Lots of debate within linguistics about the number, nature, and universality of these – AND we'll completely ignore this debate

POS Examples

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*

Why Is Part-of-speech Tagging Needed?

- Knowledge on the function of the word
- Support for the higher levels of NL Processing:
  - Phrase Bracketing (use regular expressions with POS tag matching)
  - Parsing
  - Semantics
- Applications that use POS tagging:
  - Speech synthesis - Text-to-speech (how do we pronounce “lead”?)
  - Information retrieval — selection of high-content words
  - Word-sense disambiguation
  - Sentiment detection — selection of high-opinion or emotion words

Example of POS Tag Use: Detecting Imperative Statements in Wikipedia Discussions

Examples:

- Please carefully study the General Notability Guideline (WP:GNG), which expects "significant coverage in reliable sources that are independent of the subject".(imperative)
- You must first discuss the matter there, and you need to be specific. (imperative)
- Instead of complaining, how about finding such content and improving the article? (indirect directive)

Imperative Detection Rules

1. A personal pronoun or noun (e.g., you, they, username) followed by a modal verb (e.g., should, must)
2. A verb (in its base form) as the root in the phrase structure and this particular verb has no subject child in the dependency structure.

HOW TO DO PART-OF-SPEECH (POS) TAGGING

Objective:

given sentence S = $ww0$, $ww1$, … $wwnn$,

determine tags T = $tt0$,$tt1$, … $ttnn$.

The	**grand**	**jury**	**Commented**	**on**	**a**	**number**	**of**	**other**	**topics**	**.**
DT	JJ	NN	VBD	IN	DT	NN	IN	JJ	NNS	.

Part-of-speech Tagging Is Hard

- Words may be ambiguous in different ways:
  - A word may have multiple meanings as the same part- of-speech
    - *file* – **noun**, a folder for storing papers
    - *file* – **noun**, instrument for smoothing rough edges
  - A word may function as multiple parts-of-speech
    - a *round* table: **adjective**
    - a *round* of applause: **noun**
    - to *round* out your interests: **verb**
    - to work the year *round*: **adverb**

How Hard Is POS Tagging? Measuring Ambiguity

		87-tag Original Brown		45-tag Treebank Brown	
Unambiguous (1 tag)		**44,019**		**38,857**	
Ambiguous (2–7 tags)		**5,490**		**8844**	
Details:	2 tags	4,967		6,731	
	3 tags	411		1621	
	4 tags	91		357	
	5 tags	17		90	
	6 tags	2	(*well*, *beat*)	32	
	7 tags	2	(*still*, *down*)	6	(*well*, *set*, *round*, *open*, *fit*, *down*)
	8 tags			4	(*'s*, *half*, *back*, *a*)
	9 tags			3	(*that*, *more*, *in*)

Overview Of Approaches

- Rule-based Approach
  - Simple and it doesn't require a tagged corpus, but not as accurate as other approaches
- Stochastic Approach
  - Refers to any approach which incorporates frequencies or probabilities
  - Requires a tagged corpus to learn frequencies
  - N-gram taggers
  - Hidden Markov Model (HMM) taggers
- Other Issues: unknown words and evaluation

POS Tagging: Choosing A Tagset

- There are so many parts of speech, potential distinctions we can draw
- To do POS tagging, we need to choose a standard set of tags to work with
- Could pick very coarse tagsets
  - N, V, Adj, Adv.
- More commonly used set is finer grained, the "**Penn TreeBank tagset**", 45 tags
  - PRP$, WRB, WP$, VBG
- Even more fine-grained tagsets exist, e.g., Kucera & Francis (Brown Corpus) – 87 POS tags – one of the first English POS tagging projects

Penn Treebank POS Tagset:

https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

Tag	Description	Example	Tag	Description	Example
CC	coordin. conjunction	*and, but, or*	SYM	symbol	+,%, &
CD	cardinal number	*one, two, three*	TO	“to”	*to*
DT	determiner	*a, the*	UH	interjection	*ah, oops*
EX	existential ‘there’	*there*	VB	verb, base form	*eat*
FW	foreign word	*mea culpa*	VBD	verb, past tense	*ate*
IN	preposition/sub-conj	*of, in, by*	VBG	verb, gerund	*eating*
JJ	adjective	*yellow*	VBN	verb, past participle	*eaten*
JJR	adj., comparative	*bigger*	VBP	verb, non-3sg pres	*eat*
JJS	adj., superlative	*wildest*	VBZ	verb, 3sg pres	*eats*
LS	list item marker	*1, 2, One*	WDT	wh-determiner	*which, that*
MD	modal	*can, should*	WP	wh-pronoun	*what, who*
NN	noun, sing. or mass	*llama*	WP$	possessive wh-	*whose*
NNS	noun, plural	*llamas*	WRB	wh-adverb	*how, where*
NNP	proper noun, singular	*IBM*	$	dollar sign	$
NNPS	proper noun, plural	*Carolinas*	#	pound sign	#
PDT	predeterminer	*all, both*	“	left quote	‘ or “
POS	possessive ending	*’s*	”	right quote	’ or ”
PRP	personal pronoun	*I, you, he*	(	left parenthesis	[, (, {, <
PRP$	possessive pronoun	*your, one’s*	)	right parenthesis	],), }, >
RB	adverb	*quickly, never*	,	comma	,
RBR	adverb, comparative	*faster*	.	sentence-final punc	. ! ?
RBS	adverb, superlative	*fastest*	:	mid-sentence punc	: ; ... – -
RP	particle	*up, off*			

EXAMPLES OF PENN TREEBANK TAGGING

- The/DT grand/JJ jury/NN commented/VBD on/IN a/DT number/NN of/IN other/JJ topics/NNS ./.
- Book/VB that/DT flight/NN ./.
- Does/VBZ that/DT flight/NN serve/VB dinner/NN ?/?

POS Tagging As Sequence Classification

- We are given a sentence (an "observation" or "sequence of observations")
  - *Secretariat is expected to race tomorrow*
- What is the best sequence of tags that corresponds to this sequence of observations?
- Probabilistic view:
  - Consider all possible sequences of tags
  - Out of this universe of sequences, choose the tag sequence which is most probable given the observation sequence of n words $w_1 \ldots w_n$.

ROAD TO HMMS

- We want, out of all sequences of n tags $t_1 \ldots t_n$ the single tag sequence such that $P(t_1 \ldots t_n | w_1 \ldots w_n)$ is highest.
  - $P(t_1 \ldots t_n | w_1 \ldots w_n)$: the probability of the tag sequence $t_1 \ldots t_n$ given the word sequence $w_1 \ldots w_n$

$$\hat{t}_1^n = \underset{t_1^n}{\operatorname{argmax}} P(t_1^n | w_1^n)$$

- Hat ^ means "our estimate of the best one"
- Argmax_x P(x) means "the x such that P(x) is maximized"
  - i.e. find the tag sequence that maximizes the probability

ROAD TO HMMS

- This equation is guaranteed to give us the best tag sequence

$$\hat{t}_1^n = \underset{t_1^n}{\operatorname{argmax}} P(t_1^n | w_1^n)$$

- But how to make it operational? How to compute this value?
- Intuition of Bayesian classification:
  - Use Bayes rule to transform into a set of other probabilities that are easier to compute

Thomas Bayes 1701 - 1761

USING BAYES RULE

- Bayes rule:

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)}$$

- Apply Bayes Rule

$$\hat{t}_1^n = \underset{t_1^n}{\operatorname{argmax}} \frac{P(w_1^n|t_1^n)P(t_1^n)}{P(w_1^n)}$$

- Note that this is using the conditional probability, given a tag sequence, what is the most likely word sequence with those tags.
  - Eliminate denominator as it is the same for every sequence

$$\hat{t}_1^n = \underset{t_1^n}{\operatorname{argmax}} P(w_1^n|t_1^n)P(t_1^n)$$

Likelihood and Prior

$$\hat{t}_1^n = \underset{t_1^n}{\operatorname{argmax}} \overbrace{P(w_1^n|t_1^n)}^{\text{likelihood}} \overbrace{P(t_1^n)}^{\text{prior}}$$

- Likelihood: assume that the probability of the word appearing depends only on its own POS tag

$$P(w_1^n|t_1^n) \approx \prod_{i=1}^{n} P(w_i|t_i)$$

- Prior: use the bigram assumption that the tag only depends on the previous tag

$$P(t_1^n) \approx \prod_{i=1}^{n} P(t_i|t_{i-1})$$

$$\hat{t}_1^n \approx \underset{t_1^n}{\operatorname{argmax}} \prod_{i=1}^{n} P(w_i|t_i)P(t_i|t_{i-1})$$

Two Sets Of Probabilities (1)

- Word likelihood probabilities $p(w_i | t_i)$
  - VBZ (third-person singular present verb): likely to be "is"
  - Compute P(is|VBZ) by counting in a labeled corpus:

$$P(w_i|t_i) = \frac{C(t_i, w_i)}{C(t_i)}$$

Count of is tagged with VBZ

$$P(is|VBZ) = \frac{C(VBZ, is)}{C(VBZ)} = \frac{10,073}{21,627} = .47$$

Two Sets Of Probabilities (2)

- Tag transition probabilities $p(t_i|t_{i-1})$ (priors)
  - Determiners likely to precede adjs and nouns
    - That/DT flight/NN
    - The/DT yellow/JJ hat/NN
    - So we expect P(NN|DT) and P(JJ|DT) to be high
  - Compute P(NN|DT) by counting in a labeled corpus:

$$P(t_i|t_{i-1}) = \frac{C(t_{i-1}, t_i)}{C(t_{i-1})}$$

Count of DT NN sequence

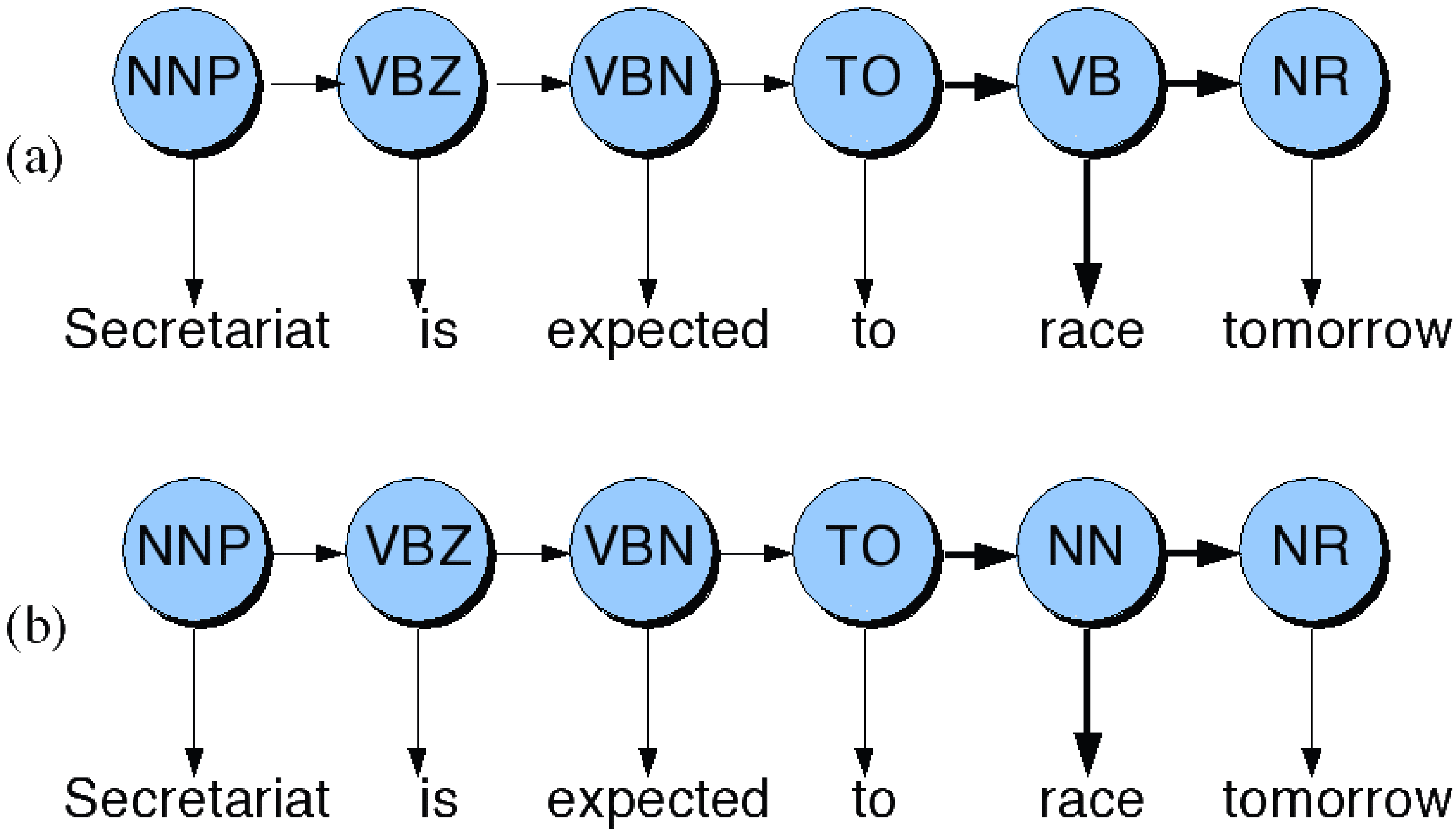
$$P(NN|DT) = \frac{C(DT,NN)}{C(DT)} = \frac{56,509}{116,454} = .49$$

An Example: The Word "Race"

- The word "race" can occur as a verb or as a noun:
  - Secretariat/NNP is/VBZ expected/VBN to/TO **race**/VB tomorrow/NR
  - People/NNS continue/VB to/TO inquire/VB the/DT reason/NN for/IN the/DT **race**/NN for/IN outer/JJ space/NN
- How do we pick the right tag?

Disambiguating "Race" (Using Brown Corpus Tagset)

Which tag sequence is most likely?



EXAMPLE

- *The equations only differ in "to race tomorrow"*
- P(NN|TO) = .00047
- P(VB|TO) = .83
- P(race|NN) = .00057
- P(race|VB) = .00012
- P(NR|VB) = .0027
- P(NR|NN) = .0012

The tag transition probabilities P(NN|TO) and P(VB|TO)

Lexical likelihoods from the Brown corpus for 'race' given a POS tag NN or VB.

Tag sequence probability for the likelihood of an adverb occurring given the previous tag verb or noun

- P(race|VB)P(VB|TO)P(NR|VB) = .00000027
- P(race|NN)P(NN|TO)P(NR|NN)=.00000000032
- *So we (correctly) choose the verb tag.*

Hidden Markov Model (HMM) Tagger

- An HMM POS tagger is a sequential classifier: - uses the previous sequence of tags as information:
  - Step 1, computes the word likelihood probabilities and tag transition probabilities for each tag from a (training) corpus
  - Step 2, for each sentence that we want to tag, it uses the Viterbi algorithm to find the path of the best sequence of tags to fit that sentence:
    - In order from left to right, use information from previous tags (tag prior probabilities) and word (word likelihood probabilities) to predict the next tag in the sequence

word_{n-1} …	word_{-2}	word_{-1}	word
$\text{tag}_{n-1..}$	tag_{-2}	tag_{-1}	XX

(Optional): The **Viterbi algorithm** is a dynamic programming **algorithm** for finding the most likely sequence of hidden states – called the **Viterbi** path – that results in a sequence of observed events, especially in the context of Markov information sources and hidden Markov models.

Evaluation: Is Our POS Tagger Any Good?

- Answer: we use a manually tagged corpus, which we will call the "Gold Standard"
  - We run our POS tagger on the gold standard and compare its predicted tags with the gold tags
  - We compute the accuracy (and other evaluation measures)
- Important: 100% is impossible even for human annotators.
  - We estimate humans can do POS tagging at about 98% accuracy.
  - Some tagging decisions are very subtle and hard to do:
    - 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
  - The "Gold Standard" will have human mistakes; humans are subject to fatigue, etc.

TREEBANK

- "a treebank is a parsed text corpus that annotates syntactic or semantic sentence structure."
  - from Wikipedia

- Penn TreeBank: the first large-scale treebank
  - https://web.archive.org/web/19970614160127/http://www.cis.upenn.edu:80/~treebank/

PENN TREEBANK – TREEBANK-2

- https://catalog.ldc.upenn.edu/LDC95T7
- A corpus containing:
  - over 1.6 million words of hand-parsed material from the Dow Jones News Service, plus an additional 1 million words tagged for part-of-speech.
  - the first fully parsed version of the Brown Corpus, which has also been completely retagged using the Penn Treebank tag set.
  - source code for several software packages which permits the user to search for specific constituents in tree structures.
- Not free

PENN TREEBANK – TREEBANK-3

- https://catalog.ldc.upenn.edu/LDC99T42

- A corpus containing:
  - One million words of 1989 Wall Street Journal material annotated in Treebank II style.
  - A small sample of ATIS-3 material annotated in Treebank II style.
  - A fully tagged version of the Brown Corpus.

- Not free

How Can We Improve Our Tagger?

- What are the main sources of information for our HMM POS tagger?
  - Knowledge of tags of neighboring words
  - Knowledge of word tag probabilities
    - *man* is rarely used as a verb….

- Unknown words (words not occurring in the training corpus) can be a problem because we don't have this information

- And we are not including information about the features of the words

Feature-based Classifiers

- A feature-based classifier is an algorithm that will take a word and assign a POS tag based on features of the word in its context in the sentence (typically feature classifier uses information from 1-3 surrounding words on either side)

$$\text{word}_{-2}\ \text{word}_{-1}\ \text{word}\ \ \text{word}_{+1}\ \ \text{word}_{+2}$$

XX

- Many algorithms are used for these traditional classifiers, just to name a few
  - Naïve Bayes
  - Maximum Entropy (MaxEnt)
  - Support Vector Machines (SVM)

Features Of Words

- Can do quite well just looking at a word by itself:
  - Word — the: the → DT (determiner)
  - Lowercased word — Importantly: importantly → RB (adverb)
  - Prefixes — unfathomable: un- → JJ (adjective)
  - Suffixes — Importantly: -ly → RB
tangential: -al → JJ
  - Capitalization — Meridian: CAP → NNP (proper noun)
  - Word shapes — 35-year: d-x → JJ
- These properties can include information about the previous or the next word(s)
  - The word *be* appears to the left of the word pretty → JJ
- No information about tags of the previous or next words, unlike HMM

Development Process For Features

- The tagged data should be separated into a training set and a test set.
  - The tagger is trained on the training set and evaluated on the test set
    - May also hold out some data for development
  - Evaluation numbers are not prejudiced by the training set

- If our feature-based tagger has errors, then we improve the features.
  - Suppose we incorrectly tag *as* as IN in the phrase *as soon as*, when it should be RB:

PRP	VBD	IN	RB	IN	PRP	VBD	.
They	left	as	soon	as	he	arrived	.

  - We could fix this with a feature that include the next word.

Overview Of POS Tagger Accuracies

- Rough accuracies (from Chris Manning): all words / unknown words
  - Most freq tag: ~90% / ~50%

Most errors on unknown words

  - Trigram HMM: ~95% / ~55%
    - HMM with trigrams
  - Maxent P(t|w): 93.7% / 82.6%
    - Feature based tagger
  - MEMM tagger: 96.9% / 86.9%
    - Combines feature based and HMM tagger
  - Bidirectional dependencies: 97.2% / 90.0%
  - Upper bound: ~98% (human agreement)

POS Taggers With Online Demos

- Many pages list downloadable taggers (and other resources) such as this page from the Stanford NLP group and George Dillon at U Washington
  - http://nlp.stanford.edu/software/tagger.shtml
  - http://faculty.washington.edu/dillon/GramResources/
- There are not too many on-line taggers available for demos, but here are some possibilities:
  - The Stanford online parser demo includes POS tags: http://nlp.stanford.edu:8080/parser/ http://nlp.stanford.edu:8080/corenlp/
  - Illinois (UIUC) tagger demo from the Cognitive Computation Group
  - http://cogcomp.cs.illinois.edu/demo/pos/?id=4

Stanford NLP Demo

Stanford CoreNLP

Output format: Visualise

Please enter your text here:

Helicopters will patrol the temporary no-fly zone around New Jersey's MetLife Stadium Sunday, with F-16s based in Atlantic City ready to be scrambled if an unauthorized aircraft does enter the restricted airspace.

Submit Clear

Part-of-Speech:

1 Helicopters/NNPS will/MD patrol/NN the/DT temporary/JJ no-fly/JJ zone/NN around/IN New/NNP Jersey/NNP 's/POS MetLife/NNP Stadium/NNP Sunday/NNP ,/, with/IN F-16s/NNS based/VBN in/IN Atlantic/NNP City/NNP ready/JJ to/TO be/VB scrambled/VBN if/IN an/DT unauthorized/JJ aircraft/NN does/VBZ enter/VB the/DT restricted/JJ airspace/NN ./.

Named Entity Recognition:

1 Helicopters will patrol the temporary no-fly zone around New Jersey[Location]'s MetLife Stadium[Org] Sunday[Date], with F-16s based in Atlantic City[Location] ready to be scrambled if an unauthorized aircraft does enter the restricted

UIUC Demo

COGNITIVE COMPUTATION GROUP
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

News | Research | People | Software | Demos | Publications | Resources | Schedule

Demo

Part of Speech Tagging Demo

About This Demo

737,457 views

If you wish to cite this work, please cite this publication.

Helicopters will patrol the temporary no-fly zone around New Jersey's MetLife Stadium Sunday, with F-16s based in Atlantic City ready to be scrambled if an unauthorized aircraft does enter the restricted airspace.

Down below, bomb-sniffing dogs will patrol the trains and buses that are expected to take approximately 30,000 of the 80,000-plus spectators to Sunday's Super Bowl between the Denver Broncos and Seattle Seahawks.

The Transportation Security Administration said it has added about two dozen dogs to monitor passengers coming

Submit

The Part-of-Speech tagger has automatically labeled the input in the following way.

NNPS/ Helicopters MD/ will NN/ patrol DT/ the JJ/ temporary JJ/ no-fly NN/ zone IN/ around NNP/ New NNP/ Jersey POS/ 's NNP/ MetLife NNP/ Stadium NNP/ Sunday ,/ , IN/ with NNP/ F-16s VBN/ based IN/ in NNP/ Atlantic NNP/ City JJ/ ready TO/ to VBN/ scrambled IN/ if DT/ an JJ/ unauthorized NN/ aircraft VBZ/ does VB/ enter DT/ the VBN/ restricted ce ./ .

Problems with our new website?

Conclusions

- Part of Speech tagging is a doable task with high performance results
  - In addition to the standard text POS taggers discussed here, there are now POS tag systems and taggers developed for social media text.

- Contributes to many practical, real-world NLP applications and is now used as a pre-processing module in most systems

- Computational techniques learned at this level can be applied to NLP tasks at higher levels of language processing

Open and Closed Classes

- Closed class: a small fixed membership
  - Prepositions: of, in, by, …
  - Auxiliaries: may, can, will had, been, …
  - Pronouns: I, you, she, mine, his, them, …
  - Usually function words (short common words which play a role in grammar)
- Open class: new ones can be created all the time
  - English has 4: Nouns, Verbs, Adjectives, Adverbs

Open Class Words

- Nouns
  - Proper nouns (Boulder, Granby, Eli Manning)
    - English capitalizes these.
  - Common nouns (the rest).
  - Count nouns and mass nouns
    - Count: have plurals, get counted: goat/goats, one goat, two goats
    - Mass: don't get counted (snow, salt, communism) (*two snows)
- Adverbs: tend to modify things
  - Unfortunately, John walked home extremely slowly yesterday
  - Directional/locative adverbs (*here,home, downhill*)
  - Degree adverbs (*extremely, very, somewhat*)
  - Manner adverbs (*slowly, slinkily, delicately*)
- Verbs
  - In English, have morphological affixes (eat/eats/eaten)

Closed Class Words

- Closed classes– words are not added to these classes:
  - determiners: a, an, the
  - pronouns: she, he, I
  - prepositions: on, under, over, near, by, …
    - over the river and through the woods
  - particles: up, down, on, off, …
    - Used with verbs and have slightly different meaning than when used as a preposition
      - she turned the paper over
- Closed class words are often function words which have structuring uses in grammar:
  - of, it , and , you
- Differ more from language to language than open class words

Open And Closed Classes

Open class (lexical) words

- Nouns
  - Proper: *IBM*, *Italy*
  - Common: *cat / cats*, *snow*
- Verbs
  - Main: *see*, *registered*
- Adjectives *old older oldest*
- Adverbs *slowly*
- Numbers *122,312* *one*
- *… more*

Closed class (functional)

- Determiners *the some*
- Conjunctions *and or*
- Pronouns *he its*
- Modals *can* *had*
- Prepositions *to with*
- Particles *off up*
- Interjections *Ow Eh*
- *… more*

Prepositions From CELEX

- Prepositions show relationships between other words
- *Charts show words from the CELEX on-line dictionary with frequencies from the COBUILD corpus*

of	540,085	through	14,964	worth	1,563	pace	12
in	331,235	after	13,670	toward	1,390	nigh	9
for	142,421	between	13,275	plus	750	re	4
to	125,691	under	9,525	till	686	mid	3
with	124,965	per	6,515	amongst	525	o'er	2
on	109,129	among	5,090	via	351	but	0
at	100,169	within	5,030	amid	222	ere	0
by	77,794	towards	4,700	underneath	164	less	0
from	74,843	above	3,056	versus	113	midst	0
about	38,428	near	2,026	amidst	67	o'	0
than	20,210	off	1,695	sans	20	thru	0
over	18,071	past	1,575	circa	14	vice	0

Charts from Jurafsky and Martin text

English Single-word Particles

- Definition of the term "particle" in linguistics **<u>varies</u>**
- Primarily words that used to provide shades of meaning to other words, particularly verbs (e.g., *put* ***away, go*** ***aboard***)

aboard	aside	besides	forward(s)	opposite	through
about	astray	between	home	out	throughout
above	away	beyond	in	outside	together
across	back	by	inside	over	under
ahead	before	close	instead	overhead	underneath
alongside	behind	down	near	past	up
apart	below	east, etc.	off	round	within
around	beneath	eastward(s),etc.	on	since	without

ADVERB PARTICLES AND PREPOSITIONS

- "While the particle is closely tied to its verb to form idiomatic expressions, the preposition is closely tied to the noun or pronoun it modifies." – From the web site: https://www.englishgrammar.org/adverb-particles-prepositions/

Pronouns In CELEX

- Personal
  - he, ours
- Demonstrative
  - that, those
- Reflexive
  - myself, ourselves
- Indefinite
  - one, neither, somebody, both

it	199,920	how	13,137	yourself	2,437	no one	106
I	198,139	another	12,551	why	2,220	wherein	58
he	158,366	where	11,857	little	2,089	double	39
you	128,688	same	11,841	none	1,992	thine	30
his	99,820	something	11,754	nobody	1,684	summat	22
they	88,416	each	11,320	further	1,666	suchlike	18
this	84,927	both	10,930	everybody	1,474	fewest	15
that	82,603	last	10,816	ourselves	1,428	thyself	14
she	73,966	every	9,788	mine	1,426	whomever	11
her	69,004	himself	9,113	somebody	1,322	whosoever	10
we	64,846	nothing	9,026	former	1,177	whomsoever	8
all	61,767	when	8,336	past	984	wherefore	6
which	61,399	one	7,423	plenty	940	whereat	5
their	51,922	much	7,237	either	848	whatsoever	4
what	50,116	anything	6,937	yours	826	whereon	2
my	46,791	next	6,047	neither	618	whoso	2
him	45,024	themselves	5,990	fewer	536	aught	1
me	43,071	most	5,115	hers	482	howsoever	1
who	42,881	itself	5,032	ours	458	thrice	1
them	42,099	myself	4,819	whoever	391	wheresoever	1
no	33,458	everything	4,662	least	386	you-all	1
some	32,863	several	4,306	twice	382	additional	0
other	29,391	less	4,278	theirs	303	anybody	0
your	28,923	herself	4,016	wherever	289	each other	0
its	27,783	whose	4,005	oneself	239	once	0
our	23,029	someone	3,755	thou	229	one another	0
these	22,697	certain	3,345	'un	227	overmuch	0
any	22,666	anyone	3,318	ye	192	such and such	0
more	21,873	whom	3,229	thy	191	whate'er	0
many	17,343	enough	3,197	whereby	176	whenever	0
such	16,880	half	3,065	thee	166	whereof	0
those	15,819	few	2,933	yourselves	148	whereto	0
own	15,741	everyone	2,812	latter	142	whereunto	0
us	15,724	whatever	2,571	whichever	121	whic[illegible]	0

Conjunctions

- Links words and phrases and gives relationship between them

and	514,946	yet	5,040	considering	174	forasmuch as	0
that	134,773	since	4,843	lest	131	however	0
but	96,889	where	3,952	albeit	104	immediately	0
or	76,563	nor	3,078	providing	96	in as far as	0
as	54,608	once	2,826	whereupon	85	in so far as	0
if	53,917	unless	2,205	seeing	63	inasmuch as	0
when	37,975	why	1,333	directly	26	insomuch as	0
because	23,626	now	1,290	ere	12	insomuch that	0
so	12,933	neither	1,120	notwithstanding	3	like	0
before	10,720	whenever	913	according as	0	neither nor	0
though	10,329	whereas	867	as if	0	now that	0
than	9,511	except	864	as long as	0	only	0
while	8,144	till	686	as though	0	provided that	0
after	7,042	provided	594	both and	0	providing that	0
whether	5,978	whilst	351	but that	0	seeing as	0
for	5,935	suppose	281	but then	0	seeing as how	0
although	5,424	cos	188	but then again	0	seeing that	0
until	5,072	supposing	185	either or	0	without	0

Auxiliary Verbs

- Auxiliary, or helping verbs, are used with main verbs to express time or mood
  - **Modal verbs** are the auxiliary verbs that express likelihood or ability
    - Can, might, must, could, should, …

can	70,930	might	5,580	shouldn't	858
will	69,206	couldn't	4,265	mustn't	332
may	25,802	shall	4,118	'll	175
would	18,448	wouldn't	3,548	needn't	148
should	17,760	won't	3,100	mightn't	68
must	16,520	'd	2,299	oughtn't	44
need	9,955	ought	1,845	mayn't	3
can't	6,375	will	862	dare	??
have	???				