NLP for Social Media: POS Tagging, Sentiment Analysis

Pawan Goyal

CSE, IITKGP

August 05, 2016

Part-of-Speech (POS) Tagging

Cant	MD
wait	VB
for	IN
the	DT
ravens	NNP
game	NN
tomorrow	NN
...	:
go	VB
ray	NNP
rice	NNP
!!!!!!!	.

Penn Treebank POS Tags

1.	CC	Coordinating conjunction	25.	TO	*to*
2.	CD	Cardinal number	26.	UH	Interjection
3.	DT	Determiner	27.	VB	Verb, base form
4.	EX	Existential *there*	28.	VBD	Verb, past tense
5.	FW	Foreign word	29.	VBG	Verb, gerund/present participle
6.	IN	Preposition/subordinating conjunction	30.	VBN	Verb, past participle
7.	JJ	Adjective	31.	VBP	Verb, non-3rd ps. sing. present
8.	JJR	Adjective, comparative	32.	VBZ	Verb, 3rd ps. sing. present
9.	JJS	Adjective, superlative	33.	WDT	*wh*-determiner
10.	LS	List item marker	34.	WP	*wh*-pronoun
11.	MD	Modal	35.	WP$	Possessive *wh*-pronoun
12.	NN	Noun, singular or mass	36.	WRB	*wh*-adverb
13.	NNS	Noun, plural	37.	#	Pound sign
14.	NNP	Proper noun, singular	38.	$	Dollar sign
15.	NNPS	Proper noun, plural	39.	.	Sentence-final punctuation
16.	PDT	Predeterminer	40.	,	Comma
17.	POS	Possessive ending	41.	:	Colon, semi-colon
18.	PRP	Personal pronoun	42.	(	Left bracket character
19.	PP$	Possessive pronoun	43.	)	Right bracket character
20.	RB	Adverb	44.	"	Straight double quote
21.	RBR	Adverb, comparative	45.	‘	Left open single quote
22.	RBS	Adverb, superlative	46.	“	Left open double quote
23.	RP	Particle	47.	’	Right close single quote
24.	SYM	Symbol (mathematical or scientific)	48.	”	Right close double quote

Part-of-Speech (POS) Tagging

Words often have more than one POS

The back door = JJ

On my back = NN

Win the voters back = RB

Promised to back the bill = VB

Part-of-Speech (POS) Tagging

Words often have more than one POS

The <u>back</u> door = JJ

On my <u>back</u> = NN

Win the voters <u>back</u> = RB

Promised to <u>back</u> the bill = VB

POS tagging problem is to determine the POS tag for a particular instance of a word.

Relevant papers and tool

- Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. *Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments*. In Proceedings of ACL 2011.
- Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider and Noah A. Smith. *Improved Part-of-Speech Tagging for Online Conversational Text with Word Clusters.* In Proceedings of NAACL 2013.
- Parts-of-Speech tagget for twitter - http://www.cs.cmu.edu/~ark/TweetNLP/

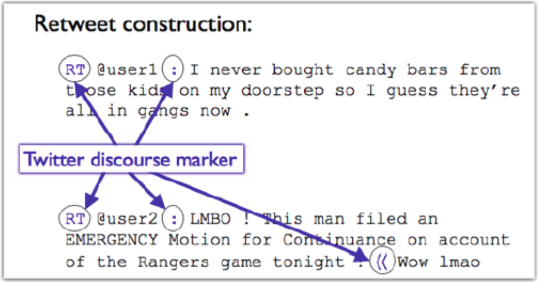
(a) @Gunservatively$_{@}$ obozo$_{\wedge}$ will$_{V}$ go$_{V}$ nuts$_{A}$ when$_{R}$ PA$_{\wedge}$ elects$_{V}$ a$_{D}$ Republican$_{A}$ Governor$_{N}$ next$_{P}$ Tue$_{\wedge}$.$_{,}$ Can$_{V}$ you$_{O}$ say$_{V}$ redistricting$_{V}$?$_{,}$

(b) Spending$_{V}$ the$_{D}$ day$_{N}$ withhh$_{P}$ mommma$_{N}$!$_{,}$

(c) lmao$_{!}$ ...$_{,}$ s/o$_{V}$ to$_{P}$ the$_{D}$ cool$_{A}$ ass$_{N}$ asian$_{A}$ officer$_{N}$ 4$_{P}$ #1$_{\$}$ not$_{R}$ runnin$_{V}$ my$_{D}$ license$_{N}$ and$_{\&}$ #2$_{\$}$ not$_{R}$ takin$_{V}$ dru$_{N}$ boo$_{N}$ to$_{P}$ jail$_{N}$.$_{,}$ Thank$_{V}$ u$_{O}$ God$_{\wedge}$.$_{,}$ #amen$_{\#}$

Twitter-specific Tags

- #hashtag
- @metion
- url
- email address
- emoticon
- discourse marker
- symbols
- ...



The case of Hashtags

Hashtags and at-mentions can also serve as words or phrases within a tweet; e.g. Is #qadaffi going down?. When used in this way, we tag hashtags with their appropriate part of speech, i.e., as if they did not start with #. Of the 418 hashtags in our data, 148 (35%) were given a tag other than **#**: 14% are proper nouns, 9% are common nouns, 5% are multi-word expressions (tagged as **G**), 3% are verbs, and 4% are something else. We do not apply this procedure to at-mentions, as they are nearly always proper nouns.

Combined forms

than for Standard English text. For example, apostrophes are often omitted, and there are frequently words like ima (short for *I'm gonna*) that cut across traditional POS categories. Therefore, we opted not to split contractions or possessives, as is common in English corpus preprocessing; rather, we introduced four new tags for combined forms: {nominal, proper noun} $\times$ {verb, possessive}.[5]

Tagging Scheme

Tag	Description	Examples	%
Nominal, Nominal + Verbal			
N	common noun (NN, NNS)	books someone	13.7
O	pronoun (personal/WH; not possessive; PRP, WP)	it you u meeee	6.8
S	nominal + possessive	books' someone's	0.1
^	proper noun (NNP, NNPS)	lebron usa iPad	6.4
Z	proper noun + possessive	America's	0.2
L	nominal + verbal	he's book'll iono (= *I don't know*)	1.6
M	proper noun + verbal	Mark'll	0.0
Other open-class words			
V	verb incl. copula, auxiliaries (V*, MD)	might gonna ought couldn't is eats	15.1
A	adjective (J*)	good fav lil	5.1
R	adverb (R*, WRB)	2 (i.e., *too*)	4.6
!	interjection (UH)	lol haha FTW yea right	2.6

Tagging Scheme

Other closed-class words			
D	determiner (WDT, DT, WP$, PRP$)	the teh its it's	6.5
P	pre- or postposition, or subordinating conjunction (IN, TO)	while to for 2 (i.e., *to*) 4 (i.e., *for*)	8.7
&	coordinating conjunction (CC)	and n & + BUT	1.7
T	verb particle (RP)	out off Up UP	0.6
X	existential *there*, predeterminers (EX, PDT)	both	0.1
Y	**X** + verbal	there's all's	0.0

Tagging Scheme

Twitter/online-specific			
#	hashtag (indicates topic/category for tweet)	#acl	1.0
@	at-mention (indicates another user as a recipient of a tweet)	@BarackObama	4.9
~	discourse marker, indications of continuation of a message across multiple tweets	RT and : in retweet construction RT @user : hello	3.4
U	URL or email address	http://bit.ly/xyz	1.6
E	emoticon	:-) :b (: <3 o_O	1.0
Miscellaneous			
$	numeral (CD)	2010 four 9:30	1.5
,	punctuation (#, $, '', (,), ,, ., :, ``)	!!! .... ?!?	11.6
G	other abbreviations, foreign words, possessive endings, symbols, garbage (FW, POS, SYM, LS)	ily (*I love you*) wby (*what about you*) 's ♫ --> awesome...I'm	1.1

Building the POS tagger

CRF model was used. Some insighful features:

- **Twitter orthography:** Features for several regular expression-style rules that detect at-mentions, hashtags, URLs etc.

Building the POS tagger

CRF model was used. Some insighful features:

- **Twitter orthography:** Features for several regular expression-style rules that detect at-mentions, hashtags, URLs etc.
- **Frequently-capitalized tokens:** Compiled gazetteers of tokens which are frequently capitalized. Features would be - membership in the top N frequently capitalized items.

Building the POS tagger

CRF model was used. Some insighful features:

- **Twitter orthography:** Features for several regular expression-style rules that detect at-mentions, hashtags, URLs etc.
- **Frequently-capitalized tokens:** Compiled gazetteers of tokens which are frequently capitalized. Features would be - membership in the top N frequently capitalized items.
- **Traditional tag dictionary:** Features for all coarse-grained tags that each word occurs with in the Penn Treebank.

Building the POS tagger

CRF model was used. Some insighful features:

- **Twitter orthography:** Features for several regular expression-style rules that detect at-mentions, hashtags, URLs etc.
- **Frequently-capitalized tokens:** Compiled gazetteers of tokens which are frequently capitalized. Features would be - membership in the top N frequently capitalized items.
- **Traditional tag dictionary:** Features for all coarse-grained tags that each word occurs with in the Penn Treebank.
- **Phonetic normalization:** Metaphone algorithm to create a coarse phonetic normalization of words to simpler keys.

Building the POS tagger

CRF model was used. Some insighful features:

- **Twitter orthography:** Features for several regular expression-style rules that detect at-mentions, hashtags, URLs etc.
- **Frequently-capitalized tokens:** Compiled gazetteers of tokens which are frequently capitalized. Features would be - membership in the top N frequently capitalized items.
- **Traditional tag dictionary:** Features for all coarse-grained tags that each word occurs with in the Penn Treebank.
- **Phonetic normalization:** Metaphone algorithm to create a coarse phonetic normalization of words to simpler keys.
- **Distributional similarity:** Distributional features from the successor and predecessor probabilities for the 10,000 most common terms.

Related Problems

Entity Recognition

- **Named Entity:** Names of people, places, organization
- Date and time

Related Problems

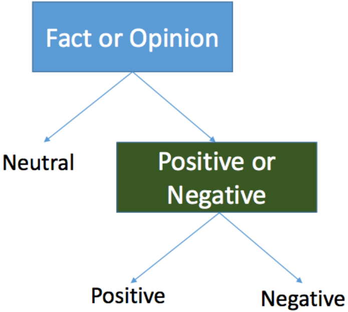
Entity Recognition

- **Named Entity:** Names of people, places, organization
- Date and time

Can you model it as a sequence labeling problem?

Sentiment Analysis

- A three way classification:
  - Positive
  - Neutral
  - Negative



Sentiment Analysis: Relevant Paper

Pak, Alexander, and Patrick Paroubek. “Twitter as a Corpus for Sentiment Analysis and Opinion Mining.” LREC. Vol. 10. 2010.

Sentiment Analysis: Examples

funkeybrewster: @redeyechicago I think Obama's visit might've sealed the victory for Chicago. Hopefully the games mean good things for the city.
vcurve: I like how Google celebrates little things like this: Google.co.jp honors Confucius Birthday — Japan Probe
mattfellows: Hai world. I hate faulty hardware on remote systems where politics prevents you from moving software to less faulty systems.
brroooklyn: I love the sound my iPod makes when I shake to shuffle it. Boo bee boo
MeganWilloughby: Such a Disney buff. Just found out about the new Alice in Wonderland movie. Official trailer: http://bit.ly/131Js0 I love the Cheshire Cat.

Data Creation

How to do that without manual labeling?

Data Creation

How to do that without manual labeling?

- Happy emoticons: ":-)", ":)", "=)", ":D" etc.
- Sad emoticons: ":-(", ":(", "=(", ";(" etc.

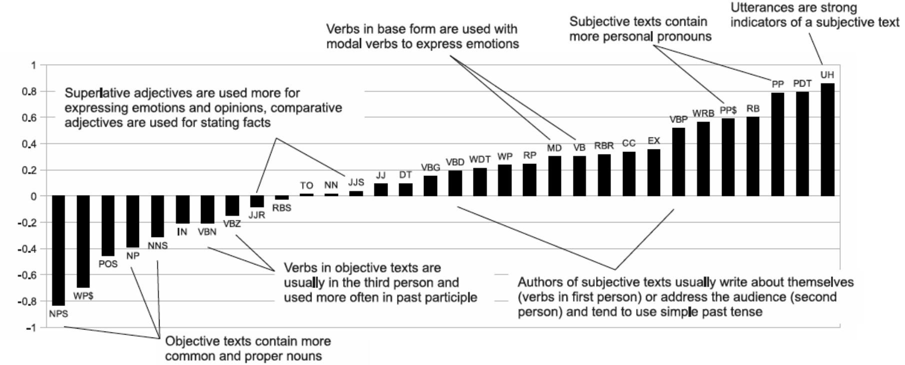
Data Creation

How to do that without manual labeling?

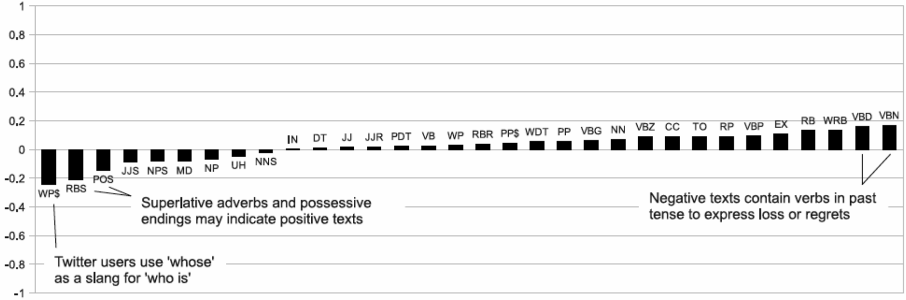
- Happy emoticons: ":-)", ":)", "=)", ":D" etc.
- Sad emoticons: ":-(", ":(", "=(", ";(" etc.

In order to collect a corpus of objective posts, we retrieved text messages from Twitter accounts of popular newspapers and magazines , such as "New York Times", "Washington Posts" etc. We queried accounts of 44 newspapers to collect a training set of objective texts.

POS tag Distribution: Subjective vs. Objective



POS tag Distribution: Positive vs. Negative



Classification Model

Naïve Bayes Model: Main features

POS tags, Word n-grams

Classification Model

Naïve Bayes Model: Main features

POS tags, Word n-grams

Using negations in Word n-grams

Constructing n-grams – we make a set of n-grams out of consecutive words. A negation (such as "no" and "not") is attached to a word which precedes it or follows it. For example, a sentence "I do not like fish" will form two bigrams: "I do+not", "do+not like", "not+like fish". Such a procedure allows to improve the accuracy of the classification since the negation plays a special role in an opinion and sentiment expression(Wilson et al., 2005).

Disregarding common n-grams

Using entropy and salience

Disregarding common n-grams

Using entropy and salience

$$entropy(g) = H(p(S|g)) = -\sum_{i=1}^{N} p(S_i|g) \log p(S_i|g)$$

Disregarding common n-grams

Using entropy and salience

$$entropy(g) = H(p(S|g)) = -\sum_{i=1}^{N} p(S_i|g) \log p(S_i|g)$$

$$salience(g) = \frac{1}{N} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} 1 - \frac{\min(P(g|s_i), P(g|s_j))}{\max(P(g|s_i), P(g|s_j))}$$

N-grams with high salience and low entropy

N-gram	Salience
so sad	0.975
miss my	0.972
so sorry	0.962
love your	0.961
i'm sorry	0.96
sad i	0.959
i hate	0.959
lost my	0.959
have great	0.958
i miss	0.957
gonna miss	0.956
wishing i	0.955
miss him	0.954
can't sleep	0.954

N-gram	Entropy
clean me	0.082
page news	0.108
charged in	0.116
so sad	0.12
police say	0.127
man charged	0.138
vital signs	0.142
arrested in	0.144
boulder county	0.156
most viewed	0.158
officials say	0.168
man accused	0.178
pleads guilty	0.18
guilty to	0.181