Improved Part-of-Speech Tagging for Online Conversational Text with Word Clusters

Olutobi Owoputi* Brendan O'Connor* Chris Dyer* Kevin Gimpel+ Nathan Schneider* Noah A. Smith*

*School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA

+Toyota Technological Institute at Chicago, Chicago, IL 60637, USA

We approach **part-of-speech tagging for informal, online conversational text** using large-scale unsupervised word clustering and new lexical features. Our system achieves state-of-the-art tagging results on both Twitter and IRC data. Additionally, we contribute the first POS annotation guidelines for such text and release a new dataset of English language tweets annotated using these guidelines.

Model

Discriminative sequence model (MEMM) with L1/L2 regularization

Tagger Features

- Hierarchical word clusters via Brown clustering (Brown et al., 1992) on a sample of 56M tweets
- Surrounding words/clusters
- Current and previous tags
- Tag dict. constructed from WSJ, Brown corpora
- Tag dict. entries projected to Metaphone encodings
- Name lists from Freebase, Moby Words, Names Corpus
- Emoticon, hashtag, @mention, URL patterns

Tagset

N	common noun
O	pronoun (personal/WH; not possessive)
^	proper noun
S	nominal + possessive
Z	proper noun + possessive
V	verb including copula, auxiliaries
L	nominal + verbal (e.g. *i'm*), verbal + nominal (*let's*)
M	proper noun + verbal
A	adjective
R	adverb
!	interjection
D	determiner
P	pre- or postposition, or subordinating conjunction
&	coordinating conjunction
T	verb particle
X	existential *there*, predeterminers
Y	X + verbal
#	hashtag (indicates topic/category for tweet)
@	at-mention (indicates a user as a recipient of a tweet)
~	discourse marker, indications of continuation across multiple tweets
U	URL or email address
E	emoticon
$	numeral
,	punctuation
G	other abbreviations, foreign words, possessive endings, symbols, garbage

Examples

Boutta	Shake	Da	Croud	So	Yall	Culd	Start	Hateing	Now
P	V	D	N	P	O	V	V	V	R

ikr	smh	he	asked	fir	yo	last	name	so	he	can	add	u	on	fb	lololol
!	G	O	V	P	D	A	N	P	O	V	V	O	P	^	!

Word Clusters

	Binary path	Top words (by frequency)
A1	111010100010	lmao lmfao lmaoo lmaooo hahahahaha lool ctfu rofl loool lmfaoo lmfaooo lmaoooo lmbo **lololol**
A2	111010100011	haha hahaha hehe hahahaha hahah aha hehehe ahaha hah hahahah kk hahaa ahah
A3	111010100100	yes yep yup nope yess yesss yessss ofcourse yeap likewise yepp yesh yw yuup yus
A4	111010100101	yeah yea nah naw yeahh nooo yeh noo noooo yeaa **ikr** nvm yeahhh nahh nooooo
A5	11101011011100	**smh** jk #fail #random #fact smfh #smh #winning #realtalk smdh #dead #justsaying
B	011101011	**u** yu yuh yhu uu yuu yew y0u yuhh youh yhuu iget yoy yooh yuo ꔹ yue juu Ʊ dya youz yyou
C	11100101111001	w fo fa fr fro ov fer **fir** whit abou aft serie fore fah fuh w/her w/that fron isn agains
D	111101011000	facebook **fb** itunes myspace skype ebay tumblr bbm flickr aim msn netflix pandora
E1	0011001	tryna gon finna bouta trynna **boutta** gne fina gonn tryina fenna qone trynaa qon
E2	0011000	gonna gunna gona gna guna gnna ganna qonna gonnna gana qunna gonne goona
F	0110110111	soo sooo soooo sooooo soooooo sooooooo soooooooo sooooooooo soooooooooo
G1	11101011001010	;) :p :-) xd ;-) ;d (; :3 ;p =p :-p =)) ;] xdd #gno xddd >:) ;-p >:d 8-) ;-d
G2	11101011001011	:) (: =) :)) :] ☺ :') =] ^_^ :))) ^.^ [: ;)) 😊 ((: ^__^ (= ^-^ :))))
G3	1110101100111	:(:/ -_- -.- :-(:'(d: :| :s -__- =(=/ >.< -___- :-/ </3 :\ -____- ;(/: :((>_< =[:[#fml
G4	111010110001	<3 ♥ xoxo <33 xo <333 ❤ ♡ #love s2 <URL-twitition.com> #neversaynever <3333

Highest Weighted Clusters

Cluster prefix	Tag	Types	Most common word in each cluster with prefix
11101010*	!	8160	lol lmao haha yes yea oh omg aww ah btw wow thanks sorry congrats welcome yay ha hey goodnight hi dear please huh wtf exactly idk bless whatever well ok
11000*	L	428	i'm im you're we're he's there's its it's
1110101100*	E	2798	x <3 :d :p :) :o :/
111110*	A	6510	young sexy hot slow dark low interesting easy important safe perfect special different random short quick bad crazy serious stupid weird lucky sad
1101*	D	378	the da my your ur our their his
01*	V	29267	do did kno know care mean hurts hurt say realize believe worry understand forget agree remember love miss hate think thought knew hope wish guess bet have
11101*	O	899	you yall u it mine everything nothing something anyone someone everyone nobody
100110*	&	103	or n & and

Tagger, tokenizer, and data all released at:
www.ark.cs.cmu.edu/TweetNLP

Datasets

	#Msg.	#Tok.	Tagset	Domain	Source
OCT27	1,827	26,594	Below	Twitter (Oct 27-28, 2010)	Gimpel et al. (2011)
DAILY547	547	7,707	Below	Twitter (Jan 2011–Jun 2012)	Annotated for this work
NPSCHAT	10,578	44,997	PTB-like	IRC (Oct–Nov 2006)	Forsyth and Martell (2007)
(w/o sys. msg.)	7,935	37,081			
RITTERTW	789	15,185	PTB-like	Twitter (dates unknown)	Ritter et al. (2011)

Results

Our tagger achieves **state-of-the-art results** in POS tagging for each dataset:

Feature set	OCT27TEST	DAILY547	NPSCHATTEST
All features	91.60	92.80	91.19
with clusters; without tagdicts, namelists	91.15	92.38	90.66
without clusters; with tagdicts, namelists	89.81	90.81	90.00
only clusters (and transitions)	89.50	90.54	89.55
without clusters, tagdicts, namelists	86.86	88.30	88.26
Gimpel et al. (2011) version 0.2	88.89	89.17	
Inter-annotator agreement (Gimpel et al., 2011)	92.2		
Model trained on all OCT27		93.2	

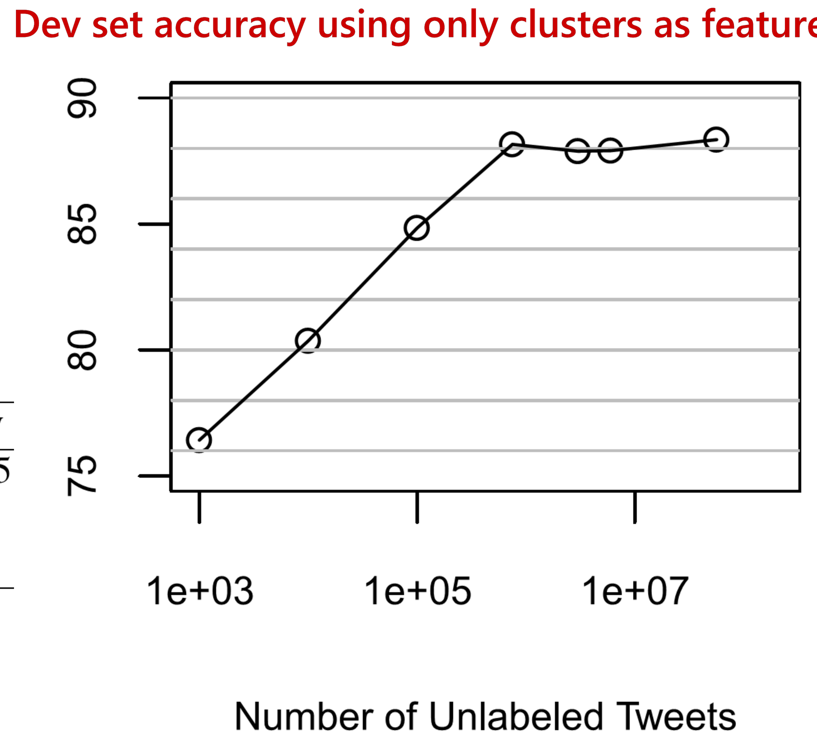
Accuracy on NPSCHATTEST corpus (incl. system messages)

Tagger	Accuracy
This work	93.4 ± 0.3
Forsyth (2007)	90.8

Accuracy on RITTERTW corpus

Tagger	Accuracy
This work	90.0 ± 0.5
Ritter et al. (2011), basic CRF tagger	85.3
Ritter et al. (2011), trained on more data	88.3

Dev set accuracy using only clusters as features



Speed

Tagger: 800 tweets/s (compared to 20 tweets/s previously)
Tokenizer: 3,500 tweets/s

Software & Data Release

- Improved emoticon detector and tweet tokenizer
- Newly annotated evaluation set, fixes to previous annotations

ARK Carnegie Mellon