Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and **Noah A. Smith**
School of Computer Science, Carnegie Mellon Univeristy, Pittsburgh, PA 15213, USA
{kgimpel,nschneid,brenocon,dipanjan,dpmills,
jacobeis,mheilman,dyogatama,jflanigan,nasmith}@cs.cmu.edu

Abstract

We address the problem of part-of-speech tagging for English data from the popular microblogging service Twitter. We develop a tagset, annotate data, develop features, and report tagging results nearing 90% accuracy. The data and tools have been made available to the research community with the goal of enabling richer text analysis of Twitter and related social media data sets.

1 Introduction

The growing popularity of social media and user-created web content is producing enormous quantities of text in electronic form. The popular microblogging service Twitter (twitter.com) is one particularly fruitful source of user-created content, and a flurry of recent research has aimed to understand and exploit these data (Ritter et al., 2010; Sharifi et al., 2010; Barbosa and Feng, 2010; Asur and Huberman, 2010; O'Connor et al., 2010a; Thelwall et al., 2011). However, the bulk of this work eschews the standard pipeline of tools which might enable a richer linguistic analysis; such tools are typically trained on newstext and have been shown to perform poorly on Twitter (Finin et al., 2010).

One of the most fundamental parts of the linguistic pipeline is part-of-speech (POS) tagging, a basic form of syntactic analysis which has countless applications in NLP. Most POS taggers are trained from treebanks in the newswire domain, such as the *Wall Street Journal* corpus of the Penn Treebank (PTB; Marcus et al., 1993). Tagging performance degrades on out-of-domain data, and Twitter poses additional challenges due to the conversational nature of the text, the lack of conventional orthography, and 140-character limit of each message ("tweet"). Figure 1 shows three tweets which illustrate these challenges.

(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$_{\#}$

Figure 1: Example tweets with gold annotations. Underlined tokens show tagger improvements due to features detailed in Section 3 (respectively: TAGDICT, METAPH, and DISTSIM).

In this paper, we produce an English POS tagger that is designed especially for Twitter data. Our contributions are as follows:

- we developed a POS tagset for Twitter,
- we manually tagged 1,827 tweets,
- we developed features for Twitter POS tagging and conducted experiments to evaluate them, and
- we provide our annotated corpus and trained POS tagger to the research community.

Beyond these specific contributions, we see this work as a case study in how to rapidly engineer a core NLP system for a new and idiosyncratic dataset. This project was accomplished in 200 person-hours spread across 17 people and two months. This was made possible by two things: (1) an annotation scheme that fits the unique characteristics of our data and provides an appropriate level of linguistic detail, and (2) a feature set that captures Twitter-specific properties and exploits existing resources such as tag dictionaries and phonetic normalization. The success of this approach demonstrates that with careful design, supervised machine learning can be applied to rapidly produce effective language technology in new domains.

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
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
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

Table 1: The set of tags used to annotate tweets. The last column indicates each tag's relative frequency in the full annotated data (26,435 tokens). (The rates for **M** and **Y** are both < 0.0005.)

2 Annotation

Annotation proceeded in three stages. For **Stage 0**, we developed a set of 20 coarse-grained tags based on several treebanks but with some additional categories specific to Twitter, including URLs and hashtags. Next, we obtained a random sample of mostly American English[1] tweets from October 27, 2010, automatically tokenized them using a Twitter tokenizer (O'Connor et al., 2010b),[2] and pre-tagged them using the WSJ-trained Stanford POS Tagger (Toutanova et al., 2003) in order to speed up manual annotation. Heuristics were used to mark tokens belonging to special Twitter categories, which took precedence over the Stanford tags.

Stage 1 was a round of manual annotation: 17 researchers corrected the automatic predictions from Stage 0 via a custom Web interface. A total of 2,217 tweets were distributed to the annotators in this stage; 390 were identified as non-English and removed, leaving 1,827 annotated tweets (26,436 tokens).

The annotation process uncovered several situations for which our tagset, annotation guidelines, and tokenization rules were deficient or ambiguous. Based on these considerations we revised the tokenization and tagging guidelines, and for **Stage 2**, two annotators reviewed and corrected all of the English tweets tagged in Stage 1. A third annotator read the annotation guidelines and annotated 72 tweets from scratch, for purposes of estimating inter-annotator agreement. The 72 tweets comprised 1,021 tagged tokens, of which 80 differed from the Stage 2 annotations, resulting in an agreement rate of 92.2% and Cohen's κ value of 0.914. A final sweep was made by a single annotator to correct errors and improve consistency of tagging decisions across the corpus. The released data and tools use the output of this final stage.

2.1 Tagset

We set out to develop a POS inventory for Twitter that would be intuitive and informative—while at the same time simple to learn and apply—so as to maximize tagging consistency within and across an-

[1] We filtered to tweets sent via an English-localized user interface set to a United States timezone.

[2] http://github.com/brendano/tweetmotif

notators. Thus, we sought to design a coarse tagset that would capture standard parts of speech[3] (noun, verb, etc.) as well as categories for token varieties seen mainly in social media: URLs and email addresses; emoticons; Twitter **hashtags**, of the form #tagname, which the author may supply to categorize a tweet; and Twitter **at-mentions**, of the form @user, which link to other Twitter users from within a tweet.

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 expresssions (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.

Another tag, **~**, is used for tokens marking specific Twitter discourse functions. The most popular of these is the RT ("retweet") construction to publish a message with attribution. For example,

> RT @USER1 : LMBO ! This man filed an EMERGENCY Motion for Continuance on account of the Rangers game tonight ! ≪ Wow lmao

indicates that the user @USER1 was originally the source of the message following the colon. We apply **~** to the RT and : (which are standard), and also ≪, which separates the author's comment from the retweeted material.[4] Another common discourse marker is ellipsis dots (. . .) at the end of a tweet, indicating a message has been truncated to fit the 140-character limit, and will be continued in a subsequent tweet or at a specified URL.

Our first round of annotation revealed that, due to nonstandard spelling conventions, tokenizing under a traditional scheme would be much more difficult 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} × {verb, possessive}.[5]

The final tagging scheme (Table 1) encompasses 25 tags. For simplicity, each tag is denoted with a single ASCII character. The miscellaneous category **G** includes multiword abbreviations that do not fit in any of the other categories, like ily (*I love you*), as well as partial words, artifacts of tokenization errors, miscellaneous symbols, possessive endings,[6] and arrows that are not used as discourse markers.

Figure 2 shows where tags in our data tend to occur relative to the middle word of the tweet. We see that Twitter-specific tags have strong positional preferences: at-mentions (**@**) and Twitter discourse markers (**~**) tend to occur towards the beginning of messages, whereas URLs (**U**), emoticons (**E**), and categorizing hashtags (**#**) tend to occur near the end.

3 System

Our tagger is a conditional random field (CRF; Lafferty et al., 2001), enabling the incorporation of arbitrary local features in a log-linear model. Our base features include: a feature for each word type, a set of features that check whether the word contains digits or hyphens, suffix features up to length 3, and features looking at capitalization patterns in the word. We then added features that leverage domain-specific properties of our data, unlabeled in-domain data, and external linguistic resources.

TWORTH: Twitter orthography. We have features for several regular expression-style rules that detect at-mentions, hashtags, and URLs.

NAMES: Frequently-capitalized tokens. Microbloggers are inconsistent in their use of capitalization, so we compiled gazetteers of tokens which are frequently capitalized. The likelihood of capitalization for a token is computed as $\frac{N_{\text{cap}}+\alpha C}{N+C}$, where

[3] Our starting point was the cross-lingual tagset presented by Petrov et al. (2011). Most of our tags are refinements of those categories, which in turn are groupings of PTB WSJ tags (see column 2 of Table 1). When faced with difficult tagging decisions, we consulted the PTB and tried to emulate its conventions as much as possible.

[4] These "iconic deictics" have been studied in other online communities as well (Collister, 2010).

[5] The modified tokenizer is packaged with our tagger.

[6] Possessive endings only appear when a user or the tokenizer has separated the possessive ending from a possessor; the tokenizer only does this when the possessor is an at-mention.

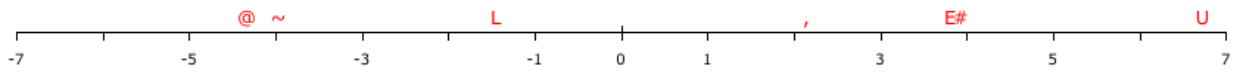

Figure 2: Average position, relative to the middle word in the tweet, of tokens labeled with each tag. Most tags fall between -1 and 1 on this scale; these are not shown.

N is the token count, N_{cap} is the capitalized token count, and α and C are the prior probability and its prior weight.[7] We compute features for membership in the top N items by this metric, for $N \in \{1000, 2000, 3000, 5000, 10000, 20000\}$.

TAGDICT: Traditional tag dictionary. We add features for all coarse-grained tags that each word occurs with in the PTB[8] (conjoined with their frequency rank). Unlike previous work that uses tag dictionaries as hard constraints, we use them as soft constraints since we expect lexical coverage to be poor and the Twitter dialect of English to vary significantly from the PTB domains. This feature may be seen as a form of type-level domain adaptation.

DISTSIM: Distributional similarity. When training data is limited, distributional features from unlabeled text can improve performance (Schütze and Pedersen, 1993). We used 1.9 million tokens from 134,000 unlabeled tweets to construct distributional features from the successor and predecessor probabilities for the 10,000 most common terms. The successor and predecessor transition matrices are horizontally concatenated into a sparse matrix $\mathbf{M}$, which we approximate using a truncated singular value decomposition: $\mathbf{M} \approx \mathbf{USV}^{\mathsf{T}}$, where $\mathbf{U}$ is limited to 50 columns. Each term's feature vector is its row in $\mathbf{U}$; following Turian et al. (2010), we standardize and scale the standard deviation to 0.1.

METAPH: Phonetic normalization. Since Twitter includes many alternate spellings of words, we used the Metaphone algorithm (Philips, 1990)[9] to create a coarse phonetic normalization of words to simpler keys. Metaphone consists of 19 rules that rewrite consonants and delete vowels. For example, in our data, {thangs thanks thanksss thanx thinks thnx} are mapped to *0NKS*, and {lmao lmaoo lmaooooo} map to *LM*. But it is often too coarse; e.g. {war we're wear were where worry} map to *WR*.

We include two types of features. First, we use the Metaphone key for the current token, complementing the base model's word features. Second, we use a feature indicating whether a tag is the most frequent tag for PTB words having the same Metaphone key as the current token. (The second feature was disabled in both $-$TAGDICT and $-$METAPH ablation experiments.)

4 Experiments

Our evaluation was designed to test the efficacy of this feature set for part-of-speech tagging given limited training data. We randomly divided the set of 1,827 annotated tweets into a training set of 1,000 (14,542 tokens), a development set of 327 (4,770 tokens), and a test set of 500 (7,124 tokens). We compare our system against the Stanford tagger. Due to the different tagsets, we could not apply the pretrained Stanford tagger to our data. Instead, we retrained it on our labeled data, using a standard set of features: words within a 5-word window, word shapes in a 3-word window, and up to length-3 prefixes, length-3 suffixes, and prefix/suffix pairs.[10] The Stanford system was regularized using a Gaussian prior of $\sigma^2 = 0.5$ and our system with a Gaussian prior of $\sigma^2 = 5.0$, tuned on development data.

The results are shown in Table 2. Our tagger with the full feature set achieves a relative error reduction of 25% compared to the Stanford tagger. We also show feature ablation experiments, each of which corresponds to removing one category of features from the full set. In Figure 1, we show examples that certain features help solve. Underlined tokens

[7] $\alpha = \frac{1}{100}, C = 10$; this score is equivalent to the posterior probability of capitalization with a Beta$(0.1, 9.9)$ prior.

[8] Both WSJ and Brown corpora, no case normalization. We also tried adding the WordNet (Fellbaum, 1998) and Moby (Ward, 1996) lexicons, which increased lexical coverage but did not seem to help performance.

[9] Via the Apache Commons implementation: `http://commons.apache.org/codec/`

[10] We used the following feature modules in the Stanford tagger: `bidirectional5words`, `naacl2003unknowns`, `wordshapes(-3,3)`, `prefix(3)`, `suffix(3)`, `prefixsuffix(3)`.

	Dev.	Test	
Our tagger, all features	88.67	89.37	
independent ablations:			
−DISTSIM	87.88	88.31	(−1.06)
−TAGDICT	88.28	88.31	(−1.06)
−TWORTH	87.51	88.37	(−1.00)
−METAPH	88.18	88.95	(−0.42)
−NAMES	88.66	89.39	(+0.02)
Our tagger, base features	82.72	83.38	
Stanford tagger	85.56	85.85	
Annotator agreement	92.2		

Table 2: Tagging accuracies on development and test data, including ablation experiments. Features are ordered by importance: test accuracy decrease due to ablation (final column).

Tag	Acc.	Confused	Tag	Acc.	Confused
V	91	N	!	82	N
N	85	^	L	93	V
,	98	~	&	98	^
P	95	R	U	97	,
^	71	N	$	89	P
D	95	^	#	89	^
O	97	^	G	26	,
A	79	N	E	88	,
R	83	A	T	72	P
@	99	V	Z	45	^
~	91	,			

Table 3: Accuracy (recall) rates per class, in the test set with the full model. (Omitting tags that occur less than 10 times in the test set.) For each gold category, the most common confusion is shown.

are incorrect in a specific ablation, but are corrected in the full system (i.e. when the feature is added).

The −TAGDICT ablation gets elects, Governor, and next wrong in tweet (a). These words appear in the PTB tag dictionary with the correct tags, and thus are fixed by that feature. In (b), withhh is initially misclassified an interjection (likely caused by interjections with the same suffix, like ohhh), but is corrected by METAPH, because it is normalized to the same equivalence class as with. Finally, s/o in tweet (c) means "shoutout", which appears only once in the training data; adding DISTSIM causes it to be correctly identified as a verb.

Substantial challenges remain; for example, despite the NAMES feature, the system struggles to identify proper nouns with nonstandard capitalization. This can be observed from Table 3, which shows the recall of each tag type: the recall of proper nouns (^) is only 71%. The system also struggles with the miscellaneous category (G), which covers many rare tokens, including obscure symbols and artifacts of tokenization errors. Nonetheless, we are encouraged by the success of our system on the whole, leveraging out-of-domain lexical resources (TAGDICT), in-domain lexical resources (DISTSIM), and sublexical analysis (METAPH).

Finally, we note that, even though 1,000 training examples may seem small, the test set accuracy when training on only 500 tweets drops to 87.66%, a decrease of only 1.7% absolute.

5 Conclusion

We have developed a part-of-speech tagger for Twitter and have made our data and tools available to the research community at `http://www.ark.cs.cmu.edu/TweetNLP`. More generally, we believe that our approach can be applied to address other linguistic analysis needs as they continue to arise in the era of social media and its rapidly changing linguistic conventions. We also believe that the annotated data can be useful for research into domain adaptation and semi-supervised learning.

Acknowledgments

We thank Desai Chen, Chris Dyer, Lori Levin, Behrang Mohit, Bryan Routledge, Naomi Saphra, and Tae Yano for assistance in annotating data. This research was supported in part by: the NSF through CAREER grant IIS-1054319, the U. S. Army Research Laboratory and the U. S. Army Research Office under contract/grant number W911NF-10-1-0533, Sandia National Laboratories (fellowship to K. Gimpel), and the U. S. Department of Education under IES grant R305B040063 (fellowship to M. Heilman).

References

Sitaram Asur and Bernardo A. Huberman. 2010. Predicting the future with social media. In *Proc. of WI-IAT*.

Luciano Barbosa and Junlan Feng. 2010. Robust sentiment detection on Twitter from biased and noisy data. In *Proc. of COLING*.

Lauren Collister. 2010. Meaning variation of the iconic deictics ^ and <— in an online community. In *New Ways of Analyzing Variation*.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

Tim Finin, Will Murnane, Anand Karandikar, Nicholas Keller, Justin Martineau, and Mark Dredze. 2010. Annotating named entities in Twitter data with crowdsourcing. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. of ICML*.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19:313–330.

Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. 2010a. From tweets to polls: Linking text sentiment to public opinion time series. In *Proc. of ICWSM*.

Brendan O'Connor, Michel Krieger, and David Ahn. 2010b. TweetMotif: Exploratory search and topic summarization for Twitter. In *Proc. of ICWSM (demo track)*.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2011. A universal part-of-speech tagset. *ArXiv:1104.2086*.

Lawrence Philips. 1990. Hanging on the Metaphone. *Computer Language*, 7(12).

Alan Ritter, Colin Cherry, and Bill Dolan. 2010. Unsupervised modeling of Twitter conversations. In *Proc. of NAACL*.

Hinrich Schütze and Jan Pedersen. 1993. A vector model for syntagmatic and paradigmatic relatedness. In *Proceedings of the 9th Annual Conference of the UW Centre for the New OED and Text Research*.

Beaux Sharifi, Mark-Anthony Hutton, and Jugal Kalita. 2010. Summarizing microblogs automatically. In *Proc. of NAACL*.

Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. 2011. Sentiment in Twitter events. *Journal of the American Society for Information Science and Technology*, 62(2):406–418.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proc. of HLT-NAACL*.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: a simple and general method for semi-supervised learning. In *Proc. of ACL*.

Grady Ward. 1996. Moby lexicon. `http://icon.shef.ac.uk/Moby`.