Part of Speech tagger

PoS tagger for English data from twitter. Accuracy: 81.4%

Part-of-speech tagging is one of the most important NLP tasks used to classify words into their part-of-speech classes.

There are two types of such classes:
1) Open classes - those for which new word are appearing.
2) Closed classes - completed classes, they will not grow in size.(e.g. articles)

Every natural language has rules. But it also has exceptions, which makes pos-tagging task difficult. There are many words that could have different PoS with the same spelling.

This PoS tagger is trained and tested on twitter data. Each line contains tag and word, separated by tab. Lines form sentences, divided by blank line.

Tags:

Tag	Example
NNP	Russia
NN	Post
:	…, -, :
CD	2010 (Cardinal Numbers)
(	(
)	)
IN	In, For (prepositions)
URL	https://google.com/
RT	RT
USR	@alex777(username)
HT	#hashtag
.	., ?, !
WRB	How, When
PRP	I, u, they (Personal pronoun)
VBP	Verbs (non 3rd person singular present)
MD	Can, must (Modal verbs)
RB	slowly, everyday, right, no== (adverb)
VB	Post - verbs(base form)
UH	Ha-ha (Interjection)
VBG	gonnauh (gerunds)
JJ	Big (Adjective)

Tag	Example
,	,
CC	And, & (Coordinating Conjunctions)
PRP$	Ur, mine (Possessive Pronouns)
DT	The (Determiner)
JJS	Most (Superlative Adjectives)
NNS	people, emails (Common Nouns (Plural))
VBZ	Shoots (Verbs (3rd person singular present))
RBR	Better (Comparative Adverbs)
VBN	Gone (Verbs (past participle))
VBD	Asked (Verbs (past tense))
TO	To
RP	Down, off (Particles)
EX	There (Existence There)
POS	’s (Possessive Endings ’s)
WP	What (Wh-pronoun)
WDT	What (Wh-determiner)
FW	Etc (Foreign Words)
JJR	Darker (Comparative Adjectives)
' '	‘, “
NNPS	Queens (Proper Nouns (Plural))
SYM	+, &lt; (symbols)
RBS	Most (Superlative Adverbs)
VPP	Please
O	“..
LS	1 (List Item Markers)
TD	A

From the table above we can identify some problems:
1) same words could be of different pos.
2) HT. hashtags could be also part of sentence.
USR/@GuardWifeL USR/@TiffanyGreen48 PRP/I VBP’ve RB/already HT/#ff’ed PRP/her PRP/she VBZ/’s RB/not JJ/worthy IN/of CD/two IN/in CD/one NN/day ./! UH/Lol
3) Use of not-correct spelled words: 2moro, 4, sk8, jst, u, ur

To create pos tagger we should train classifier. The main problem here is defining features.
As we can see, it is not enough to predict pos of the word from only this word. So we should also consider the context.

E.g. we can look at previous and next words. But this works bad in practice, because the pos of a word is not depends on the word itself, but it depends on pos of the previous or next word. So we will look at the most informative feature of these words:
2-letter and 3-letter suffixes.
2-letter suffix could define VBD, which always ends with -ed, RB which often ends with -ly etc.
3-letter suffix helps to define VBG (-ing)

As classifier I use DecisionTreeClassifier. The goal of Decision Tree is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

Advantages:
Requires little data preparation. Other techniques often require data normalisation, dummy variables need to be created and blank values to be removed.

Able to handle both numerical and categorical data.

Disadvantages:
Decision-tree learners can create over-complex trees that do not generalise the data well.

Decision trees can be unstable because small variations in the data might result in a completely different tree being generated.

Decision tree learners create biased trees if some classes dominate. It is therefore recommended to balance the dataset prior to fitting with the decision tree.

Possible Improvements:

- Find more informative features.
- Normalize data.