CS60075 Natural Language Processing Assignment 4: POS Tagging of Universal Dependency Hindi corpus

Aditya Sawant 17CS10060

Features chosen:

word – the word itself
word.Lower() - the word is reduced to lowercase
word.isTitle() - Boolean True if only the first character is in uppercase and rest are lowercase
word.isUpper() - Boolean True if all characters of string are uppercase
word.isDigit() - Boolean True if all characters of string are Digits
Prefix-1 - word[0] first character of word
Prefix-2 - word[0:2] first 2 characters of word
Prefix-3 - word[0:3] first 3 characters of word
Suffix-1 - word[-1] last character of word
Suffix-2 - word[-2:] last 2 characters of word
Suffix-3 - word[-3:] last 3 characters of word
has_Hyphen - Boolean True if word has hyphen in it

BOS - If word is the Beginning of the Sentence
-1:word.Lower() - previous word reduced to lowercase
-1:word.isTitle() - Boolean True if only first character of previous word is uppercase and rest are lowercase
-1:word.isUpper() - Boolean True if all characters of the previous word are uppercase
-1:postag - pos tag of previous word

EOS - If word is end of the sentence
+1:word.Lower() - next word reduced to lowercase
+1:word.isTitle() - Boolean True if only first character of the next word is in uppercase and rest in lowercase
+1:word.isUpper() - Boolean True if all characters of next word are in uppercase
+1:postag - pos tag of next word

Top 10 Most Common POS Transition Features

VERB => AUX	2.20021
AUX => AUX	1.55228
NUM => NOUN	1.47586
ADJ => NOUN	1.37534
PROPN => ADP	1.35694
PROPN => PROPN	1.33343
NOUN => ADP	1.28564
VERB => SCONJ	1.25156
DET => NOUN	1.22831
NOUN => VERB	1.09714

Top 10 Least Common POS Transition Features

AUX => ADP	-1.06835
COMMA => ADP	-1.08245
NUM => PRON	-1.10012
PUNCT => PUNCT	-1.10883
ADP => COMMA	-1.22990
DET => CCONJ	-1.33245
CCONJ => AUX	-1.47313
ADJ => PRON	-1.84110
ADJ => ADP	-2.05786
DET => ADP	-2.30424

Model Prediction on Training Data

	precision	recall	f1-score	support
ADJ	1.00	1.00	1.00	570
ADP	1.00	1.00	1.00	1387
ADV	0.99	0.99	0.99	111
AUX	0.99	1.00	0.99	730
CCONJ	1.00	1.00	1.00	150
COMMA	1.00	1.00	1.00	114
DET	1.00	0.99	0.99	231
NOUN	1.00	1.00	1.00	1597
NUM	1.00	1.00	1.00	152
PART	1.00	1.00	1.00	163
PRON	1.00	1.00	1.00	431
PROPN	1.00	1.00	1.00	708
PUNCT	1.00	1.00	1.00	564
SCONJ	0.98	1.00	0.99	61
VERB	1.00	0.98	0.99	640

Model Prediction on Test Data

	precision	recall	f1-score	support
ADJ	0.74	0.78	0.76	94
ADP	0.96	0.98	0.97	309
ADV	0.71	0.48	0.57	21
AUX	0.96	0.97	0.97	139
CCONJ	1.00	1.00	1.00	25
COMMA	-	-	-	-
DET	0.89	0.92	0.90	36
NOUN	0.83	0.89	0.86	329
NUM	1.00	0.92	0.96	25
PART	1.00	0.97	0.98	33
PRON	0.92	0.88	0.90	65
PROPN	0.81	0.69	0.74	145
PUNCT	1.00	1.00	1.00	135
SCONJ	0.60	1.00	0.75	3
VERB	0.90	0.88	0.89	99

Metrics obtained for training set

precision: 0.9976580869578112
recall: 0.9976350019708317
f1-score: 0.9976342225110107
accuracy: 0.9976350019708317

Metrics obtained for the test set

precision: 0.8967440932480188
recall: 0.897119341563786
f1-score: 0.8957270489881956
accuracy: 0.897119341563786