CRFSuite POS Tagger

Bulat Nasrulin

Innopolis University

Our approach

CRF- suite is fast training and tagging tools with implemented state-of-the-art methods:

1. Limited-memory BFGS (L-BFGS) [Nocedal 80]
2. Orthant-Wise Limited-memory Quasi-Newton (OWL-QN) method [Andrew 07]
3. Stochastic Gradient Descent (SGD) [Shalev-Shwartz 07]
4. Averaged Perceptron (AP) [Collins 02]
5. Passive Aggressive (PA) [Crammer 06]
6. Adaptive Regularization Of Weight Vector (AROW) [Mejer 10]

We evaluate all these methods in terms of item accuracy (words correctly tagged) and instance accuracy (the whole sentence correctly tagged).

English Twitter dataset

To use crfsuite for tagging, we need firstly choose good features.

In the table features are presented. For each feature we also provide it's importance (how much it changes the accuracy).

Feature	Description	Importance
word	the whole word in lower case	very important
suffix	last two letters	important
prefix	first letter(or symbol) in the word	important
first_captital	is the first symbol - upper case	important
length	word length	important for inst. accr.
number	is word a number	slightly increases item accuracy
number	is word a number	slightly increases item accuracy
in_dict	is word presented in brown dictionary	little improvement
next	next word after current	important
prev	word before current	no positive effect
ht	first symbol hashtag	no positive effect
VBP	is word one of 're, 'm, 's, etc.	no positive effect
case	All letters are upper case	no positive effect

Best models are PA and AP with:

Item accuracy: 1886 / 2211 (0.8530)

Instance accuracy: 22 / 116 (0.1897)

Performance by label (match, model, ref) (precision, recall, F1):
NNP : $(128, 175, 169)(0.7314, 0.7574, 0.7442)$
NN : $(222, 279, 286)(0.7957, 0.7762, 0.7858)$
: $(77, 82, 81)(0.9390, 0.9506, 0.9448)$
CD : $(33, 42, 35)(0.7857, 0.9429, 0.8571)$
(: $(3, 3, 4)(1.0000, 0.7500, 0.8571)$
) : $(6, 6, 6)(1.0000, 1.0000, 1.0000)$
IN : $(161, 175, 167)(0.9200, 0.9641, 0.9415)$
URL : $(20, 24, 23)(0.8333, 0.8696, 0.8511)$
RT : $(24, 24, 24)(1.0000, 1.0000, 1.0000)$
USR : $(63, 63, 64)(1.0000, 0.9844, 0.9921)$
HT : $(26, 26, 26)(1.0000, 1.0000, 1.0000)$
. : $(123, 128, 123)(0.9609, 1.0000, 0.9801)$
WRB : $(22, 23, 22)(0.9565, 1.0000, 0.9778)$
PRP : $(159, 166, 164)(0.9578, 0.9695, 0.9636)$
VBP : $(63, 82, 82)(0.7683, 0.7683, 0.7683)$
MD : $(25, 27, 27)(0.9259, 0.9259, 0.9259)$
RB : $(82, 101, 92)(0.8119, 0.8913, 0.8497)$
VB : $(68, 90, 89)(0.7556, 0.7640, 0.7598)$
UH : $(46, 68, 63)(0.6765, 0.7302, 0.7023)$
VBG : $(31, 34, 35)(0.9118, 0.8857, 0.8986)$
JJ : $(63, 86, 116)(0.7326, 0.5431, 0.6238)$
,: $(40, 40, 40)(1.0000, 1.0000, 1.0000)$
CC : $(42, 45, 43)(0.9333, 0.9767, 0.9545)$
PRP: (26, 30, 32) (0.8667, 0.8125, 0.8387)
DT: (106, 114, 115) (0.9298, 0.9217, 0.9258)
JJS: (2, 3, 3) (0.6667, 0.6667, 0.6667)
NNS: (50, 62, 64) (0.8065, 0.7812, 0.7937)
VBZ: (47, 56, 58) (0.8393, 0.8103, 0.8246)
RBR: (4, 4, 6) (1.0000, 0.6667, 0.8000)
VBN: (9, 16, 18) (0.5625, 0.5000, 0.5294)
VBD: (34, 48, 44) (0.7083, 0.7727, 0.7391)
TO: (41, 41, 42) (1.0000, 0.9762, 0.9880)
RP: (11, 12, 14) (0.9167, 0.7857, 0.8462)
EX: (3, 3, 4) (1.0000, 0.7500, 0.8571)
POS: (4, 4, 4) (1.0000, 1.0000, 1.0000)
WP: (9, 9, 10) (1.0000, 0.9000, 0.9474)
WDT: (1, 4, 1) (0.2500, 1.0000, 0.4000)
JJR: (2, 4, 4) (0.5000, 0.5000, 0.5000)

Russian LiveJournal Dataset

To further extend this work we repeat same experiment but for LiveJournal dataset with Russian Language. **Best model:** Averaged Perseptron:

Item accuracy: 10831 / 11548 (0.9379)
Instance accuracy: 570 / 1025 (0.5561)

Feature	Description	Importance
word	the whole word in lower case	very important
prefix	first 2 letter(or symbols) in the word	important
first_captital	is the first symbol - upper case	important
length	word length	important for inst. accr.
number	is word a number	slightly increases item accuracy
next	next word after current	slightly important
prev	word before current	no positive effect
case	All letters are upper case	no positive effect

Accuracy for specific labels:
N (Noun): (2401, 2655, 2566) (0.9043, 0.9357, 0.9197)
.: (2145, 2145, 2145) (1.0000, 1.0000, 1.0000)
M (Numeral): (154, 158, 154) (0.9747, 1.0000, 0.9872)
C (Conjunction): (648, 723, 693) (0.8963, 0.9351, 0.9153)
Q (Particle): (401, 427, 460) (0.9391, 0.8717, 0.9042)
V (Verb): (1406, 1572, 1526) (0.8944, 0.9214, 0.9077)
P (Pronoun): (1390, 1418, 1434) (0.9803, 0.9693, 0.9748)
A (Adjective): (640, 739, 788) (0.8660, 0.8122, 0.8382)
R (Adverb): (452, 482, 554) (0.9378, 0.8159, 0.8726)
S (Preposition): (925, 937, 928) (0.9872, 0.9968, 0.9920)
H (Parenthetical phrase): (40, 46, 52) (0.8696, 0.7692, 0.8163)
I (Interjection): (20, 31, 28) (0.6452, 0.7143, 0.6780)
X (Residual): (167, 168, 172) (0.9940, 0.9709, 0.9824)
W (Predicative): (42, 47, 48) (0.8936, 0.8750, 0.8842)

Final notes

CRF- suite is a powerful tool for achieving a good accuracy. Although we provide some useful features, it is hard to design and evaluate the 'hand-crafted' features.

References

1. CRFSuite, $http://www.chokkan.org/software/crfsuite/$
2. Russian LiveJournal dataset, $http://www.webcorpora.ru/news/282$