

Writeup: Assignment 3

- Dhruvin Patel

2012CS50293

Q1)

Trained pos and ner model using Model.traindev

Fscore for pos: 0.59

Fscore for ner: 0.18

Q2) Changes are made appropriately to matlab files.

Q3)

Final Fscore for POS Tags (training on pos.traindev and testing on pos.test): 0.738145

Precision = 0.804282, Recall = 0.682058

Final Fscore for NER Tags (training on pos.traindev and testing on pos.test): 0.360302

Precision = 0.652044, Recall = 0.248926

Features tried were corresponding to prefix of token, suffix of token, length of token, Next_token, Prev_token, ISCAPITAL, HasPunctuation, HAS@, HAS#, HASurl, HAS: , HAS~, HASRT, whetherlengthequalsone, LENGTHGREATERTHANX, LENGTHLESSTHANX, (X was varied from 1 to 10), nltk features: lammetizer, Browncluster feature index (cluster id to which a token belongs), List of pos tags obtained from various public domains, List of ner tags such as PER, LOC, ORG, etc.

Ablation study for POS tags

(Note: Only those features are shown in ablation study which are retained in the final model, because few features had negative contribution to the accuracy)

Feature	Fscore
Baseline (consisiting of input token)	0.5907
3Prefix + 2Prefix	0.6113
3Suffix + 2Suffix	0.6187

Nextword	0.6191
Prevword	0.6199
ISFIRSTWORDCAPITAL	0.6202
HASPUNCTUATIONS ('(', ')', '_, '>', '<', '=')	0.6632
HAS@	0.6743
HAS#	0.6796
HAS:	0.6855
HASurl	0.6904
LENGTHGREATERTHAN8	0.6911
HAS~	0.7123
HASRT	0.7189
LENGTHEQUALSONE	0.7314
POSTAGLISTFROMDOMAINS	0.7543
NERLISTFROMDOMAINS	0.7580

Ablation study for NER tags

(Note: Only those features are shown in ablation study which are retained in the final model, because few features had negative contribution to the accuracy)

Feature	Fscore
Baseline (consisiting of input token)	0.1856
3Prefix + 2Prefix	0.2145
3Suffix + 2Suffix	0.2211
Nextword	0.2498
Prevword	0.2643
ISWORDCAPITAL	0.2788
HASPUNCTUATIONS ('(', ')', '_, '>', '<', '=')	0.2806
HAS@	0.2834
HAS#	0.2834
HAS:	0.2839
HASurl	0.2849
HAS~	0.2865
HASRT	0.2876
LENGTHEQUALSONE	0.2907

POSTAGLISTFROMDOMAINS	0.2909
NERLISTFROMDOMAINS	0.3603

Features which didn't work:

Brown cluster (id) (Possibly because number of ids is very large around 2^{13} , that crf donot have sufficient data to learn each of those feature)

LENGTHGREATERTHANX (for X about 3,4,5) Because this feature donot lend any information.

Dataset References:

NER TAG LIST: <http://www.clips.uantwerpen.be/conll2003/ner/>

POS TAG LIST: <http://downloads.gate.ac.uk/twitie/twitie-tagger.zip>

BROWN CLUSTER IDS: <http://www.ark.cs.cmu.edu/TweetNLP/clusters/>