PGM Assignment 3B Report

submitted by : K.Sai Vignan 2012CS50289

In this assignment I have tried to experiment with CRF model and learnt to deal with mallet code In general I have done assignment with order 1 i.e a linear chain CRF which will be similar to HMM.The main difference between the two is the ability to add interdependent and arbitrary features in CRF compared to HMM.Experiment results are given according to the sequence asked in assignment file

PART 1:

I have trained the CRF model with the default order of 1 on both POS.traindev and NER.traindev without adding any features and after order of 2 has been experimented.order of 3 has taken a very long time (>28hrs) .So, results of Markov order 3 are not mentioned.Number of iterations used to train data is default 500 and threads of 4

POS

	macroPrecision	macroRecall	**macroF_Score**
Order 1	0.660922	0.456428	**0.539962**
Order 2	0.520531	0.408060	**0.457484**

NER

	macroPrecision	macroRecall	**macroF_Score**
Order 1	0.85	0.0068	**0.013580**
Order 2	0.85	0.0019	**0.003779**

PART 2: HMM

I edited the part of SimpleTagger to obtain HMMSimpleTagger.java and the main edit will be Using the HMM class in HMM.java instead of CRF class and to rectify some error of array out of bounds I edited HMM.java.I took help of an implementation of HMMSimpleTagger main differnce is removing of the threads and used classes already built.iterations 500

	macroPrecision	macroRecall	**macroF_Score**
POS	0.515298	0.120532	**0.189329**
NER	1	0	**0**

We can see the advantage of CRF,the inclusion of arbitrary features over HMM with these results.

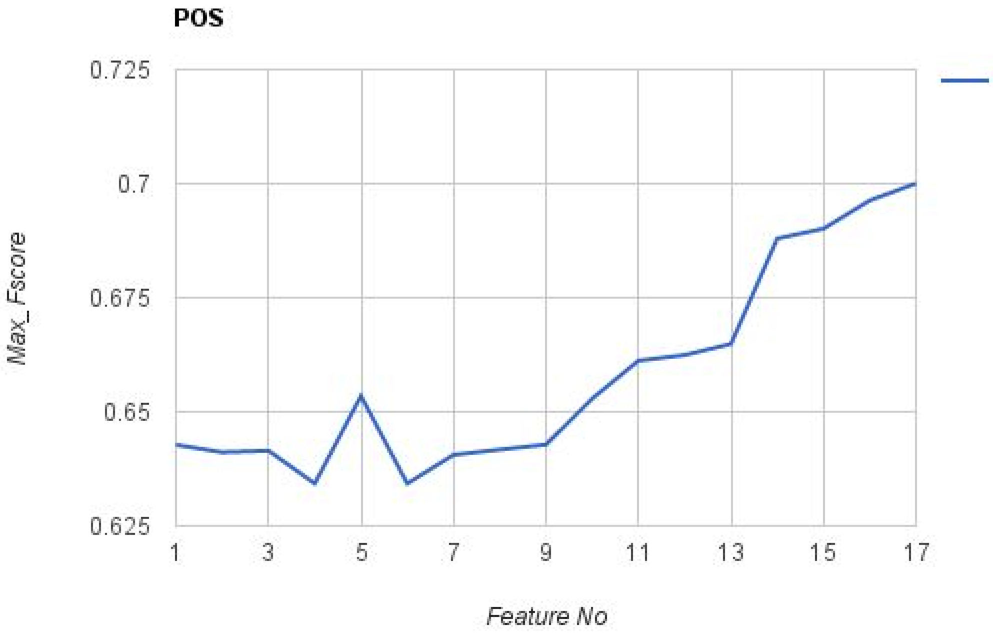
PART3: Adding Features

POS

I have tried various features adding one after another.basic features like capitals,starting with capital,contains punctuation mark,numeral etc are motivated from the paper given as reference for POS.Total features I have tried are **17**
Some of the important features added by me are maintaining a list of stopwords (words which can be removed and sentence can still makes sense,i.e generally over repetitive words),if the word is any one of the RT,@,url,# then only the feature added by me for it will be the RT,AT,URL,Hash as they can not be other things.By doing this more weight for this label(Misc(G)).Other important feature is Prefix and suffixes I mainly tried this to cluster words which are most likely having same label.Other feature is words ending with ‘ing’,’ed’,’es’,’ly’ etc. Other dictionaries maintained are list of emoticons to strengthen feature strength for label E. Study of F_score adding one after another feature can be seen in below table,Number of iterations is default 500 and order is of one

Feature added at current step	**Max_FScore**
Array of some symmetric tags	0.642760
Symbols like RT,@,#,url	0.641141
ALL_CAPS	0.641489
IS_CAPITALIZED	0.634281
IS_NUM	0.653474
SINGLEDIGIT	0.634273
DOUBLEDIGIT	0.640568
HASDASH	0.641713
PUNCTUATION	0.642821
STOP_WORD	0.652893

EMOTICON	0.661202
END=ed	0.662438
END=ly	0.664863
Prefix and Suffixes	0.687966
END=ing	0.690121
END=es	0.696311
END=ss	0.700041

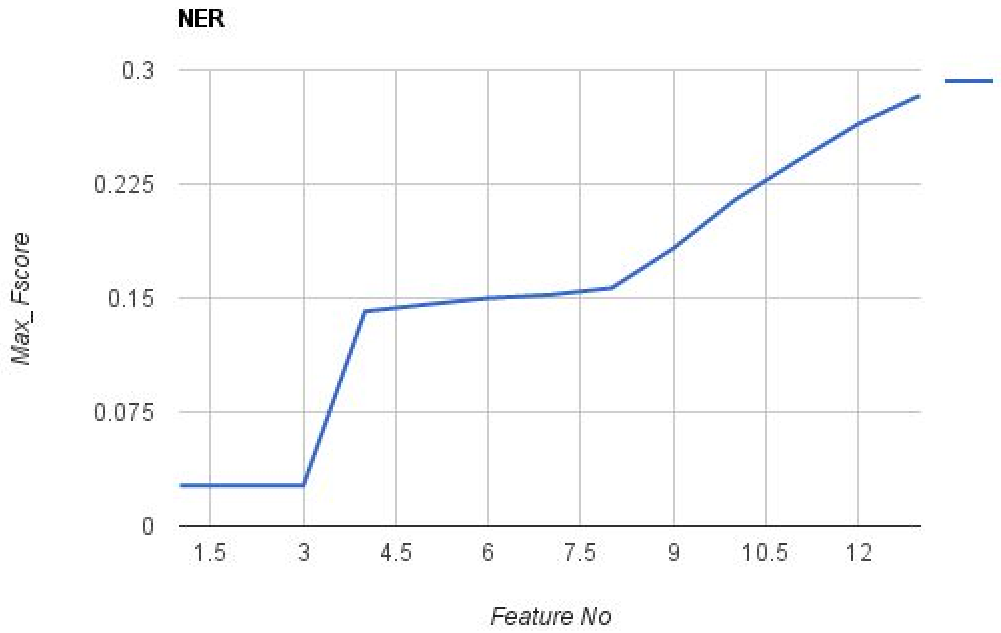


Some of the major observations we can see here are the drastic increase at the feature prefix and suffixes and a steady increase at endings,dictionaries which indicates these are important features.The fluctuations can be associated with fact that the feature we have added hasn't contributed in increasing the strength of labels and unnecessarily regularizing the weights to other features thus decreasing the strength of before labels and in turn decreasing Fscore

NER

NER has very less labels compared to POS labels Mainly here the task is to identify person,company,location etc.I have applied some features as similar to POS and mainly added a database of company,facility,geoloc,person,product database has been extracted from the development files given and some online resources.

Feature added at current step	**Max_FScore**
Symbols like RT,@,#,url,PUNCTUATION,HASDASH	0.026504
IS_NUM	0.026504
ALL_CAPS	0.026504
IS_CAPITALIZED	0.141170
STOP_WORD	0.145677
EMOTICON	0.149898
END=ed,END=ing,END=ly,END=es,END=ss	0.152031
Prefix and Suffixes	0.156531
Company	0.182567
Facility	0.214533
Geo loc	0.240011
Person	0.264514
Product	0.283212



We can see the basic features at the start hasn't contributed much as the features can't contribute to strengthen the feature weight as features are not related to our label set
But the Feature starting with capital which is the main characteristic of all the labels has increased score from 0.02 to 0.14 and next also the same score trend continued but a increase with almost similar width can be seen as they are the features associated with label directly.

Resources:

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Je rey Flanigan, and Noah A Smith.
Part-of-speech tagging for twitter: Annotation, features, and experiments. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies
https://github.com/chbrown/nlp/blob/master/src/main/java/nlp/lm/HMMSimpleTagger.java