Junyang Xin

N11176767

In my solution to this assignment, I have included 8 feature sets. They are listed below:

Feature set 1:

current=currentToken pos=currentTokenPOS

Feature set 2:

current=currentToken pos=currentTokenPos isInitialCap=True/False

Feature set 3:

prev=previousToken prevPos=previousTokenPOS current=currentToken pos=currentTokenPos isInitialCap=True/False next=nextToken nextPos=nextTokenPOS

Feature set 4:

prev=previousToken prevPos=previousTokenPOS current=currentToken pos=currentTokenPos isInitialCap=True/False isSuffixIng=True/False next=nextToken nextPos=nextTokenPOS

Feature set 5:

prev=previousToken prevPos=previousTokenPOS current=currentToken pos=currentTokenPos isInitialCap=True/False isSuffixIng=True/False isAllUpper=True/False isAllLower=True/False isMixedCase=True/False capInitWithPeriod=True/False endWithDigit=True/False hasHyphen=True/False next=nextToken nextPos=nextTokenPOS

Feature set 6:

prev=previousToken prevPos=previousTokenPOS current=currentToken pos=currentTokenPos isInitialCap=True/False isSuffixIng=True/False isAllUpper=True/False isAllLower=True/False isMixedCase=True/False capInitWithPeriod=True/False endWithDigit=True/False hasHyphen=True/False stemmed=stemmedCurrentToken next=nextToken nextPos=nextTokenPOS

Feature set 7:

prev=previousToken prevPos=previousTokenPOS prevTag=previousTokenTag current=currentToken pos=currentTokenPos isInitialCap=True/False isSuffixIng=True/False isAllUpper=True/False isAllLower=True/False isMixedCase=True/False capInitWithPeriod=True/False endWithDigit=True/False hasHyphen=True/False stemmed=stemmedCurrentToken next=nextToken nextPos=nextTokenPOS nextTag=nextTokenTag

Feature set 8:

prev=previousToken prevPos=previousTokenPOS prevTag=previousTokenTag current=currentToken pos=currentTokenPos isInitialCap=True/False isSuffixIng=True/False next=nextToken nextPos=nextTokenPOS nextTag=nextTokenTag

Please note that during tagging, the nextTag is always set to null to mimic the fact that only prior state has been seen.

Besides the MaxEnt tagging, I also have implemented 2 versions of Viterbi and applied them with model trained with features set 8, aka. model8. The complex one should strictly mimic the MEMM in the textbook, and the simple one has some relaxation. The simple one is not 100% correct viterbi implementation (the previous state is always set to null) so it is more or less a placeholder simply for performance comparison purpose. However its performance is not bad at all.

Below are the results I got for each feature set. The measures I used are **number of correct tags** and **precision, recall, F0.5, F1, F2** of ngroup chunking.

Model1: correct tags 9186

Model2: correct tags 9186

Model3: correct tags 9356

Model4: correct tags 9355

Model5: correct tags 9355

Model6: correct tags 9351

Model7: correct tags 9345

Model8: correct tags 9325

Model8 with Viterbi-simple: correct tags 9319

Model8 with Viterbi-complex: correct tags 9331

Measures of model1: precision:0.8182861514919664 recall:0.8404715127701375 F1:0.8292304710215158 F0.5:0.8226290285362664 F2:0.8359387212755979

Measures of model2: precision:0.8182861514919664 recall:0.8404715127701375 F1:0.8292304710215158 F0.5:0.8226290285362664 F2:0.8359387212755979

Measures of model3: precision:0.8953307392996109 recall:0.9041257367387033 F1:0.8997067448680351 F0.5:0.8970760233918128 F2:0.9023529411764706

Measures of model4: precision:0.8954758190327613 recall:0.9021611001964637 F1:0.898806028577021 F0.5:0.896804937114288 F2:0.9008160703075958

Measures of model5: precision:0.8946957878315133 recall:0.9013752455795678 F1:0.898023096496379 F0.5:0.8960237481446761 F2:0.9000313873195229

Measures of model6: precision:0.8924814959096221 recall:0.900196463654224 F1:0.8963223787167449 F0.5:0.8940138921407943 F2:0.8986428179179414

Measures of model7: precision:0.8895729126587149 recall:0.9084479371316306 F1:0.8989113530326595 F0.5:0.8932849084305696 F2:0.904609124344628

Measures of model8: precision:0.8867996930161166 recall:0.9080550098231827 F1:0.8973014948553678 F0.5:0.8909707764669595 F2:0.9037228218363835

Measures of model8 with Viterbi-simple: precision:0.880061115355233 recall:0.9053045186640472 F1:0.8925043579314351 F0.5:0.8849965429822539 F2:0.90014064697609

Measures of model8 with Viterbi-complex: precision:0.8875239923224568 recall:0.9084479371316306 F1:0.8978640776699028 F0.5:0.891631315079059 F2:0.9041845913179508

**Conclusion**: Feature conjunction does help improved the performance. Including prior state as a feature also helped improve the performance. The best performance I could get is with feature set 8 using MaxEnt and Viterbi-complex using feature set 8. Viterbi did improved the performance but just a little bit.