Approaches used and comparisons between CRF++ and HMM models

Hidden Markov Models (HMM):

Hidden Markov Models (HMMs) have been widely used in various sequence modeling tasks, including part-of-speech (POS) tagging. HMMs are based on the principle of modeling the probability distribution over sequences of observations (words) and hidden states (POS tags).

One of the core assumptions of HMMs is the Markov property, which states that the probability of a particular state (POS tag) at time t depends only on the state at time t-1. In other words, HMMs assume that the current state is independent of all previous states given the state at the previous time step, making them first-order Markov models. This assumption simplifies the modeling process and reduces the computational complexity of training and inference.

HMM-based POS tagging typically involves two sets of probabilities: transition probabilities and emission probabilities. Transition probabilities describe the likelihood of transitioning from one POS tag to another, while emission probabilities model the likelihood of observing a word given a particular POS tag. These probabilities are estimated from annotated training data using techniques such as maximum likelihood estimation (MLE) or smoothing methods like Laplace smoothing or Good-Turing smoothing.

During inference, the Viterbi algorithm is commonly employed to find the most likely sequence of hidden states (POS tags) given the observed sequence of words. The algorithm efficiently computes the most probable sequence by considering both the transition probabilities between POS tags and the emission probabilities of observing words given POS tags.

While HMMs have been widely used for POS tagging and other sequence labeling tasks due to their simplicity and computational efficiency, they do have limitations. One major

limitation is their inability to capture long-range dependencies and complex interactions between distant words in a sequence. Additionally, HMMs rely on the independence assumption between observations, which may not hold true in all cases.

Conditional Random Fields (CRF):

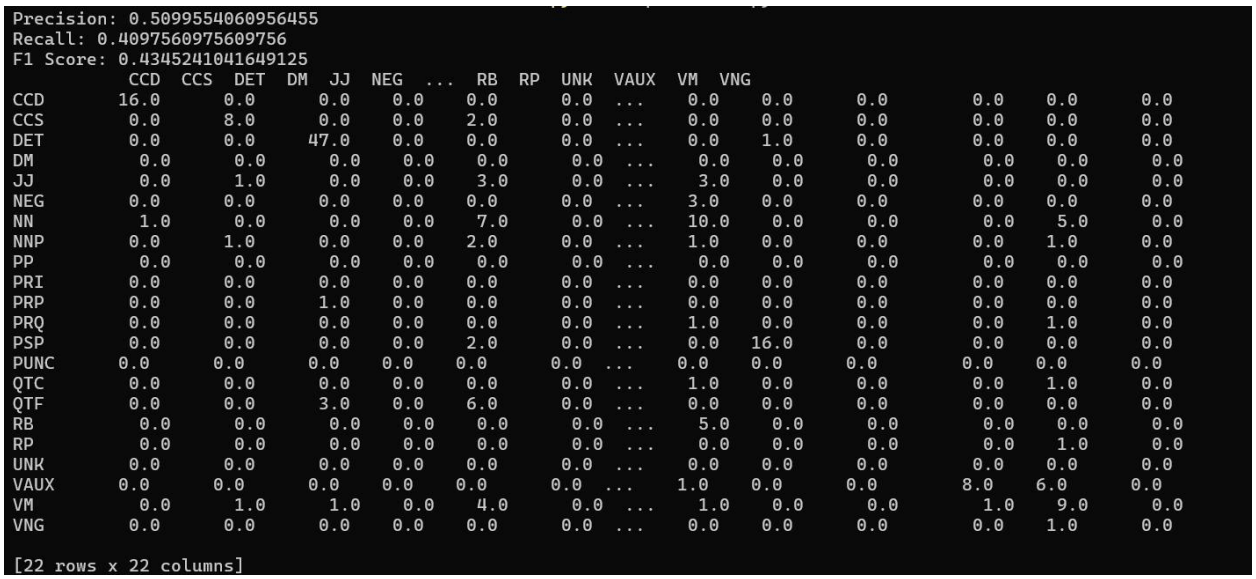
Conditional Random Fields (CRFs) are discriminative sequence models that directly model the conditional probability of a sequence of labels (POS tags) given an input sequence of observations (words). Unlike HMMs, CRFs do not make explicit assumptions about the independence of observations or impose constraints on the transition probabilities between labels.

CRFs model dependencies between neighboring labels directly by incorporating a set of feature functions that capture relevant information from the input sequence. These features may include word identities, suffixes, prefixes, capitalization patterns, contextual information from neighboring words, and more. The model parameters (weights) associated with these features are learned from annotated training data using optimization algorithms such as gradient descent or L-BFGS.

One of the key advantages of CRFs over HMMs is their ability to capture complex dependencies and interactions between features, leading to improved performance, especially in tasks requiring context sensitivity. By modeling dependencies beyond adjacent words, CRFs can effectively capture long-range dependencies and contextual information, resulting in more accurate POS tagging.

Observations (Performance of CRF++ model vs HMM model with regards to various language sample data)

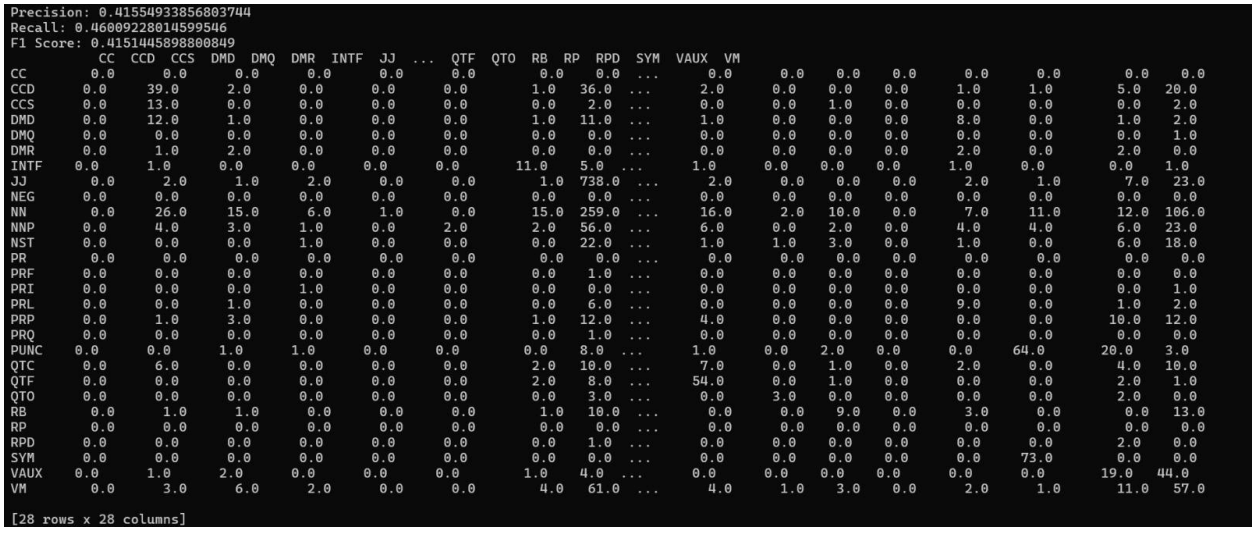
1. CRF++ English



```
Precision: 0.5099554060956455
Recall: 0.4097560975609756
F1 Score: 0.4345241041649125
          CCD  CCS  DET  DM  JJ  NEG  ...  RB  RP  UNK  VAUX  VM  VNG
CCD      16.0  0.0  0.0  0.0  0.0  0.0  ...  0.0  0.0  0.0  0.0  0.0  0.0
CCS       0.0  8.0  0.0  0.0  2.0  0.0  ...  0.0  0.0  0.0  0.0  0.0  0.0
DET       0.0  0.0 47.0  0.0  0.0  0.0  ...  0.0  1.0  0.0  0.0  0.0  0.0
DM        0.0  0.0  0.0  0.0  0.0  0.0  ...  0.0  0.0  0.0  0.0  0.0  0.0
JJ        0.0  1.0  0.0  0.0  3.0  0.0  ...  3.0  0.0  0.0  0.0  0.0  0.0
NEG       0.0  0.0  0.0  0.0  0.0  0.0  ...  3.0  0.0  0.0  0.0  0.0  0.0
NN        1.0  0.0  0.0  0.0  7.0  0.0  ... 10.0  0.0  0.0  0.0  5.0  0.0
NNP       0.0  1.0  0.0  0.0  2.0  0.0  ...  1.0  0.0  0.0  0.0  1.0  0.0
PP        0.0  0.0  0.0  0.0  0.0  0.0  ...  0.0  0.0  0.0  0.0  0.0  0.0
PRI       0.0  0.0  0.0  0.0  0.0  0.0  ...  0.0  0.0  0.0  0.0  0.0  0.0
PRP       0.0  0.0  1.0  0.0  0.0  0.0  ...  0.0  0.0  0.0  0.0  0.0  0.0
PRQ       0.0  0.0  0.0  0.0  0.0  0.0  ...  1.0  0.0  0.0  0.0  1.0  0.0
PSP       0.0  0.0  0.0  0.0  2.0  0.0  ...  0.0 16.0  0.0  0.0  0.0  0.0
PUNC      0.0  0.0  0.0  0.0  0.0  0.0  ...  0.0  0.0  0.0  0.0  0.0  0.0
QTC       0.0  0.0  0.0  0.0  0.0  0.0  ...  1.0  0.0  0.0  0.0  1.0  0.0
QTF       0.0  0.0  3.0  0.0  6.0  0.0  ...  0.0  0.0  0.0  0.0  0.0  0.0
RB        0.0  0.0  0.0  0.0  0.0  0.0  ...  5.0  0.0  0.0  0.0  0.0  0.0
RP        0.0  0.0  0.0  0.0  0.0  0.0  ...  0.0  0.0  0.0  0.0  1.0  0.0
UNK       0.0  0.0  0.0  0.0  0.0  0.0  ...  0.0  0.0  0.0  0.0  0.0  0.0
VAUX      0.0  0.0  0.0  0.0  0.0  0.0  ...  1.0  0.0  0.0  8.0  6.0  0.0
VM        0.0  1.0  1.0  0.0  4.0  0.0  ...  1.0  0.0  0.0  1.0  9.0  0.0
VNG       0.0  0.0  0.0  0.0  0.0  0.0  ...  0.0  0.0  0.0  0.0  1.0  0.0

[22 rows x 22 columns]
```

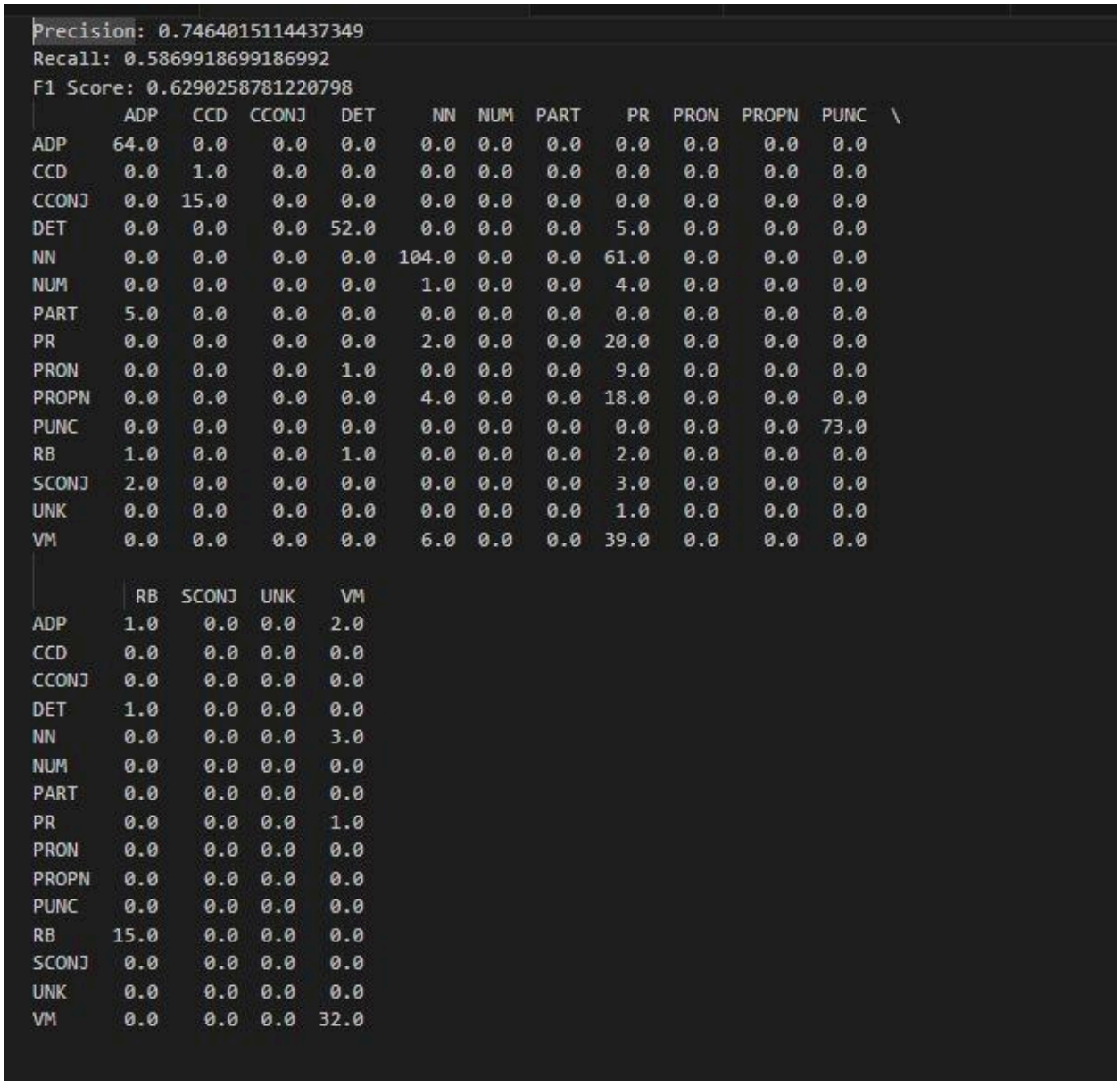
2. CRF++ Marathi



```
Precision: 0.41554933856803744
Recall: 0.46009228014599546
F1 Score: 0.4151445898800849
          CC  CCD  CCS  DMD  DMQ  DMR  INTF  JJ  ...  QTF  QTO  RB  RP  RPD  SYM  VAUX  VM
CC       0.0  0.0  0.0  0.0  0.0  0.0   0.0  0.0  ...  0.0  0.0  0.0  0.0  0.0  0.0   0.0  0.0
CCD      0.0 39.0  2.0  0.0  0.0  0.0   1.0 36.0  ...  2.0  0.0  0.0  0.0  1.0  1.0   5.0 20.0
CCS      0.0 13.0  0.0  0.0  0.0  0.0   0.0  2.0  ...  0.0  0.0  1.0  0.0  0.0  0.0   0.0  2.0
DMD      0.0 12.0  1.0  0.0  0.0  0.0   1.0 11.0  ...  1.0  0.0  0.0  0.0  8.0  0.0   1.0  2.0
DMQ      0.0  0.0  0.0  0.0  0.0  0.0   0.0  0.0  ...  0.0  0.0  0.0  0.0  0.0  0.0   0.0  1.0
DMR      0.0  1.0  2.0  0.0  0.0  0.0   0.0  0.0  ...  0.0  0.0  0.0  0.0  2.0  0.0   2.0  0.0
INTF     0.0  1.0  0.0  0.0  0.0  0.0  11.0  5.0  ...  1.0  0.0  0.0  0.0  1.0  0.0   0.0  1.0
JJ       0.0  2.0  1.0  2.0  0.0  0.0   1.0 738.0  ...  2.0  0.0  0.0  0.0  2.0  1.0   7.0 23.0
NEG      0.0  0.0  0.0  0.0  0.0  0.0   0.0  0.0  ...  0.0  0.0  0.0  0.0  0.0  0.0   0.0  0.0
NN       0.0 26.0 15.0  6.0  1.0  0.0  15.0 259.0  ... 16.0  2.0 10.0  0.0  7.0 11.0  12.0 106.0
NNP      0.0  4.0  3.0  1.0  0.0  2.0   2.0 56.0  ...  6.0  0.0  2.0  0.0  4.0  4.0   6.0 23.0
NST      0.0  0.0  0.0  1.0  0.0  0.0   0.0 22.0  ...  1.0  1.0  3.0  0.0  1.0  0.0   6.0 18.0
PR       0.0  0.0  0.0  0.0  0.0  0.0   0.0  0.0  ...  0.0  0.0  0.0  0.0  0.0  0.0   0.0  0.0
PRF      0.0  0.0  0.0  0.0  0.0  0.0   0.0  1.0  ...  0.0  0.0  0.0  0.0  0.0  0.0   0.0  0.0
PRI      0.0  0.0  0.0  1.0  0.0  0.0   0.0  0.0  ...  0.0  0.0  0.0  0.0  0.0  0.0   0.0  1.0
PRL      0.0  0.0  1.0  0.0  0.0  0.0   0.0  6.0  ...  0.0  0.0  0.0  0.0  9.0  0.0   1.0  2.0
PRP      0.0  1.0  3.0  0.0  0.0  0.0   1.0 12.0  ...  4.0  0.0  0.0  0.0  0.0  0.0  10.0 12.0
PRQ      0.0  0.0  0.0  0.0  0.0  0.0   0.0  1.0  ...  0.0  0.0  0.0  0.0  0.0  0.0   0.0  0.0
PUNC     0.0  0.0  1.0  1.0  0.0  0.0   0.0  8.0  ...  1.0  0.0  2.0  0.0  0.0 64.0  20.0  3.0
QTC      0.0  6.0  0.0  0.0  0.0  0.0   2.0 10.0  ...  7.0  0.0  1.0  0.0  2.0  0.0   4.0 10.0
QTF      0.0  0.0  0.0  0.0  0.0  0.0   2.0  8.0  ... 54.0  0.0  1.0  0.0  0.0  0.0   2.0  1.0
QTO      0.0  0.0  0.0  0.0  0.0  0.0   0.0  3.0  ...  0.0  3.0  0.0  0.0  0.0  0.0   2.0  0.0
RB       0.0  1.0  1.0  0.0  0.0  0.0   1.0 10.0  ...  0.0  0.0  9.0  0.0  3.0  0.0   0.0 13.0
RP       0.0  0.0  0.0  0.0  0.0  0.0   0.0  0.0  ...  0.0  0.0  0.0  0.0  0.0  0.0   0.0  0.0
RPD      0.0  0.0  0.0  0.0  0.0  0.0   0.0  1.0  ...  0.0  0.0  0.0  0.0  0.0  0.0   2.0  0.0
SYM      0.0  0.0  0.0  0.0  0.0  0.0   0.0  0.0  ...  0.0  0.0  0.0  0.0  0.0 73.0   0.0  0.0
VAUX     0.0  1.0  2.0  0.0  0.0  0.0   1.0  4.0  ...  0.0  0.0  0.0  0.0  0.0  0.0  19.0 44.0
VM       0.0  3.0  6.0  2.0  0.0  0.0   4.0 61.0  ...  4.0  1.0  3.0  0.0  2.0  1.0  11.0 57.0

[28 rows x 28 columns]
```

3. HMM English



```
Precision: 0.7464015114437349
Recall: 0.5869918699186992
F1 Score: 0.6290258781220798
        ADP   CCD  CCONJ   DET     NN  NUM  PART    PR  PRON  PROPN  PUNC  \
ADP    64.0   0.0    0.0   0.0    0.0  0.0   0.0   0.0   0.0    0.0   0.0
CCD     0.0   1.0    0.0   0.0    0.0  0.0   0.0   0.0   0.0    0.0   0.0
CCONJ   0.0  15.0    0.0   0.0    0.0  0.0   0.0   0.0   0.0    0.0   0.0
DET     0.0   0.0    0.0  52.0    0.0  0.0   0.0   5.0   0.0    0.0   0.0
NN      0.0   0.0    0.0   0.0  104.0  0.0   0.0  61.0   0.0    0.0   0.0
NUM     0.0   0.0    0.0   0.0    1.0  0.0   0.0   4.0   0.0    0.0   0.0
PART    5.0   0.0    0.0   0.0    0.0  0.0   0.0   0.0   0.0    0.0   0.0
PR      0.0   0.0    0.0   0.0    2.0  0.0   0.0  20.0   0.0    0.0   0.0
PRON    0.0   0.0    0.0   1.0    0.0  0.0   0.0   9.0   0.0    0.0   0.0
PROPN   0.0   0.0    0.0   0.0    4.0  0.0   0.0  18.0   0.0    0.0   0.0
PUNC    0.0   0.0    0.0   0.0    0.0  0.0   0.0   0.0   0.0    0.0  73.0
RB      1.0   0.0    0.0   1.0    0.0  0.0   0.0   2.0   0.0    0.0   0.0
SCONJ   2.0   0.0    0.0   0.0    0.0  0.0   0.0   3.0   0.0    0.0   0.0
UNK     0.0   0.0    0.0   0.0    0.0  0.0   0.0   1.0   0.0    0.0   0.0
VM      0.0   0.0    0.0   0.0    6.0  0.0   0.0  39.0   0.0    0.0   0.0

        RB  SCONJ  UNK    VM
ADP    1.0    0.0  0.0   2.0
CCD    0.0    0.0  0.0   0.0
CCONJ  0.0    0.0  0.0   0.0
DET    1.0    0.0  0.0   0.0
NN     0.0    0.0  0.0   3.0
NUM    0.0    0.0  0.0   0.0
PART   0.0    0.0  0.0   0.0
PR     0.0    0.0  0.0   1.0
PRON   0.0    0.0  0.0   0.0
PROPN  0.0    0.0  0.0   0.0
PUNC   0.0    0.0  0.0   0.0
RB    15.0    0.0  0.0   0.0
SCONJ  0.0    0.0  0.0   0.0
UNK    0.0    0.0  0.0   0.0
VM     0.0    0.0  0.0  32.0
```

4. HMM Marathi

```
Precision: 0.6236160959179685
Recall: 0.5584326148336891
F1 Score: 0.5585092561036178
          CC    CCD   CCS   DMD  DMQ   DMR  INTF     JJ  NEG    NST  ...    QTF  \
CC       0.0    1.0   0.0   0.0  0.0   0.0   0.0    0.0  0.0    0.0  ...    0.0
CCD      0.0  219.0  11.0   0.0  0.0   0.0   0.0    0.0  0.0    0.0  ...    0.0
CCS      0.0   16.0  38.0   0.0  0.0   0.0   0.0    0.0  0.0    0.0  ...    0.0
DMD      0.0    0.0   0.0  79.0  0.0  22.0   0.0    4.0  0.0    0.0  ...    0.0
DMQ      0.0    0.0   0.0   0.0  0.0   0.0   0.0    0.0  0.0    0.0  ...    0.0
DMR      0.0    2.0   0.0  15.0  0.0   2.0   0.0    0.0  0.0    0.0  ...    0.0
INTF     0.0    0.0   0.0   0.0  0.0   0.0  11.0   10.0  0.0    0.0  ...   49.0
JJ       0.0    0.0   0.0   4.0  0.0   0.0   0.0  676.0  0.0   13.0  ...   31.0
NEG      0.0    0.0   0.0   0.0  0.0   0.0   0.0    0.0  2.0    0.0  ...    0.0
NST      0.0    0.0   0.0   0.0  0.0   0.0   0.0    4.0  0.0  104.0  ...    0.0
PR       0.0    0.0   0.0   0.0  0.0   0.0   0.0    0.0  0.0    0.0  ...    0.0
PRF      0.0    0.0   0.0   0.0  0.0   0.0   0.0    0.0  0.0    0.0  ...    0.0
PRI      0.0    0.0   0.0   0.0  0.0   0.0   0.0    0.0  0.0    0.0  ...    0.0
PRL      0.0    0.0   0.0   1.0  0.0   0.0   0.0    0.0  0.0    0.0  ...    0.0
PRP      0.0    3.0   0.0   1.0  0.0   0.0   0.0    0.0  0.0    0.0  ...    0.0
PRQ      0.0    0.0   0.0   0.0  1.0   0.0   0.0    0.0  0.0    0.0  ...    0.0
PUNC     0.0    0.0   0.0   0.0  0.0   0.0   0.0    0.0  0.0    0.0  ...    0.0
QTC      0.0    0.0   0.0   0.0  0.0   0.0   0.0    1.0  0.0    0.0  ...    8.0
QTF      0.0    0.0   0.0   0.0  0.0   0.0   1.0    4.0  0.0    0.0  ...  123.0
QTO      0.0    1.0   0.0   1.0  0.0   0.0   0.0   15.0  0.0    5.0  ...    2.0
QTOP     0.0    0.0   0.0   0.0  0.0   0.0   0.0    2.0  0.0    0.0  ...    0.0
RB       0.0    0.0   0.0  14.0  0.0   0.0   0.0   11.0  0.0    8.0  ...   10.0
RP       0.0    0.0   0.0   0.0  0.0   0.0   0.0    0.0  0.0    0.0  ...    0.0
RPD      0.0    0.0   4.0   0.0  0.0   0.0   0.0    0.0  0.0    0.0  ...    0.0
SYM      0.0    0.0   0.0   0.0  0.0   0.0   0.0    0.0  0.0    0.0  ...    0.0
VAUX     0.0    0.0   0.0   0.0  0.0   0.0   0.0    1.0  1.0    0.0  ...    0.0
VM       0.0    0.0   0.0   0.0  0.0   0.0   0.0    1.0  0.0    1.0  ...    0.0
वाजता   0.0    0.0   0.0   0.0  0.0   0.0   0.0    0.0  0.0    0.0  ...    0.0

            QTO  QTOP    RB   RP  RPD    SYM  VAUX      VM  वाजता
CC          0.0   0.0   0.0  0.0  0.0    0.0   0.0     0.0    0.0
CCD        56.0   0.0   0.0  0.0  3.0    0.0   0.0     2.0    0.0
CCS        35.0   0.0   0.0  0.0  6.0    0.0   0.0     0.0    0.0
DMD        72.0   0.0   0.0  0.0  0.0    0.0   0.0     3.0    0.0
DMQ         1.0   0.0   0.0  0.0  0.0    0.0   0.0     0.0    0.0
DMR        15.0   0.0   0.0  0.0  0.0    0.0   0.0     1.0    0.0
INTF       12.0   0.0   0.0  0.0  0.0    0.0   0.0     0.0    0.0
JJ        617.0   0.0   4.0  0.0  0.0    0.0   0.0     0.0    0.0
NEG         0.0   0.0   0.0  0.0  0.0    0.0   0.0     0.0    0.0
NST       169.0   0.0   1.0  0.0  0.0    0.0   0.0     0.0    0.0
PR          1.0   0.0   0.0  0.0  0.0    0.0   0.0     0.0    0.0
PRF         8.0   0.0   0.0  0.0  0.0    0.0   0.0     0.0    0.0
PRI         1.0   0.0   0.0  0.0  0.0    0.0   0.0     0.0    0.0
PRL        57.0   0.0   0.0  0.0  0.0    0.0   0.0     0.0    0.0
PRP       140.0   0.0   0.0  0.0  0.0    0.0   0.0     0.0    0.0
PRQ         9.0   0.0   0.0  0.0  0.0    0.0   0.0     0.0    0.0
PUNC       25.0   0.0   0.0  0.0  0.0  119.0   0.0     0.0    0.0
QTC       103.0   0.0   0.0  0.0  0.0    0.0   0.0     0.0    0.0
QTF        23.0   0.0   0.0  0.0  0.0    0.0   0.0     0.0    0.0
QTO      3348.0   0.0   2.0  0.0  0.0    0.0   1.0     3.0    0.0
QTOP     1560.0   0.0   0.0  0.0  0.0    0.0   0.0     0.0    0.0
RB         87.0   0.0  11.0  0.0  0.0    0.0   3.0     2.0    0.0
RP          0.0   0.0   0.0  0.0  1.0    0.0   0.0     0.0    0.0
RPD         7.0   0.0   0.0  0.0  1.0    0.0   0.0     0.0    0.0
SYM         2.0   0.0   0.0  0.0  0.0   45.0   0.0     0.0    0.0
VAUX       52.0   0.0   0.0  0.0  0.0    0.0  74.0   377.0    0.0
VM        549.0   0.0   0.0  0.0  0.0    0.0  28.0  1242.0    0.0
वाजता      0.0   0.0   0.0  0.0  0.0    1.0   0.0     0.0    0.0

[28 rows x 28 columns]
```

Comparative Study and Observations:

1. Model Complexity and Flexibility: HMMs are generally simpler and have fewer parameters compared to CRFs. CRFs are more flexible and can capture complex dependencies between features, leading to potentially higher accuracy.

2. Training Efficiency: Due to their simpler structure, HMMs are often faster to train compared to CRFs. However, CRFs may require more computational resources and time for training, especially when dealing with large feature sets and complex interactions.

3. Performance: CRFs tend to outperform HMMs in many cases, particularly in tasks where context sensitivity and capturing long-range dependencies are crucial. However, the performance difference between the two models may vary depending on the specific characteristics of the data and the complexity of the task.

4. Data Requirements: CRFs may require more annotated training data to effectively learn complex feature interactions compared to HMMs, which rely on simpler probabilistic assumptions. However, with sufficient training data, CRFs have the potential to achieve higher accuracy by leveraging more information from the input sequence.

Overall Analysis

Due to its greater precision, recall and F1, and also due to other technical factors, HMM turns out to be a better method when it comes to the models I have developed, and how they work for my datasets.