NATURAL LANGUAGE PROCESSING

Chunking using CRF Model and Rule Based Model

For MALAYALAM DATA

A.Hari Narayana (201201111) G.Sai Vishwa Teja (201201176)

Abstract: Conditional Random Fields (CRFs) is a framework for building probabilistic models to segment and label sequence data. CRFs offer several advantages over hidden Markov models (HMMs) and stochastic grammars for such tasks, including the ability to relax strong independence assumptions made in those models. Here we have implemented chunking using various templates for CRF and another model which is rule based. In the Rule Based Model based on the train data we implemented chunking was performed. The best template in the CRF chunking gave about 91.66 % accuracy and Rule based model produced an accuracy about 73.1 %.

Keywords: Conditional Random Fields, Chunking, Language Models, Rule Based Model, Accuracy.

Introduction:

Chunking or shallow parsing is the task of identifying and segmenting the text into syntactically correlated word groups like noun phrase, verb phrase etc. It is considered as an intermediate step towards full parsing. This projects presents the exploitation of CRFs for Chunking of Malayalam language. A rule based chunker is also used for chunking the data where the rules are written using the train data. Then we compare the results of each chunker.

1. Conditional Random Fields:

Conditional random fields (CRFs) are undirected graphical models developed for labeling sequence data [19]. CRFs directly model $p(x \mid z)$, the *conditional* distribution over the hidden variables x given observations z. This model is different from generative models such as Hidden Markov Models or Markov Random Fields, which apply Bayes rule to infer hidden states [20]. CRFs can handle arbitrary dependencies between the observations z, which gives them substantial flexibility in using high-dimensional feature vectors.

The nodes in a CRF represent hidden states, denoted $x = < x_1, x_2, \ldots, x_n >$, and data, denoted z. The nodes x_i, along with the connectivity structure represented by the undirected edges between them, define the conditional distribution $p(x \mid z)$ over the hidden states x. Let C be the set of cliques (fully connected subsets) in the graph of a CRF. Then, a CRF factorizes the conditional distribution into a product of *clique potentials* $\phi_c(z, {}^{x}_{c})$, where every c ε C is a clique in the graph and z and ${}^{x}_{c}$ are the observed data and the hidden nodes in the clique c, respectively. Clique potentials are functions that map variable configurations to non-negative numbers. Intuitively, a potential captures

the "compatibility" among the variables in the clique: the larger the potential value, the more likely the configuration. Using clique potentials, the conditional distribution over hidden states is written as

$$p(x \mid z) = 1/Z(z)\left(\prod_{c \in C} \phi_c(z, x_c)\right) \quad (1)$$

where $Z(z) = \sum_{x} \prod_{c \in C} \phi_c(z, x_c)$ is the normalizing partition function.

The computation of this partition function can be exponential in the size of x. Hence, exact inference is possible for a limited class of CRF models only. Potentials $\phi_c(z, x_c)$ are described by log-linear combinations of *feature functions* f_c i.*e,*

$\phi_c(z, x_c) = \exp(w_c^T . f_c(z, x_c))$ --- (2) where w_c^T is called as a weight vector.

$f_c(z, x_c)$ is a function that extracts a vector of features from the variable values. Using feature functions, we rewrite the conditional distribution (1) as

$$P(x|z) = 1/Z(z) \exp\{\sum_{\mathbf{c \in C}} Wc^T . f_c(Z, X_c)\}$$

OUTPUTS:

Here we ran it for three different template files and given below are outputs and accuracies. These were calculated using the software CRF ++ which takes a train data and an feature file and generates a model file which will be used for testing test data. Following are the results of various feature files used in CRF++.

Template 1:

#Unigram
U00:%x[-2,0]
U01:%x[-1,0]

Results:

accuracy: 49.46%; **precision**: 36.14%; **recall**: 39.92%; FB1: 37.94
BLK: precision: 18.23%; recall: 68.63%; FB1: 28.81 192
CCP: precision: 0.00%; recall: 0.00%; FB1: 0.00 5
NP: precision: 48.58%; recall: 49.68%; FB1: 49.12 317
RBP: precision: 0.00%; recall: 0.00%; FB1: 0.00 0
VGF: precision: 38.10%; recall: 14.81%; FB1: 21.33 21
VGINF: precision: 37.50%; recall: 23.08%; FB1: 28.57 8
VGNF: precision: 42.86%; recall: 10.34%; FB1: 16.67 14
VGNN: precision: 0.00%; recall: 0.00%; FB1: 0.00 13

First 5 sentences:

1) [പോരായ്മകള്ക്ക്] [ഉള്ള] [പരിഹാരങ്ങളും] [കണ്ടെത്തിയിരുന്നു] [।]
2) [അതിന്റെ ഫലം] [ആയി "] [തോമസ് കുക്ക്] [റെയില്വേകൂപ്പണും "] [ഹോട്ടല്കൂപ്പണും] [ആവിഷ്ക്കരിക്കപ്പെട്ടു] [।]
3) [1866-ല്] [കുക്ക്] [അമേരിക്കയില്] [എത്തി] [।]
4) [ആ വര്ഷം തന്നെ] [അമേരിക്കയിലേയ്ക്ക്] [ഉള്ള] [യാത്രാസംഘത്തെ] [അയയ്ക്കാനും] [സാധിച്ചു] [।]
5) [യാത്രകള്] [വിപുലവും] [പ്രവര്ത്തനങ്ങള്] [വ്യാപകവും] [ആയതോടെ] [1867-ല്] [തോമസ് കുക്ക് ആന്ഡ് സണ്] [എന്ന്] [ഒരു കമ്പനി] [സ്ഥാപിക്കപ്പെട്ടു] [।]

Template 2:

Unigram
U00:%x[-2,0]
U01:%x[-1,0]
U02:%x[0,0]
U03:%x[1,0]

Results :

accuracy: 74.11%; **precision**: 64.88%; **recall**: 65.50%; FB1: 65.19
BLK: precision: 53.12%; recall: 100.00%; FB1: 69.39 96
CCP: precision: 91.67%; recall: 84.62%; FB1: 88.00 12
NP: precision: 66.91%; recall: 59.35%; FB1: 62.91 275
RBP: precision: 0.00%; recall: 0.00%; FB1: 0.00 7

VGF: precision: 86.21%; recall: 92.59%; FB1: 89.29 58
VGINF: precision: 15.00%; recall: 23.08%; FB1: 18.18 20
VGNF: precision: 86.05%; recall: 63.79%; FB1: 73.27 43
VGNN: precision: 20.00%; recall: 13.33%; FB1: 16.00 10

First 5 sentences:

1) [പോരായ്മകള്ക്ക്] [ഉള്ള] [പരിഹാരങ്ങളും] [കണ്ടെത്തിയിരുന്നു] [।]
2) [അതിന്റെ ഫലം] [ആയി "] [തോമസ് കുക്ക്] [റെയില്വേകൂപ്പണും "] [ഹോട്ടല്കൂപ്പണും] [ആവിഷ്ക്കരിക്കപ്പെട്ടു] [।]
3) [1866-ല്] [കുക്ക്] [അമേരിക്കയില്] [എത്തി] [।]
4) [ആ വർഷം തന്നെ] [അമേരിക്കയിലേയ്ക്ക്] [ഉള്ള] [യാത്രാസംഘത്തെ] [അയയ്ക്കാനും] [സാധിച്ചു] [।]
5) [യാത്രകള്] [വിപുലവും] [പ്രവർത്തനങ്ങള്] [വ്യാപകവും] [ആയതോടെ] [1867-ല്] [തോമസ് കുക്ക് ആന്ഡ് സണ്] [എന്ന്] [ഒരു കമ്പനി] [സ്ഥാപിക്കപ്പെട്ടു] [।]

Template 3 :

Unigram
U00:%x[-2,0]
U01:%x[-1,0]
U02:%x[0,0]
U03:%x[1,0]
U04:%x[2,0]
U05:%x[-1,0]/%x[0,0]
U06:%x[0,0]/%x[1,0]
U10:%x[-2,1]
U11:%x[-1,1]
U12:%x[0,1]
U13:%x[1,1]
U14:%x[2,1]
U15:%x[-2,1]/%x[-1,1]
U16:%x[-1,1]/%x[0,1]
U17:%x[0,1]/%x[1,1]
U18:%x[1,1]/%x[2,1]
U20:%x[-2,1]/%x[-1,1]/%x[0,1]
U21:%x[-1,1]/%x[0,1]/%x[1,1]
U22:%x[0,1]/%x[1,1]/%x[2,1]

RESULTS :

accuracy: 91.78%; **precision**: 87.64%; **recall**: 87.98%; FB1: 87.81
BLK: precision: 98.08%; recall: 100.00%; FB1: 99.03 52
CCP: precision: 100.00%; recall: 84.62%; FB1: 91.67 11
NP: precision: 84.11%; recall: 81.94%; FB1: 83.01 302

RBP: precision: 100.00%; recall: 50.00%; FB1: 66.67 1
VGF: precision: 96.36%; recall: 98.15%; FB1: 97.25 55
VGINF: precision: 85.71%; recall: 92.31%; FB1: 88.89 14
VGNF: precision: 90.62%; recall: 100.00%; FB1: 95.08 64
VGNN: precision: 73.68%; recall: 93.33%; FB1: 82.35 19

First 5 sentences:

1) [പോരായ്മകള്ക്ക്] [ഉള്ള] [പരിഹാരങ്ങളും] [കണ്ടെത്തിയിരുന്നു] [।]
2) [അതിന്റെ] [ഫലം] [ആയി "] [തോമസ് കുക്ക്] [റെയില്വേകൂപ്പണും "] [ഹോട്ടല്കൂപ്പണും] [ആവിഷ്ക്കരിക്കപ്പെട്ടു] [।]
3) [1866-ല്] [കുക്ക്] [അമേരിക്കയില്] [എത്തി] [।]
4) [ആ വര്ഷം തന്നെ] [അമേരിക്കയിലേയ്ക്ക്] [ഉള്ള] [യാത്രാസംഘത്തെ] [അയയ്ക്കാനും] [സാധിച്ചു] [।]
5) [യാത്രകള്] [വിപുലവും] [പ്രവര്ത്തനങ്ങള്] [വ്യാപകവും] [ആയതോടെ] [1867-ല്] [തോമസ് കുക്ക്] [ആന്ഡ് സണ് എന്ന്] [ഒരു കമ്പനി] [സ്ഥാപിക്കപ്പെട്ടു] [।]

OBSERVATIONS :

As we see here based on the feature template modifications the accuracy in our results changes. The state features and transition features in the feature file has a huge effect. If we take more features we are getting better results.

2) RULE BASED CHUNKING

The procedure here we follow is observe the train data and write all the possible chunks and define the rules. Based on these observations and rules the test data will be chunked and will give the output. Here an algorithm runs which will calculate the best chunk in a sequence of words and give the result.

Following are the rules that we derived from the given training data:

['N_NN', 'PR_PRQ', 'RD_PUNC', 'N_NNPC RD_PUNC N_NNPC', 'QT_QTC N_NN RD_PUNC', 'QT_QTF N_NN', 'QT_QTF N_NN RD_PUNC', 'V_VM_VNF', 'V_VM_VF', 'PR_PRP', 'V_VN', 'DM_DMD N_NN PSP', 'PR_PRP QT_QTF', 'PR_PRP N_NN', 'RP_RPD QT_QTF N_NN', 'V_VM_VF RD_PUNC', 'PR_PRP N_NN RP_RPD', 'N_NN PSP RP_RPD', 'RB', 'PR_PRP RP_RPD', 'CC_CCS', 'V_VM_VF RP_RPD', 'CC_CCS_UT', 'N_NN PSP', 'JJ QT_QTF N_NN', 'DM_DMD N_NN', 'JJ QT_QTF N_NN RD_PUNC', 'PR_PRP PSP', 'PR_PRF N_NNC N_NNC RD_PUNC', 'N_NN RD_PUNC', 'V_VM_VINF', 'V_VN RD_PUNC', 'N_NN RP_RPD', 'N_NN N_NST RP_RPD', 'N_NN N_NST', 'QT_QTC N_NN', 'RD_PUNK', 'QT_QTC N_NN RP_RPD', 'V_VM_VNF RD_PUNC', 'QT_QTF QT_QTC N_NN', 'RP_RPD QT_QTF V_VM_VINF', 'RP_INTF QT_QTF N_NN', 'QT_QTC QT_QTF N_NN', 'V_VN N_NST', 'N_NNP', 'QT_QTF QT_QTF N_NN', 'QT_QTF QT_QTF N_NN PSP', 'JJ N_NN', 'CC_CCS RD_PUNC', 'RD_SYM RP_RPD N_NN', 'V_VM_VF RD_SYM', 'PR_PRF', 'RP_UNK', 'RP_RPD JJ N_NN', 'N_NNC N_NNC', 'V_VN PSP', 'V_VN RP_RPD', 'N_NN N_NST PSP RD_PUNC', 'V_VM_VNF RP_RPD', 'V_VM_VINF RD_PUNC', 'PR_PRP N_NN PSP', 'N_NNP RD_PUNC', 'N_NNP RP_RPD', 'JJ JJ N_NN', 'N_NNP PSP', 'RP_RPD N_NN', 'PR_PRF N_NN', 'RP_INTF N_NN RP_RPD', 'CC_CCD RD_PUNC', 'CC_CCS RP_RPD', 'PR_PRP N_NST RP_RPD', 'PR_PRC', 'PR_PRP N_NST', 'JJ N_NN PSP', 'QT_QTF N_NNC N_NNC', 'QT_QTF CC_CCD', 'PR_PRF N_NN N_NST', 'QT_QTF N_NN PSP', 'QT_QTF PR_PRQ RP_RPD', 'RP_INTF N_NN', 'QT_QTF N_NN RP_RPD', 'N_NNC N_NNC PSP', 'QT_QTF PR_PRP', 'V_VN PSP RP_RPD', 'N_NNC RD_SYM N_NNC N_NN', 'QT_QTF JJ N_NN', 'N_NNPC N_NNPC N_NNPC', 'N_NN RP_RPD PSP', 'RD_PUNC N_NN', 'N_NNPC N_NNPC RP_RPD', 'PR_PRP PR_PRP', 'V_VNV_VN', 'V_VM_VF RP_RPD RP_RPD RD_PUNC', 'CC_CCD', 'QT_QTF', 'V_VN PSP RD_PUNC', 'JJ QT_QTF N_NN RP_RPD', 'RP_UNK RD_PUNC', 'JJ N_NN RD_PUNC', 'JJ N_NN RP_RPD', 'JJ JJ N_NN RD_PUNC', 'PR_PRP N_NN RD_PUNC', 'PR_PRQ N_NN V_VM_VF RD_PUNCN_NN', 'QT_QTC QT_QTC N_NN', 'PR_PRF N_NN PSP', 'PR_PRF RP_INTF N_NN', 'PR_PRP PSP RP_RPD', 'QT_QTF JJ N_NN RD_PUNC', 'RP_RPD N_NNP RD_PUNC', 'N_NNC N_NNC RD_PUNC', 'N_NNC N_NNC RP_RPD', 'DM_DMD N_NN RP_RPD', 'QT_QTC N_NN N_NST', 'N_NN RP_RPD RD_PUNC', 'JJ RP_RPD N_NN', 'QT_QTO N_NN', 'QT_QTF N_NNC JJ N_NNC', 'CC_CCS_UT RP_RPD', 'RP_RPD V_VM_VF RD_PUNC', 'V_VM_VF PSP', 'PR_PRP JJ N_NN', 'DM_DMD JJ N_NN', 'QT_QTF N_NN N_NST', 'QT_QTF V_VN N_NST RP_RPD', 'QT_QTO N_NNC N_NNC', 'QT_QTC N_NNC N_NNC', 'N_NNPC N_NNPC RD_PUNC', 'N_NNPC N_NNPC', 'QT_QTC N_NN PSP', 'RP_RPD RB', 'QT_QTC', 'RB RD_PUNC', 'V_VM_VINF PSP', 'PR_PRP N_NNQT_QTF N_NN', 'RP_INTF JJ N_NN']

First 10 chunks sentences for given test data:

- [പോരായ്മകള്ക്ക്] [ഉള്ള] [പരിഹാരങ്ങളും] [കണ്ടെത്തിയിരുന്നു ।]
- [അതിന്റെ ഫലം] [ആയി "] [തോമസ്] [കുക്ക്] [റെയില്വേകൂപ്പണും "] [ഹോട്ടല്കൂപ്പണും] [ആവിഷ്ക്കരിക്കപ്പെട്ടു ।]
- [1866-ല്] [കുക്ക്] [അമേരിക്കയില്] [എത്തി ।]
- [ആ] [വര്ഷം തന്നെ] [അമേരിക്കയിലേയ്ക്ക്] [ഉള്ള] [യാത്രാസംഘത്തെ] [അയയ്ക്കാനും] [സാധിച്ചു ।]
- [യാത്രകള്] [വിപുലവും] [പ്രവര്ത്തനങ്ങള്] [വ്യാപകവും] [ആയതോടെ] [1867-ല്] [തോമസ്] [കുക്ക്] [ആന്ഡ്] [സണ്] [എന്ന്] [ഒരു] [കമ്പനി] [സ്ഥാപിക്കപ്പെട്ടു ।]

- [അപ്പോഴേയ്ക്കും] [അത് ഒരു] [വലിയ പ്രസ്ഥാനം] [ആയി] [വളര്ന്ന്കഴിഞ്ഞിരുന്നു ।]
- [അവര്ക്ക് അധികാരസ്ഥാനങ്ങളില്] [അംഗീകാരവും] [ജനങ്ങള്ക്ക് ഇടയില്] [വിശ്വാസ്യതയും] [കൈവന്നിരുന്നു ।]
- [തോമസ്] [കക്കിന്റെ] [അംഗീകാരവും] [പ്രശസ്തിയും] [വര്ദ്ധിച്ചതോട് ഒപ്പം] [എതിര്പ്പുകളും] [വിമര്ശനങ്ങളും] [ഇല്ലാതെയും] [ഇരുന്നില്ല ।]
- [1872-ല്] [തോമസ്] [കുക്ക്] [തന്റെ] [ആദ്യത്തെ] [ആഗോളയാത്ര] [നടത്തി] [.]
- [എന്നാല്] [ടൂറിസ്റ്റുകളുടെ] [ആഗോളയാത്ര] [ആദ്യം] [ആയി] [സംഘടിക്കപ്പെട്ടത്] [1878-ല്] [സ്റ്റാനഞെന്] [എന്ന] [ജര്മ്മന്] [കമ്പനി] [ആയിരുന്നു)] [ലോകയാത്രയുടെ]

Accuracy:

Total chunks: 580

Correct chunks formed: 424

Accuracy: **73.11 %**

3) CONCLUSION:

The models designed for chunking based on CRF gave higher performance as compared to Rule based model which is written by noting down rules in train data. The templates are very important in designing the CRF as they highly effect the performance of the chunking. CRF is best way as compared to Rule Based Model.

REFERENCES:

FOR CRF++:

http://crfpp.googlecode.com/svn/trunk/doc/index.html?source=navbar