Part of Speech Tagger

Kunal Sachdeva
kunal.sachdeva@students.iiit.ac.in

In this document we explain the steps involved in training a CRF(conditional random fields) based part of speech tagger on treebank data.

Assumptions

- You have a working CRF++ package . For further reading and downloading the latest version refer to the below mentioned link.
  - http://crfpp.googlecode.com/svn/trunk/doc/index.html?source=navbar#source
  - In case you get an error “error while loading shared libraries” , give the exact location of scripts used for training and testing.
- The data being used is from treebank in SSF format, although you can change the scripts according to your needs.
- A new line should be inserted after end of each sentence.

Description of feature set

- We are describing our experiments on Hindi Treebank data in 'wx' format, considering the below mentioned features. You can use any encoding scheme along with suitable features. You need to make changes in the feature computation scripts to account for desirable features.
- This document is an extension of postagger-hindi 2.2 (used in ILMT), where we have not mentioned about the training procedure.
- For defining the features used for training, a separate 'template' file is provided during training of model. For our purpose we have considered a window of 2 words i.e 2 words after and 2 words before any target word for prediction.
- We have considered a suffix length of 5 words and a prefix length of 7 words. For words less than this threshold a NULL will be inserted .
- We require two steps for conversion of data to desired input format for CRF++.
  - **SSFtoTnt:** In this process we extract the relevant information (word and gold POS) from the SSF format.
  - **Extrafeatures:** In this process we compute the features from the word which are mentioned in the template file.

Example:

In template file, we mention the features as below. In this example we are using the feature of context window.

```
U00:%x[-2,0] -Second previous word
U01:%x[-1,0] -Previous word
U02:%x[0,0]  - Current word
U03:%x[1,0]  - Next word
U04:%x[2,0]  -Second next word
```

Example:

In training file we mention the features as below. Each new sentence should be preceeded by a new line character. The last column contains the gold part of speech tag which will be used as a label in training.

aMwima MORE wima ima ma a LL aMwima aMwim aMwi aMw aM a JJ
(word) (5 length suffix) (7 length prefix) (Gold POS)

Results

- We have used ~14,000 sentences(~3,00,000 tokens) for training and ~3000 sentences for testing the module.
- We achieved an acuuracy of 95.7 % on the test set.