IDENTIFYING CODE SWITCHING IN TWEETS

GROUP NO. 8

Submitted By:

Himanshu Verma
Nitesh Sekhar
Kalyan Kumar
Riya Bubna
Sandeep Pan

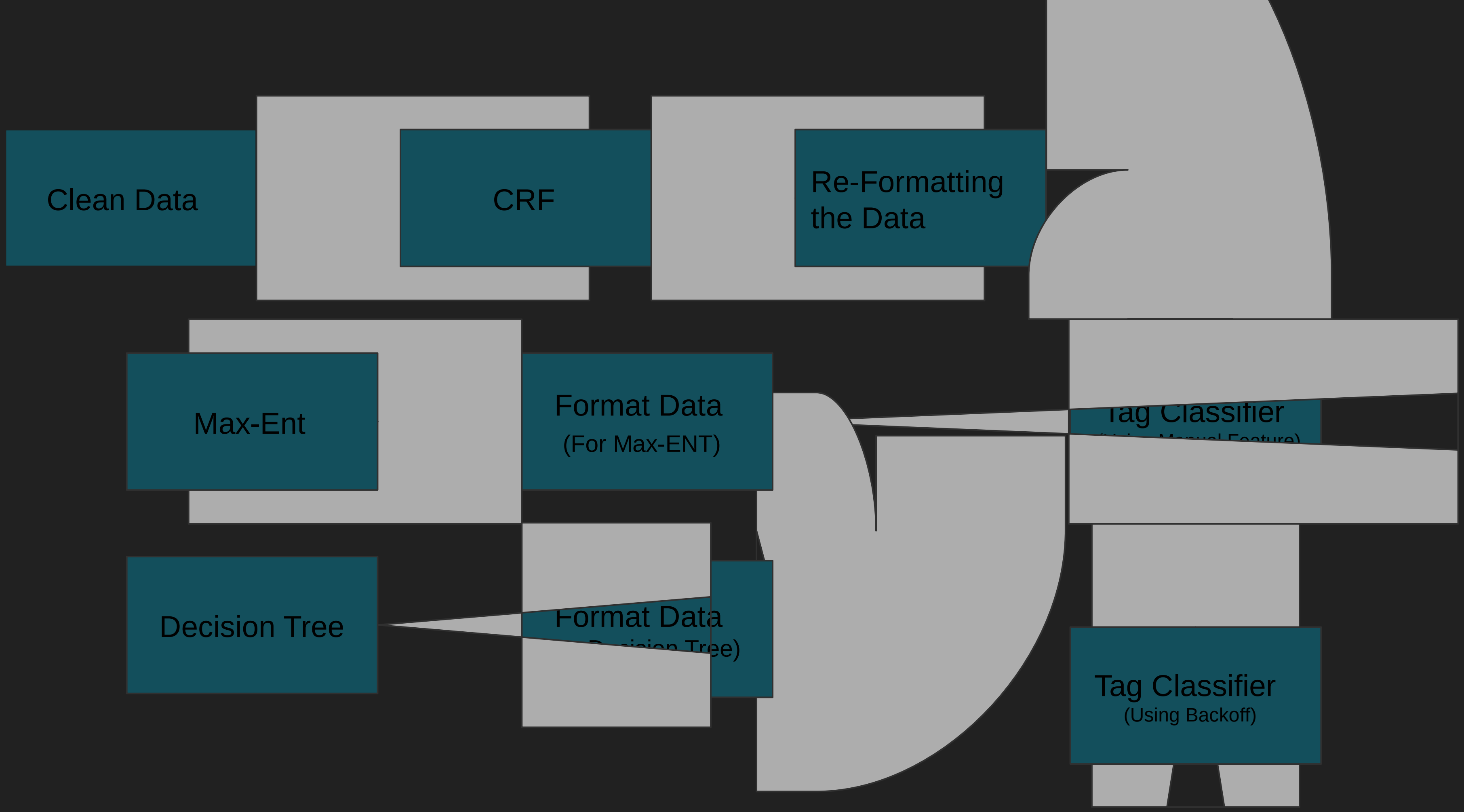
Shrey Garg
Shubham Jain

Mentored By :

Dr. Monojit Choudhury
Pawan Goyal
Koustav Rudra

Objective

Given tweets written in a mixture of English and Hindi, identifying the points where there is a change in language. If a language change is found, then identifying whether it is a code-switched point or not.



Clean Data

CRF

Re-Formatting the Data

Max-Ent

Format Data
(For Max-ENT)

Tag Classifier
(Using Manual Feature)

Decision Tree

Format Data
(For Decision Tree)

Tag Classifier
(Using Backoff)

Cleaning the original Tweets

We cleaned the whole data in 8 steps :

1. Remove all the tweet id’s

Regular Expression : \d{5,}

2. Remove all the usernames

Regular Expression : (\@(\w|_)+(\:|))

3. Remove all the hashtags

Regular Expression : \#(\w|_)+

4. Remove all the urls

Regular Expression : (http|https|ftp|mailto|tel):\S+[/a-zA-Z0-9]

Contd...

5. Replacing multiple occurrences of character or punctuation in word by twice the same character or punctuation

   Regular Expression : (.)\1+ ---> \1\1

6. Remove digits as we don't need them

   Regular Expression : \d+[^\s]*

7. Remove all these unnecessary punctuations

   Regular Expression : \~|\@|\#|\$|\%|\^|\&|*|\[|\]|\{|\}|\\|V|\;|\:|\>|\<|\-|_|\=|\+|\-|*|\'|\"|\|

8. Convert these punctuation in dots

   Regular Expression : \!|\(|\)|\.|\,|\? ---> \.

CONDITIONAL RANDOM FIELD (CRF)

CRF was used to make a language model for checking that the word is English or Hindi .The template file was made which had different features which related the current word with the tag of the previous two words and the next two words to mark the tag of the current word. CRF model was created by using that template file and train file which contained around 24000 tweets with around 5 lakh words.

Features

To train the data, we needed some features to identify when there is a case of code switching. These features include:

- A binary value to denote the existence of trigrams, bigrams and unigrams before and after any point of language change.
- The presence of punctuation after any word.
- Any word has language english or hindi.
- If language change is encountered after any word.
- If a language change is encountered at position ‘i’, then the present word is present at a position at least ‘i+3’

RESULT OF DETERMINISTIC MODEL

We used a test-file which contained code switched Data to test our features.

The accuracy came out to be around 30.7 %

However, we realized that there were a lot of sentences in the file which were neither code-switching or mixing :

eg : mein school jaa rha hun .

Here, we obtained change tags after mein and school .

However such points can’t be classified as language-change points .

Hence after removing these points, out accuracy came to be about 70.82% (even 83% in one test file)

Max Ent Model

Features :

For each tag, distance from the previous and the next tag .

Sliding window centered around the current word and encoding this information in the form of an integer.

Result :

If we consider the entire file as code switching,accuracy is around 51%.

However considering code-mixing in the data, accuracy comes to around 80% .

Decision Tree

A decision tree is a graph that uses a branching method to illustrate every possible outcome of a decision. Decision Trees are excellent tools for helping you to choose between several courses of action. They provide a highly effective structure within which you can lay out options and investigate the possible outcomes of choosing those options .

In our project, we used a python implementation of a decision tree model to determine whether or not a word is a code switching point or not. We used the following attributes while training the decision tree and for testing the output.

Decision Tree Results

Attributes for Decision Tree :

Initial_trigram? , Initial_bigram? , Initial_unigram? , End_trigram? , End_bigram? , End_unigram? , LanguageEng? , distance_from_prev_change , Punctuation?

Result

Considering all the language change points as code switch, the accuracy had come to around 24% . But as we had seen earlier , removing change points with only two words between them is sure to increase the accuracy by at least 20% more.

Input

```
Indian batsman itna slow khel rahe he ki lagta he is se fast to out of form sehwag and sachin khel
lete... #IndvsSA
@NitishKumar Sir Aaj aap jit kaye aur hum bihari har gaye, was just reading comment from people who are
not bihari and got d same feeling of 90s
```

Output for Deterministic Model

```
indian <xx> batsman <xx> itna <xx> slow <cs> khel rahe <xx> he <xx> ki lagta <cs> he is <cs> se <xx> fast
to out of form <cs> sehwag <xx> and <xx> sachin khel lete
sir <xx> aaj aap jit kaye aur <xx> hum <xx> bihari har gaye <cs> was just reading comment from people who
are not <cs> bihari <xx> and got d same feeling of
```

Output for Max-Ent Model

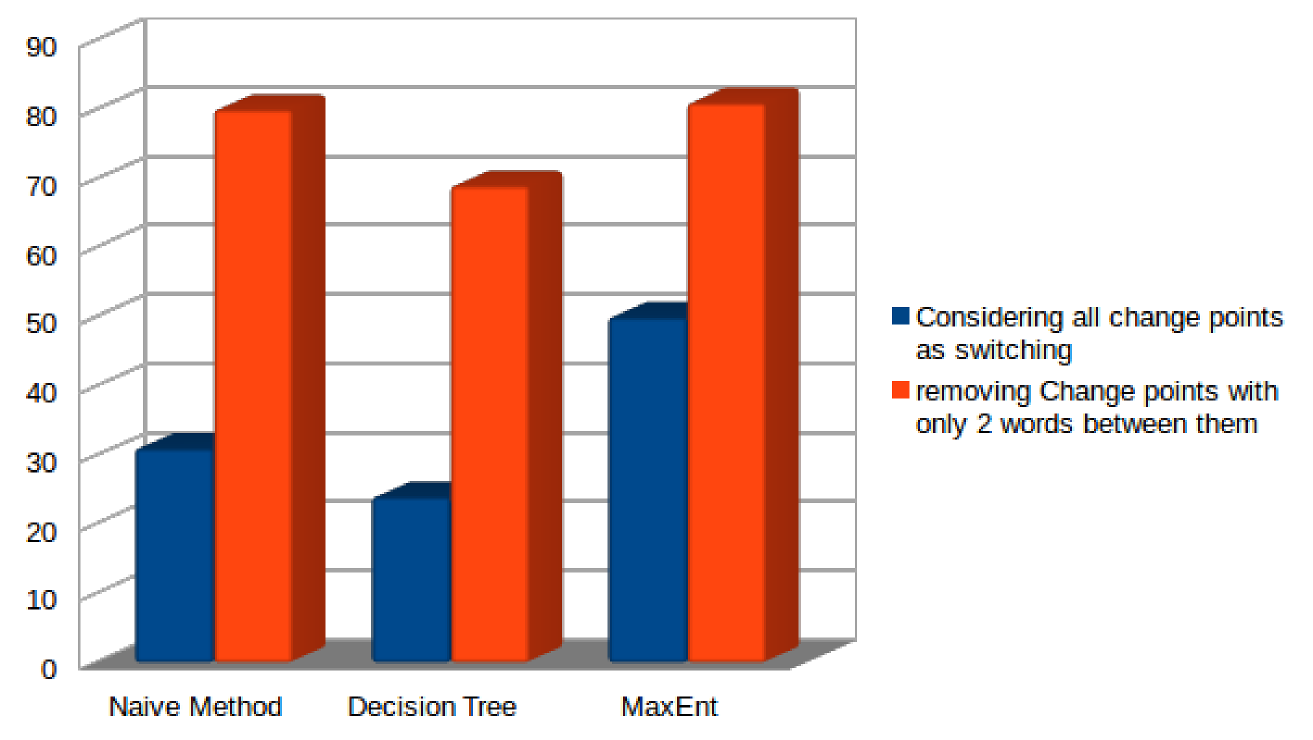
```
indian <Code-Mixed> batsman <Code-Mixed> itna <Code-Mixed> slow <Code-Mixed> khel rahe <Code-Mixed> he
<Code-Mixed> ki lagta <Code-Mixed> he is <Code-Mixed> se <Code-Mixed> fast to out of form <Code-Mixed>
sehwag <Code-Mixed> and <Code-Mixed> sachin khel lete .
sir <Code-Mixed> aaj aap jit kaye aur <Code-Mixed> hum <Code-Mixed> bihari har gaye . <Code-Switched> was
just reading comment from people who are not <Code-Switched> bihari <Code-Switched> and got d same
feeling of
```

Output for Decision tree

```
word = no
indian = no
batsman = ?
itna = no
slow = ?
khel = no
rahe = no
he = no
ki = no
lagta = no
he = no
is = ?
se = no
fast = no
to = no
out = no
of = no
form = no
sehwag = no
and = no
sachin = no
khel = no
lete = ?
```

```
sir = no
aaj = no
aap = no
jit = no
kaye = ?
aur = no
hum = ?
bihari = no
har = no
gaye = yes
was = no
just = no
reading = ?
comment = no
from = no
people = no
who = no
are = no
not = ?
bihari = no
and = no
got = no
d = no
same = no
feeling = no
of = no
```

Conclusion


90
80
70
60
50
40
30
20
10
0
Naive Method
Decision Tree
MaxEnt
🟦 Considering all change points as switching
🟥 removing Change points with only 2 words between them

THANKS!