Classification of Hindi Literature according to Author Writing Style

Dhruv Anand 11251
Srijan Shetty 11727
Motivation

➔ Document Fraud Detection
➔ Classifying works from unknown authors
➔ From a Literary perspective
  ◆ Repeating trends of authors
  ◆ Adopting styles of popular authors
Previous Work

➔ Extensive work done on Author Attribution for English (using domain-specific datasets like blogs, emails, forum posts, short stories and novels)

➔ No work has been done on Hindi datasets

➔ Various lexical and syntactic features have been tried by researchers in this field
Challenges

➔ Non-uniform data for Hindi
➔ Variance of writing style markers in Hindi Literature
➔ Multiple derivative words that must be aggregated without any pre-programmed tool for lemmatization. (The language is morphologically rich.)
Problem Statement

➔ Apply known methods of Author Attribution to a Hindi dataset
➔ Analyse difference in effectiveness of various methods between English and Hindi
➔ Exploring new types of lexical and syntactic features to give better results for Hindi Literature
Methodology
Proposed Features

 ➔ Word n-grams
   ◆ Stemmed/non-stemmed unigrams
   ◆ Collocations (bigrams)
 ➔ Character n-grams
 ➔ Sentence length distribution
 ➔ Word length distribution
 ➔ Feature word frequency distribution
<table>
<thead>
<tr>
<th>Features</th>
<th>Required tools and resources</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Lexical</strong></td>
<td></td>
</tr>
<tr>
<td>Token-based (word length, sentence length, etc.)</td>
<td>Tokenizer, [Sentence splitter]</td>
</tr>
<tr>
<td>Vocabulary richness</td>
<td>Tokenizer</td>
</tr>
<tr>
<td>Word frequencies</td>
<td>Tokenizer, [Stemmer, Lemmatizer]</td>
</tr>
<tr>
<td>Word n-grams</td>
<td>Tokenizer</td>
</tr>
<tr>
<td>Errors</td>
<td>Tokenizer, Orthographic spell checker</td>
</tr>
<tr>
<td><strong>Character</strong></td>
<td></td>
</tr>
<tr>
<td>Character types (letters, digits, etc.)</td>
<td>Character dictionary</td>
</tr>
<tr>
<td>Character n-grams (fixed-length)</td>
<td>-</td>
</tr>
<tr>
<td>Character n-grams (variable-length)</td>
<td>Feature selector</td>
</tr>
<tr>
<td>Compression methods</td>
<td>Text compression tool</td>
</tr>
<tr>
<td><strong>Syntactic</strong></td>
<td></td>
</tr>
<tr>
<td>Part-of-Speech</td>
<td>Tokenizer, Sentence splitter, POS tagger</td>
</tr>
<tr>
<td>Chunks</td>
<td>Tokenizer, Sentence splitter, [POS tagger], Text chunker</td>
</tr>
<tr>
<td>Sentence and phrase structure</td>
<td>Tokenizer, Sentence splitter, POS tagger, Text chunker, Partial parser</td>
</tr>
<tr>
<td>Rewrite rules frequencies</td>
<td>Tokenizer, Sentence splitter, POS tagger, Text chunker, Full parser</td>
</tr>
<tr>
<td>Errors</td>
<td>Tokenizer, Sentence splitter, Syntactic spell checker</td>
</tr>
</tbody>
</table>

*image from [Sta09]*
Classification

→ Supervised
  ◆ SVMs
  ◆ Bayesian Multinominal Regression (BMR)

→ Unsupervised
  ◆ K-means clustering
Framework

Stage 1
Feature Extraction

Stage 2
Classification

Stage 3
Evaluation

Results

Text Snippets

Feature Specification

Feature Vectors

Label Assignment
A bit of theory
In the beginning God created
the heaven and the earth.
And the earth was without form,
and void; and darkness was
upon the face of the deep.
And the Spirit of God moved
upon the face of the waters.
And God said, Let there be
light: and there was light.

<table>
<thead>
<tr>
<th></th>
<th>Representation</th>
</tr>
</thead>
<tbody>
<tr>
<td>beginning</td>
<td>1</td>
</tr>
<tr>
<td>earth</td>
<td>2</td>
</tr>
<tr>
<td>God</td>
<td>3</td>
</tr>
</tbody>
</table>
K Means

(http://www.mathworks.com/matlabcentral/fileexchange/screenshots/2240/original.jpg)
\[
\Pr(Y_i = 1) = \frac{e^{\beta_1 \cdot X_i}}{1 + \sum_{k=1}^{K-1} e^{\beta_k' \cdot X_i}}
\]

\[
\Pr(Y_i = K - 1) = \frac{e^{\beta_{K-1}' \cdot X_i}}{1 + \sum_{k=1}^{K-1} e^{\beta_k' \cdot X_i}}
\]

\[
\Pr(Y_i = K) = \frac{1}{1 + \sum_{k=1}^{K-1} e^{\beta_k' \cdot X_i}}
\]
Where do we stand
Dataset Compilation

➔ No standard dataset for classical/contemporary hindi authors (novels and stories)

➔ Scraped HindiSamay.com manually to build a database of Classical Hindi literature.
  ◆ 5 authors
  ◆ 2-4 lakh words per author

➔ Each author’s work has been divided into multiple snippets of 500 words.
Belief: Authors repeat the same set of words

Stemming: BOW using all tokens and BOW using 4500 most frequent words (>20 frequency in the entire corpus)

Classification: K-means on 3 classes (RNT, Premchand, V.N.Rai) and on 5 classes.

Results for 3 classes:
- Average Precision: 50% (v/s baseline of 33%)
- Average Recall: 48% (v/s baseline of 33%)
# Results with 5 authors

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>Snippets</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNT</td>
<td>111</td>
<td>14</td>
<td>20</td>
<td>0</td>
<td>6</td>
<td>151</td>
<td>22.65%</td>
<td>73.5%</td>
</tr>
<tr>
<td>Prem</td>
<td>108</td>
<td>21</td>
<td>58</td>
<td>0</td>
<td>211</td>
<td>398</td>
<td>71.77%</td>
<td>53.01%</td>
</tr>
<tr>
<td>Dharamvir</td>
<td>11</td>
<td>24</td>
<td>14</td>
<td>150</td>
<td>2</td>
<td>201</td>
<td>100%</td>
<td>74.6%</td>
</tr>
<tr>
<td>Sarat</td>
<td>142</td>
<td>332</td>
<td>3</td>
<td>0</td>
<td>65</td>
<td>542</td>
<td>82.19%</td>
<td>61.25%</td>
</tr>
<tr>
<td>VN</td>
<td>118</td>
<td>13</td>
<td>277</td>
<td>0</td>
<td>10</td>
<td>418</td>
<td>74.46%</td>
<td>66.26%</td>
</tr>
</tbody>
</table>
Corpus has mostly stories for Rabindranath Tagore, both recall and precision for him are low indicating that across multiple works frequent words used by author change.

Corpus contained only novels for Premchand and so both recall and precision for him were high > 70%.

The corpus contained essays by V.N.Rai, indicating high amount of content words.
Future Work
In the coming weeks

➔ Use collocations (bigrams) as a feature.

➔ Analyzing sentence structure:
  ◆ Sentence lengths
  ◆ Number of subjects, verbs, objects in a sentence (instead of POS tagging we will lookup common words from HindiWordNet)

➔ Reducing dimensionality using PCA.

➔ Training on multiple features together (using multivariate discriminant analysis)

➔ Improving results by tuning snippet length and parameters used in classification.
In the future

➔ Exploring the possibility of using a morphological tagger to get more accurate style measures for authors.
➔ Extending the method to Hindi tweets, forum comments and messages to compare accuracy.
References


Tools Used

➔ ZSH
➔ Python Modules
  ◆ indicngram
  ◆ nltk, scipy, scikit-learn
➔ Snippets of code have been taken from
  ◆ http://www.csc.villanova.edu/~matuszek/spring2012/snippets.html

*www.python.org
THANK YOU!

Questions?