Classification of Hindi Literature according to Author Writing Style

Srijan R. Shetty (11727) [*1] and Dhruv Anand (11251) [†1]

[1]Department of Computer Science and Engineering, Indian Institute of Technology Kanpur

February 28, 2014

1 Problem Statement

The objective of author attribution is to map a given piece of text to its author. In this project, we explore features which when used in unison changed significantly with the change of the author. This kind of analysis, while prevalent for English corpora, has never been tried for an Indian language such as Hindi. The deterrents for Indian Languages, we believe are their morphologically rich structure as well as the lack of a standard corpora.

2 Motivation

While we are employing our methods for author attribution of well-known authors, we believe that given enough data of any user of the language, we can profile him. This *signature* can then be used for online *scam* and *fraud detection*. On the flip side, the same methodology can be used to *uncover anonymous writers* or journalists who work under an alias to protect themselves. This project can therefore serve as a reminder to these authors that internet anonymity has limitations and they should obfuscate their work before publishing. That being said, we hope that the results of our project can be used for greater good.

3 Data Set

- **HindiSamay**
  The aforementioned website contains poems, stories and epics from myriad known and relatively unknown authors (classical and contemporary). We intend to scrape this website to build a corpora with labelled data of authors and their texts.
- **IIIT Indian Languages Corpus**
  Language Technologies Research Center, IIIT-H, allows its classical Hindi Literature corpus to be used for academic purposes.
- **Emille Corpus**
  As stated in their website

  > The EMILLE Corpus has been constructed as part of a collaborative venture between the EMILLE project (Enabling Minority Language Engineering), Lancaster University, UK, and the Central Institute of Indian Languages (CIIL), Mysore, India. EMILLE is distributed by the European Language Resources Association. The corpus consists of three components: monolingual, parallel and annotated corpora. The EMILLE/CIIL Corpus (ELRA-W0037) is distributed free of charge for use in non-profit-making research only.

[*]srijans@iitk.ac.in
[†]adhruv@iitk.ac.in

4 Methodology

A non exhaustive and probable set of features that we intend to use for the classification of authors is:

- **Word Frequency**
  A *bag-of-words* aprroach using a term frequency-inverse document frequency function etc.
- **n-grams**
  N gram analysis will be performed first using only the tokens and then later on if possible using morphological units.
- **Word Length**
  Histogram vector for each story.
- **Sentence Length**
  Histogram vector for each story.
- **Vocabulary Richness/Diversity**

The vector formed for each story by this feature set could be processed using the following machine learning methods:

- **k-means clustering** (*Unsupervised*)
  The labels of the data will be ignored in this analysis and will be only used in the evaluation phase to test accuracy.
- **Support Vector Machine** (*Supervised*)
  This method can be used by using some of our labelled data as test data (by the validation set method). The comparison between both these types of methods could bring more insight into the problem of author attribution.

References

[KSA09] Moshe Koppel, Jonathan Schler, and Shlomo Argamon. Computational methods in authorship attribution. *J. Am. Soc. Inf. Sci. Technol.*, 60(1):9–26, January 2009.

[KSA11] Moshe Koppel, Jonathan Schler, and Shlomo Argamon. Authorship attribution in the wild. *Lang. Resour. Eval.*, 45(1):83–94, March 2011.

[Sta09] Efstathios Stamatatos. A survey of modern authorship attribution methods. *J. Am. Soc. Inf. Sci. Technol.*, 60(3):538–556, March 2009.