Author Identification in Malayalam using n-grams

Bindu Baby Thomas
bbtuce@gmail.com
Research Scholar, Dept of Computer Science,
CUSAT, Cochin

Sindhu L.
Sindhul_cep@yahoo.co.in
Research Scholar, Dept of Computer Science,
CUSAT, Cochin

Dr Sumam M Idicula
sumam@gmail.com
Reader, Dept of Computer Science, CUSAT,
Cochin

Abstract -**Author identification is the problem of identifying the author of an anonymous text or text whose authorship is in doubt from a given set of authors. The works by different authors are strongly distinguished by quantifiable features of the text. This paper deals with the attempts made on identifying the most likely author of a text in Malayalam from a list of authors. Malayalam is a Dravidian language with agglutinative nature and not much successful tools have been developed to extract syntactic & semantic features of texts in this language. We have done a detailed study on the various stylometric features that can be used to form an authors profile and have found that the frequencies of word collocations can be used to clearly distinguish an author in a highly inflectious language such as Malayalam. In our work we try to extract the word level and character level features present in the text for characterizing the style of an author. Our first step was towards creating a profile for each of the candidate authors whose texts were available with us, first from word n-gram frequencies and then by using variable length character n-gram frequencies. Profiles of the set of authors under consideration thus formed, was then compared with the features extracted from anonymous text, to suggest the most likely author.**

Keywords – **stylometrics, feature extraction, author profile, lexical features, character features, collocations, classification, n-grams, distance measure.**

1. INTRODUCTION

Author identification is the task of identifying the most likely author of an anonymous text or text whose authorship is in doubt from among the list of known candidates. Scientific investigation regarding the authorship of texts has been done in many languages. These works were done under the notion that works by different authors are strongly distinguished by features of the text that can be quantified [1]. This line of research is known as stylometrics.

These techniques used for authorship attribution rely mainly on the fact that the authors exhibit some semantic, syntactic, lexicographic and morphological features [4] which can be used for profile creation. But detecting and extracting these features from the texts and forming a profile for each of the authors becomes a tedious language dependent task. In an effort to construct an automated authorship attributor for documents in Malayalam (a Dravidian Language) a study of the various stylometric features that can be used for forming an authors profile was done. As in the current scenario, very few feature extraction tools are available in this highly inflectious language the study was limited to Lexical & Character level features. A comparison between the character level features and word level features were done based on the sample corpus.

2. STYLOMETRY

Since 1964, the research in authorship attribution is mainly supported by attempts to define features for quantifying writing style. This line of research is known as ‘stylometry’. These quantifiable features include sentence length, word length, word frequencies, character frequencies, vocabulary richness functions etc.

2.1 Lexical features

The word length, sentence length, vocabulary richness, word frequencies, word n-grams, spelling errors etc., are some of the features which can be used for the profile creation of an author. These are all simple, yet powerful in representing the characteristics of an author. If word frequencies are used, then the most common words (articles, prepositions, pronouns etc) are found to be the best for representing an author’s style. Another measure is a set of spelling errors (eg : letter omissions and insertions), formatting errors, usage of some abbreviations etc, which are very specific to a particular author and can be extracted using spell checkers. Extraction of these class of features need tools like tokenizer, sentence splitter, stemmer, lemmatizer, spell checkers, case converters (not needed in Malayalam), detectors of homographic forms etc most of which are language specific.

2.2 Character Features.

The features can be character-level also, like alphabetic characters count, digit characters count, uppercase and lowercase characters count (in case of case sensitive languages-not applicable here), letter frequencies, punctuation marks count etc. These are easy to be measured. Another feature is character n- gram, which involves the extraction of the frequencies of n-grams at the character level. Another case of using character information is the compression based approach, where the compression model acquired from one text used to compress another text. If two texts are written by the same author, the size of the resulting file will be relatively low because of the repetitions. Some amount of success has been reported in this case [10].

2.3 Syntactic Features

Authors tend to use similar syntactic patterns repeatedly. Extracting information about POS, sentence and phrase structures, common grammatical errors etc, can give more information about the author's style. But they need very accurate NLP tools like sentence splitter, POS tagger, Parser, syntax checker etc. to perform syntactic analysis of texts. Some features that can be used are noun phrase counts, verb phrase counts, length of noun phrases, length of verb phrases, presence of errors, sentence fragments, mismatched tense etc.

2.4 Semantic features

Only a few attempts have been made for the extraction of semantic features. Extraction of binary semantic features (number and person of nouns, tense and aspect of verbs, etc) and semantic modification relations (the syntactic and semantic relations between a node of the graph and its daughters) can be considered as semantic features. The information about an author can also be represented by the usage of synonyms, hyponyms of words.

2.5 Application Specific Features

Application-specific measures can be defined in order to better represent the style of an author. Some such measures include the use of greetings and farewells in the messages (as in The New Testament letters written by St Luke), types of signatures, use of indentation, paragraph length, etc. If the texts in question are in HTML form, measures related to HTML tag distribution, font color counts, and font size counts can also be defined. In order to better capture the properties of an author's style within a particular text domain, content-specific keywords can be used.

3. FEATURE SELECTION

After looking into the different groups of stylometric features, now which of these features are to be selected for representing an authors writing style was the question. The answer to this mainly depended on two things, the availability of extraction methods and the frequency of the feature. The more frequent a feature the more stylistic variation it captures.

The simplest extraction methods are for the lexical and character features. Houvards and Stamatos [2] proposed an approach for extracting character n-grams of variable length using frequency information only. The frequency based feature set was more accurate for feature sets comprising up to 4,000 features [3]. Thus the frequencies of word n-gram, or character n-gram can be chosen as the best and many works have been successfully done using these. [9]

Language independent extraction methods can be used for character n-grams by using the byte level information. Before collecting the statistics of n-grams only simple preprocessing has to be done and the extraction of each gram involves no complicated process as the character length of each gram is the same. As we collect character n-grams we can count the number of instances of each one in the corpus. Also it is not affected by noises easily, like the presence of spelling and grammatical errors. Even if there are such errors, there will still be a lot of common n-grams. Finally for languages where the tokenization procedure is quite hard, character n-gram approach is more suitable.

Word n-gram approach needs a tokenizer, to separate the words. Also it may require additional tools (language dependent) like stemmers, lemmatizers, and detectors of common homographic forms for the extraction of words. But it has been proved that the frequency of word collocations can be best used to distinguish an authors writing style. Also in style based text categorization the presence of uncommon n-grams caused by spelling errors, could be considered as personal traits of the author. Thus if a good tokenizer is available for the language, word n-gram frequencies can be considered to create an author's profile.

Peng, Shuurmans, Keselj & Wang [4], Keselj and Stamatatos [5] reported very good results using character n-gram information. Moreover, one of the best performing algorithms in an authorship attribution competition organized in 2004 was also based on a character *n*-gram representation [6] ,[7].

4. FEATURE EXTRACTION FROM DOCUMENTS IN MALAYALAM

Malayalam is one of the Dravidian languages, a morphologically rich and highly agglutinative language. The verb takes tense, aspect, mood and does not take person, number, and gender marker. The Language Processing works in this language is still in its early stages. Morphological Analyzer and POS taggers are still under construction and not yet available for use. Hence the extraction of stylometric features like syntactic features & semantic features from texts in Malayalam cannot be considered at all. Now that we are left out with character features and Lexical features.

4.1 Character Feature Extraction

As already stated, language independent extraction methods can be used for forming a feature set based on frequency of character n-grams. Frequency of occurrences of character groups can be chosen for representing an authors profile, if a purely language independent approach is needed. Extraction of information is easier as only little preprocessing is needed to be done. But the volume of information generated is very huge hence the classification problem will be more complex. Again the size of the feature set has to be shortened by introducing many filtering measures.

If we convert strings with only letters in the English alphabet into 3-grams, we get a 26^3-dimensional space (the first dimension measures the number of occurrences of "aaa", the second "aab", and so forth for all possible combinations of three letters). Using this representation, we lose information about the string. However, we know empirically that if two strings of real text have a similar vector representation then they are likely to be similar.

When we went into the character level frequencies of documents in Malayalam, it was seen that the most frequent character groups or bi-grams generated, comprised mainly of the pure consonants (chillaksharam). In the language there are 56 letters including 15 long and short vowels, 36 consonants and 5 pure consonants. If we look at all possible combination of these characters the feature set size will be too large compared to that of English. Also being an agglutinative language a substantial number of words begin and end with vowels and these words on joining with affixes changes their form.

4.2 Word Frequency Extraction

Frequency of occurrences of individual words or group of words can be considered as the feature set of an author. The most frequent words and the most rarely used words both can together represent an author uniquely. This has been under consideration since the time of Zipf.

4.2.1 Zipf's Law

Zipf argues in his book "Human Behaviour and the Principle of Least Effort" that people will act so as to minimize their probable average rate of work. If we have the frequency of occurrences of a list of words in a large corpus as $\mathbf{f}_i$ and ranks of each in the list as $\mathbf{r}_i$, then Zipf's law states that

$$f \quad \alpha \quad 1/r$$

This is a rough description of the frequency distribution of words in human languages. According to him the speaker/ writer try to minimize his effort by having a small vocabulary of common words. Many further studies based on his findings have been done and proved the validity of this statement to some extend.

4.2.2 Collocations

A collocation is any term or phrase or accepted usage where somehow the whole is perceived to have an existence beyond the sum of the parts. It is an expression consisting of two or more words that corresponds to some conventional way of saying things. We are interested in collocations because they show the frequent forms in which a word is used, that is contextual information is also available for the word. Above this, the frequent use of some collocation gives us a clear guess about the author's profile.

Collocations may be several words long, but we have restricted our study to 2-4 word collocations as they tend to have frequent usage. The simplest method for finding collocations in a text corpus is counting. If two words occur together a lot then that is evidence that they have a special function.

4.2.3 Compounding of Words

Compounding is one of the most important features of Malayalam Language word formation process. Two types of compounding are done mainly, noun-noun compounding and noun-verb compounding. The usual word formation is done when nouns are freely combined with other nouns.

Eg: neela + aakaasham = neelaakaasham
blue sky blue sky

Kadalas + thoni = kadalasthoni
Paper boat paper boat

When nouns are combined with verbs more grammatical features have to be considered.

Eg Uunjaal + Aadi = Uunjaalaadi
Swing(N) swing(V) to swing

Poo + Iruthu = Pooviruthu
Flower plucked plucked flower

In Malayalam it is possible to form lengthy compound words but governed by sandhi rules. The formation of such lengthy words usually portray the authors style and thus the frequency of occurrence of such words in a document can easily discriminate between the documents of different authors.

5. FEATURE SET FORMATION FROM TEXTS IN MALAYALAM

After looking into the character level structure of the texts in this language it has become evident that the symbols which contribute to the vowel sounds in the language, which may come in different combinations contribute more to the character frequencies. This has to be solved by adding filters which should be selected with much care. Also single character frequency cannot be considered as an efficient one. N-gram character frequency can be used, again with much care given while selecting the value of ‘n’, considering the size of the feature set to be formed.

Lexical features like word frequencies or frequencies of collocations can be considered, but the language permits to construct very large words by compounding many words. Extraction of root forms from these requires good stemmers or lemmatizers. Not many tools are available in Malayalam to help us in extracting lexical features except word frequencies.

Keeping these in mind, a combination of both single word and group word frequencies were used to create an author profile. After doing some level of preprocessing on the documents taken from the web tokenization was done and 1- 4 gram word frequencies of documents in Malayalam have been created. It was observed that single word frequencies and double word frequencies together efficiently represented an author’s style while 3- 4 word frequencies contributed less. This would have been more effective if a lemmatizer was available & the frequencies of most commonly used suffixes were known.

6. CNG APPROACH

The Common N-Grams (CNG) approach to author identification [5] is based on constructing author profiles consisting of the most frequent character n-grams found in the text. This method has achieved good results in many experiments [6]. In this approach each text sample is considered as a bag of n-grams and author’s profile is created. The profile comprises of pairs of values ie; most frequent n-grams of the text and their normalized frequencies of occurance. A test document is assigned to an author using a distance measure performed between the test document profile and the profiles of the various candidate authors. The distance measure used in this approach to find d_0 is

$$d_0\,(P(x), P(T_a)) = \sum_{g \in P(x) \cup P(T_a)} \left[\frac{2(f_x(g) - f_{T_a}(g))}{f_x(g) + f_{T_a}(g)}\right]^2$$

P(x) & P(T_a) are the profiles of test document and the author a , f_x(g) and f_{Ta}(g) are the frequencies of the n-gram g in the test text and authors profile. If g is not in P then f(g) = 0. The author with whose profile minimum distance measure obtained is considered to be the most likely author.

7. EXPERIMENT RESULTS

The first step performed was preprocessing of the texts obtained for each of the authors. The first preprocessing step was making sure that the texts collected were all encoded in UTF-8 as the tool was written to accept UTF-8 encoding. ocessing include skipping of extra spaces, punctuation marks, numeric values and other non Malayalam characters. After that the given text was tokenized and stored in a separate file.

The next step is Profile generation. Corpus of four authors were obtained and the preprocessing steps were performed. Each authors profile was generated then first with word uni-grams and then with variable length character n-grams(3<=n<=8). The size of authors profile was 1000 to 2000 for three of them. The profile created from word uni grams were considerably smaller in size and produced less than 10% accuracy. So we proceeded with character n-gram profiles. Also this set was made up of mainly the common inflictions used in the text. The profiles thus created gave a better result of 76% accuracy for one of the authors, 58 – 60 % accuracy for two authors and 36% accuracy for the author with very short profile [8]. Better results were obtained for authors who had larger training texts, all the test documents had a tendency towards the larger profile.

8. CONCLUSION

Studies of the different stylometric features that can be used for representing an authors writing style have been done. Most of the features needed strongly language dependent and higher level tools for their extraction. It can be seen that to derive syntax level and semantic level features which can represent stylistic information greatly, more language level works has to be done. But for developing an author attribution tool its language independent nature is an important criterion. Keeping this in mind n-gram character frequency or word frequency can be chosen. We selected the word frequencies as character frequency feature set forming involved decision of proper filters. Also after obtaining the frequency details of single to 4 word groups we decided to choose the value of ‘n’ in n-gram word frequency approach, as 1 and 2 as we found them to be suitable for forming author’s feature set. 3-gram and 4-gram word frequencies contributed only very little. Any classification algorithm can be chosen to work upon such feature sets formed for a set of authors and then attribute a test document to one of the authors. It was observed that for Malayalam documents the inflections added by an author can be used as his style marker and the author who had a larger profile of character n-grams when used can be used for author identification.

REFERENCES

[1] J Burrows. Computation into Criticism: A study of Jane Austen’s Novels and an Experiment in Method. Clarendon Press, Oxford, 1987.

[2] Houvardas, J., & Stamatatos E. (2006). N-gram feature selection for authorship identification. In Proceedings of the 12th International Conference on Artificial Intelligence: Methodology,Systems, Applications, (pp. 77-86), Springer.

[3] Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. Journal of Machine Learning Research, 3, 1289-1305.

[4] Peng, F., Shuurmans, D., Keselj, V., & Wang, S. (2003). Language independent authorship attribution using character level language models. In Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (pp. 267-274).

[5] Keselj, V., Peng, F., Cercone, N., & Thomas, C. (2003). N-gram-based author profiles for authorship attribution. In Proceedings of the Pacific Association for Computational Linguistics (pp. 255-264).

[6] Juola, P. (2004). Ad-hoc authorship attribution competition. In Proceedings of the Joint Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing (pp. 175-176).

[7] Juola, P. (2006). Authorship attribution for electronic documents. In M. Olivier and S. Shenoi (eds.) Advances in Digital Forensics II (pp. 119-130) Springer.

[8] Sanderson, C., & Guenter, S. (2006). Short text authorship attribution via sequence kernels, Markov chains and author unmasking: An investigation. In Proceedings of the International Conference on Empirical Methods in Natural Language Engineering (pp. 482-491).

[9]Patrick Juola (2009). JGAAP : A system for comparative evaluation of Authorship Attribution. JDHCS 2009 Vol 1, No1.

[10]D Pavelec, L.S.Oliviera, E. Justino, L.V.Batista (2009). Author Identification Using Compression Models. ICDAR’09 Proceedings.