A Novel Approach on Tamil Text Classification Using C-Feature

ArunaDevi, K.¹ Saveeth, R²

¹,²Department of Computer Science Engineering
¹²Coimbatore Institute of Technology, India.

Abstract—Text Classification is one of the central issues in the information systems dealing with text data because of the increasing amount of information stored in a digital form. Text Classification Techniques have been applied on Tamil language to extract meaningful information and knowledge from unstructured Tamil text. Tamil language is a morphologically rich Dravidian language, classifying a Tamil document is different than classifying an English text. In order to enhance the effectiveness of information extraction, we have compared feature extraction techniques and text classifiers and suggested an efficient C-feature (Compound feature), using C-feature we can create an efficient vocabulary set for Tamil text classification.

Key words: Tamil Text Classification, Feature Classification, Vocabulary set or bag-of-words, Text Mining, Natural Language Processing

I. INTRODUCTION

Text mining or knowledge discovery is a process of finding useful information from text documents. However, unlike data mining, which mainly focuses on the structured information, text mining focuses on the text data that is in unstructured form. Text mining deals with the analysis of text data by using the support of machine. Research works in text mining solves the problem of text representation, text classification, text clustering information extraction and modeling of hidden information. We have used data mining methods and statistics to handle their specific tasks in the areas like information retrieval, natural language processing and information extraction. In Information Retrieval, we find the documents which is containing the answers and questions that is systems that retrieve the documents based on the keywords. In Information Extraction, the extraction of specific information from text documents ie it embraces all activities regarding document processing namely Namely Entity Recognition. Natural Language Processing (NLP) provides a significant way for understanding natural languages by the use computers.

Tamil is one of the morphologically richest and longest surviving classical languages in the world. Tamil is always a head-final language. The word order will be Subject-Object-Verb(SOV). Not all the Tamil sentence should contain subjects, objects and verb, we can have a Tamil sentence without anyone of these three in a meaningful way without violating the grammar.

Tamil Text Classification is the automatic process of assigning Tamil documents into the various classes or group of documents based on their contents. The enormous amount of information stored in a digital form are in unstructured texts cannot simply be used by computers, which normally handle text data as simple sequence of strings. Therefore we need a specific preprocessing methods are required in order to extract meaningful information from the documents. For Tamil documents, few researchers have used Vector Space Model and Artificial Neural Network for Tamil text classification [1].

The automatic classification of Tamil text plays an important role in building Tamil corpus. The first corpus for modern written Tamil was built in the Central Institute of Indian Languages (CIIL), Mysore, consists of around 3.6 million words of written Tamil. The Mozhi Truest has also build a corpus of around 3 million words of written Tamil.

A. STEPS IN TEXT CLASSIFICATION

The process of automatic Tamil text classification contains the three major steps namely

- Document representation
- Dimensionality reduction
- Choice of classifier

Tamil Text Classification involves Document representation, stop word removal, Feature extraction, Vocabulary set formation and choosing an classifier for text categorization.

1) DOCUMENT REPRESENTATION

In document representation, a tokenization process is required that is a text document is split into a stream of words by removing punctuation marks and a single white space is used to replace the tabs and other non text characters. This tokenized representation of data is used for future processing in text classification.

2) DIMENSIONALITY REDUCTION (DR)

Dimensionality reduction (DR) is the process of identifying key attributes and thereby eliminating redundant and irrelevant keywords or attributes from the training corpus. This process aims at reducing the complexity of text classification. A number of machine learning, knowledge engineering, and probabilistic based classifier-induction methods have been proposed for text classification. The most popular methods include Naïve Bayes Classifier ie Bayesian probabilistic methods, regression models, example-based classification, decision trees, decision rules, Rocchio method, support vector machines and association rule mining.

3) DIMENSIONALITY REDUCTION (DR)

A number of feature selection methods based on entropy, statistics and optimization techniques had been proposed for text classification. Some important popular Feature Selection methods such as document frequency, information gain (IG), mutual information (MI), (CHI) statistics, and term strength (TS).

- Feature extraction: It is a process that tries to generate a set of “synthetic” terms T from the original set T which will maximize effectiveness of classifier used in text classification.
II. FEATURE EXTRACTION

In a traditional feature extraction method, a bag-of-words will contain the combination of single terms as a keyword which is used to represent a particular document, but here we have proposed a C-feature extraction method which will make the vocabulary set or bag-of-words containing a compound-feature to represent a document. Extraction of C-features is composed of two terms that occur in Tamil documents is done without any restriction on distance or group within the document using the pair of terms will identify the document category. The following keywords are examples of C-feature.

Fig.1: C-Feature from Tamil Documents

From the above Fig 1, we can understand the usage of compound feature i.e. C-Feature extraction in Tamil text classification. The first keyword example represents (eye, pain) and the second example represent (place, path). The first keyword represents the document related to medicine i.e. having pain in the eye and the second keyword represents the document related to travel i.e. identifying a route for a particular place.

Fig.2: Steps in Creating Vocabulary Set from Tamil Documents

The influence of a feature \( f \) in a class \( c \) is the conditional probability of \( c \) given the occurrence of \( f \) estimated the training set. Let \( F = \{ f_1, f_2, \ldots, f_n \} \) be the set of features associated with a collection \( C = \{ c_1, c_2, \ldots, c_m \} \) be the set of categories or classes that occur in a document collection, \( df(f_i, c_j) \) be the number of training documents associated with the class \( c_j \) which contain the feature \( f_i \) and \( Inf_{fact}(f_i, c_j) \) be the number of training documents associated with the class \( c_j \) which contain the feature \( f_i \). We define the influence of factor as

\[
Inf_{fact}(f_i, c_j) = df(f_i, c_j) \sum_{i=1}^{n} df(f_i, c_j) \quad -- (1.1)
\]

To obtain a high discriminative C-features that are pairs of s-features (terms) using (1.1). There are three steps are followed:

(i) we select the bests- feature \( s \) that will be used to build the C-features using information gain as the most important measure to rank the s-features.

(ii) In the selection step we select the generated C-features that will be used to increase the documents of the training and test Tamil sets i.e., vocabulary set of training and test Tamil documents. As the selection criterion we use C-features with high dominance for a given category.

(iii) Finally, in the augmentation step we insert only those C-features that have high dominance in the class of the training document for increasing the training documents. The increase of a test document is done by inserting all high dominance c-features that occur in the document.

The smaller the number of distinct classes where a C-feature occurs, the higher the influence factor. Influence factor or Dominance is important because it can be used to filter C-feature distributed unequaly in various classes and directly quantifies the relevance of which classes should or should not have their documents increased with a given C-feature.

The third step is to increase the document collection, which aims to add C-features that help the classifier in performing its task. We first perform the extension in the training set. An important thing here is to determine whether to include a C-feature to a document. We use influence factor to determine whether to include a C-feature in the training set. Suppose that a C-feature \( f \) is composed of \( s \)-features, where \( T \) is the set of s-features. Let \( C^{\phi}\{c_1, c_2, \ldots, c_k\} \) be the set of classes that occur in the collection and let influence factor \( f(c, e) \) be the influencing values of C-feature \( f \) in each class where \( c_j \) of \( C \).

Finally, we can use the classifiers to classify the Tamil text document using the classifiers vector space model, Naive Bayes classifier, and k-nn classifier. Using C-feature, we can easily assign a Tamil test document to its predefined category by its contents.

III. CONCLUSION

In this paper, it was proposed that a efficient method for extracting C-feature for classifying Tamil text documents. Since, there are more number of digital documents available in English, the classification of text documents are performed in English and also in some other languages. Using the C-feature extraction, we can easily classify the documents because C-feature will contain a pair of terms to classify a document to a predefined category. As future work, we have planned to implement the extraction of C-features from Tamil text documents and classify the document using the classifiers k-nn and SVM classifier to retrieve the Tamil text document efficiently.

REFERENCES


