Support Vector Machine Based Classification Of Hindi MWEs

Rakhi Joon, Archana Singhal

Abstract: Multi Word Expressions (MWEs) are one of the important concept of text processing. These are the frequently repeated word combinations used in spoken as well as written languages. The meaning of a phrase cannot be accurately determined without the correct identification of MWEs in a particular language. The classification of MWEs is very tedious task and it depends on various factors like the Part of Speech (POS) tags, grammatical constructs, and the linguistic information. A lot of work has already been done on MWEs in various languages but the Hindi MWEs have not gained much attention. In this paper, a Support Vector Machine (SVM) based classification for Hindi MWEs has been proposed. Hindi language is a morphological rich language and has very large set of grammar constructs. SVM provides a good platform for analysis of Hindi MWEs due to their sentence structure and idiomatic rich literature. SVM based classification is very efficient and preferred over other conventional methods due to robustness and automatic parameter selection techniques. The suggested approach provides efficient classification as new categories can be added in future if identified with ease.

Index Terms: MWEs; Hindi Multwords; Support Vector Machine; statistical measures; MWEs analysis; feature vector.

1. INTRODUCTION

Huge amount of text is generated from various sources now a days, and it is very important to analyze the text properly for extracting useful information. This useful information can be in form of Summary, Keywords, Multiwords, and many more. For extracting this information, the text should be managed properly. The management of large number of text documents and their classification require efficient systems to be developed. The text document classification systems are classified under two categories, first, information retrieval and second text classification [1]. In information retrieval systems, the text documents are identified by words or phrases extracted from the text. Whereas, text classification is the process of categorizing text documents collected from different sources into various categories. Text is evaluated on the basis of various features of text. MWEs are one of the text forms which has been discussed in the chapter. The text categorization will further lead to the concept of categorizing other forms of text like, MWEs. The extraction and classification of the Hindi MWEs had already been discussed in [2], [3] and [4] where various categories of Hindi MWEs had been suggested. In this chapter the classification of Hindi MWEs is revisited with the help of SVM for efficient representation. It is preferred over other classification methods due to its feature selection criteria and proved to be a very promising and easy to use classifier for Hindi MWEs.

2 LITERATURE SURVEY

The plain text Hindi documents were processed and analyzed for various applications by the researchers, but Hindi MWEs have not been much focused. Various authors have discussed about the classification approaches, like in [5], a formal classification of Hindi MWEs was suggested. In the proposed work this classification has been extended by adding new categories. Fuzzy-SVM based system was designed for the Indic languages by the authors in [6]. Fuzzy set models were applied for text classification of Hindi sentences using keyword extraction in [7]. Various classification models have been used for finding the context meaning of open class Hindi words in [8]. The distinction between types of words in Hindi language had been discussed in [9]. Comparison of various classifier approaches with the baseline methods for English-Hindi text simplification was done by [10]. Later the applications of classifiers were used in [11] and temporal expression recognition in Hindi was done. In [12], the authors have discussed the SVM based classification for Hindi sentences. There is no standard classification exists for Hindi, so the authors have suggested a standard classification of Hindi MWEs in [2] and [3]. In the proposed work, the SVM based classification has been proposed and designed for the Hindi MWEs.

3 SUPPORT VECTOR MACHINE

The concept of SVM is based on the supervised learning methods. Supervised learning methods perform classification of the data by constructing hyper plane which is used to separate the data. The hyper plane can be extended from 2-dimensional to n-dimensional depending on the complexity of the problem. The approach based on the SVM is very efficient for classification as well as for training data as SVMs have ability to generalize well in high dimensional feature spaces. SVM based classification is preferred over other conventional methods due to robustness and automatic parameter selection techniques. SVMs show better performance in the experiments. All these qualities make SVMs a very promising and easy to use method for developing language classifiers. In the proposed approach, a SVM based system has been developed to classify the Hindi MWEs using the feature vector which are further classified into various classes generating the functional classification of Hindi MWEs.

4 PROPOSED METHODOLOGY

The proposed methodology mainly deals with the classification of the Hindi MWEs using the Support Vector Machine. The dataset used for the experimental purpose is taken from the Hindi novel written by Munshi Premchand Ji, “Godaan”. It is based on the spoken language collected from conversation of

---

• Rakhi Joon is currently pursuing Ph.D. in Computer Science in University of Delhi, India. E-mail: rjoon30@gmail.com

• Archana Singhal is working as an associate professor in IP College for Women, University of Delhi, Delhi, India. E-mail: archanasinghal1970@gmail.com
the rural area people and thus contains different types of MWEs in various forms.

**Hindi MWEs SVM Classification Model**

![Diagram of SVM based classification model for Hindi MWEs]

*Fig. 1. SVM based classification model for Hindi MWEs*
The process model used for the classification of Hindi MWEs using SVM is shown in figure 1 and the various steps are explained below:

1. The text documents from the corpus are collected and the MWEs are extracted using the Hindi MWEs extraction process.
2. From the list of extracted Hindi MWEs, the feature vector is created using the frequency count of each category of MWE in text document. This frequency count of specific category of MWEs is taken as the feature value for that particular category.
3. The complete list of extracted MWEs is scrutinized on the basis of categories of Hindi MWEs existing in a particular Hindi Text Document (HTD) and the feature matrix is created for the HTDs and the corresponding feature values (Fn).
4. The Hindi text documents are evaluated by creating the corresponding feature matrix using the frequency count of the individual MWEs. For example, “पौलो पौताक” belongs to the Hindi MWEs “N+N” and it has been extracted from the HTD and added in the feature matrix for feature “F3” with frequency “1” in HTD1. Similarly, “मुनजन-मजन” also belongs to the same functional category and represented by feature “F9” with frequency “2” in HTD2.
5. The feature vectors belonging to common categories are combined to form the particular class of Hindi MWEs (Cn) extracted from the feature matrix. For example, the above feature vectors belonging to “N+N” are combined to for class “C1” which represents the functional category of the Hindi MWEs. The relevance of the particular class has been measured by its frequency count. More the frequency count, more relevant is the category of Hindi MWEs.

As shown in figure 1, the Feature matrix is mapped with the number of classes. There can be more than one feature vector mapped to same class and thus classification proceeds. The classified categories are based on the functional classification of Hindi MWEs. The classification of Hindi MWEs has already been proposed in [3], [4] and [5] but that was solely based on the POS patterns. SVM based classification provides a better alternative to accurately classify the Hindi MWEs, because the MWEs are divided into feature vectors on the basis of the frequency counts and the feature matrix is then mapped to the classes of functional classification of Hindi MWEs. The classification is suggested on the basis of feature vector created for each type of MWE, so it will include the categories which might have been skipped in other approaches. New categories can also be easily added in future using this classification technique. This is a step towards development of more accurate and efficient classification system. In the next section, the experimental results have been discussed.

5 Experimental Results
The experiments are carried out in order to classify Hindi MWEs. First the MWEs are extracted from the collected Hindi text and then extracted list is processed using the SVM for classification process. Each MWE corresponds to a feature value and the feature vector is created using the list of extracted MWEs. The feature vector mapping is shown graphically, where the feature value selected for a given class of MWE is mapped to the corresponding value of the frequency count of that feature. The frequency value selected for MWEs is called the Word Vector score. The feature mapping value is generated randomly on the basis of maximum number of features which can exist for a particular class and the word vector is selected for the actual frequency count value selected for the corresponding feature. For the Hindi text documents various features are shown using the different class representation in figure 2.

![Fig. 2. SVM based classification for Feature and word mapping](image-url)

The experimental results are based on a sample data collected from Hindi Novels. This approach can be applied to large dataset to get more accurate classification. Addition of more number of classes will not affect the existing system and thus the developed system become more efficient in terms of modifications as well as usage.

6 Conclusion
The SVM based Hindi MWEs classification system generates a wide range of applications in text processing. The proposed SVM classifier for Hindi MWEs classify MWEs into various classes based on the feature vector. The systems like ontology, focused on the classification only, but in case of SVM based classification the features of various types have been combined to classify the Hindi MWEs. The proposed approach can be easily adapted in terms of modifiability, as more classes can be added in the existing classification without modifying the existing system. This approach can further be extended to implement the SVM based classification model for Hindi MWEs for large corpus to get better accuracy and can be further analyzed using the multiple threshold method.

REFERENCES


