A NOVEL TOOL FOR COGNIZANCE MULTILIGUISTIC IMAGE ANALYSIS USING NLP

Mrs. R. MEENA
Assistant Professor, Department of Computer Science and Engineering
Sri Bharathi Engineering College for Women, Kaikkurichi, Pudukkottai.
r.meena.cse.90@gmail.com

***Abstract* - Language Identification is the process of determining in which natural language the contents of the text is written. Language identification is a fascinating field to be studied due to increased demand of natural language processing applications. Language Identification can be done using two types of techniques: computational techniques and non-computational techniques. Computational techniques are based on statistical methods and require large set of training data for each of the language while non- computational techniques require that researcher must have extensive knowledge about the language to be identified. In this work, a novel method is proposed to identify the language of the text using SVM classifiers. Once the language is identified, classes based on the frequency of their co – occurrence with other words are considered for further processes. Content Based Image Retrieval is used to display the images that are tagged for the words. While tagging, the language of individual words is identified using language models and dictionaries. When a word is displayed, the proposed method searches for the alphabets available in the dataset dictionary. It is followed by separation of the word into alphabets. Then, the tagged images for the appropriate alphabets will be displayed from the dataset. The proposed method is done for four languages viz., English, Tamil, Hindi and Malayalam. A dataset context is incorporated to improve the performance of the proposed method. This method is implemented and the results show that the proposed method is more efficient than the existing methods.**

INTRODUCTION

Research in recent years has given a lot of interest to textual data processing and especially to multilingual textual data. This is for several reasons: a growing collection of networked and universally distributed data, the development of communication infrastructure and the Internet, the increase in the number of people connected to the global network and whose mother tongue is not English. This has created a need to organize and process huge volumes of data. The manual processing of these data (expert , or knowledge based systems) is very costly in time and personnel, they are inflexible and generalization to other areas are virtually impossible, so we try to develop automatic methods.

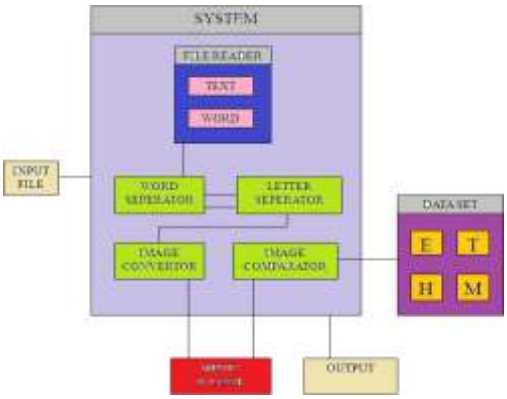

Figure 1.1.1 Overview of the process

ALGORITHMS AND TECHNIQUES

In order to perform computation, create four folders viz. English, Tamil, Hindi and Malayalam. Enrich all the folders with their respective alphabets' images. These folders are substitutes of training corpus. Another folder named Input folder is created to store input images

1. Get the input file. The file can be a text or word document.

```
Firstclass1
{
String name;   // gets the input file
}
```

2. Access the file and read it word by word.

```
Wordsep1
{
Fileread() // reads text file word by word
}

Readwordfile
{
Filereadword() // reads a word file word by word
}
```

3. For each word i, do the following:

- Create a new folder input$_i$ inside the Inputfolder.
- Separate the letters of this word.
- Convert each letter into images and store them in input$_i$.
- Compare input$_i$'s images with the images of predefined folders (English, Tamil, Hindi and Malayalam).
- If all the images of input$_i$matches a particular language, then that is the desired language.

```
Letterimg
{
Lettersep() //separates letters of each word
Texttoimg()//converts letter to an image and stores in inputi
}
Tclass1
{
Imgcomp()  // creates thread for image comparison
search()  // compares inputi  images with predefined folders.
}
```

4. The User Interface for the system is accomplished using the following classes.

```
Frontend
{
//calls the SecondFrame class
}
SecondFrame
{
// gets input for the system and activates Firstclass1
}
Sample
{
  // lists the set of languages and its words in the document
}
```

2. LITERATURE REVIEW

2.1 Title: "A survey of Language Identification Techniques and Applications", A Journal of Emerging Technologies in Web Intelligence, Vol 6, no 4, pp 388, November 2014. Author: Archana Garg, Vishal Gupta and Manish Jindal

Language Identification is the process of determining in which natural language the contents of the text is written. Language identification is always been an important research area which has been carried out from early 1970's. Still it is a fascinating field to be studied due to increased demand of natural language processing applications. In many applications, it works as a primary step of some larger process. A number of applications are outlined where language identification is working successfully. Language Identification can be done using two types of techniques: computational techniques and non-computational techniques. Computational techniques are based on statistical methods and require large set of training data for each of the language while non-computational techniques require that researcher must have extensive knowledge about the language to-be-identified. In this paper, a brief review of the few papers is presented which outlines the various statistical and non-statistical techniques that have been applied by the different researchers for language identification. Besides it, different researchers performed language identification for different type of documents such as monolingual, multilingual, long and short and for a particular set of languages.

2.2 Title: "Automatic Language Identification: An alternative unsupervised approach using a new Hybrid algorithm", International Journal of Computer and Applications, Techno-Mathematics Research Foundation, Vol. 7, No. 1, pp 94-107, 2010. Author: Abdelmalek Amine, Zakaria Elberrichi and Michel Simonet

It deals with research on unsupervised classification for automatic language identification purpose. The study of this new hybrid algorithm shows that the combination of the K means and the artificial ants and taking advantage of an n-gram text representation is promising. We propose an alternative approach to the standard use of both algorithms. A multilingual text corpus is used to assess this approach. Taking into account that this method does not require a priori information (number of classes, initial partition), it is able to quickly process large amount of data and that the results can also be visualized. We can say that, these results are very promising and offer many perspectives.

2.3 Title: "Automatic Language Identification using both N-gram and word information", International Conference on Statistical Analysis of Textual Data (JADT), Vol.2, pp 263-268, December 23, 1998. Author: Bruno M. Schulze

The predominant language of a Sample text is automatically identified using probability data that include N-gram probability data for at least one language and word probability data for at least one language. The N-gram probability data of a language indicate, for each N-gram, the probability that it occurs if the language is predominant. Similarly, the word probability data of a language indicate, for each word, the probability that it occurs if the language is predominant. The probability data are used to automatically obtain Sample probability data for at least two languages. The sample probability data include N-gram probability information for at least one language and word probability information for at least one language. The Sample probability data are used to automatically obtain language identifying data identifying the language whose Sample probability data indicate the highest probability. The N-grams can be trigrams, while the words can be short words of no more than five characters. Some languages can have both trigram and word probabilities, while some can have only trigram probabilities.

3. RELATED WORK

Multilingual posts can potentially affect the outcomes of content analysis on micro blog platforms. To this end, language identification can provide a monolingual set of content for analysis.

The unedited and idiomatic language of micro blogs is found to be challenging for state-of-the- art language identification methods. To account for this, we identify five micro blog characteristics that can help in language identification: the language profile of the blogger (blogger), the content of an attached hyperlink (link), the language profile of other users mentioned (mention) in the post, the language profile of a tag (tag), and the language of the original post (conversation), if the post we examine is a reply. Further, the methods combine these priors in a post-dependent and post-independent way. We present test results on 1,000 posts from five languages (Dutch, English, French, German, and Spanish), which show that our priors improve accuracy by 5 % over a domain specific baseline, and show that post-dependent combination of the priors achieves the best performance. When suitable training data does not exist, our methods still outperform a domain unspecific baseline. We conclude with an examination of the language distribution of a million tweets, along with temporal analysis, the usage of twitter features across languages, and a correlation study between classifications made and geo-location and language metadata fields.

Multilingual speakers switch between languages in online and spoken communication. Analyses of large scale multilingual data require automatic language identification at the word level. For experiments with multilingual online discussions, we first tag the language of individual words using language models and dictionaries. Secondly, incorporate context to improve the performance. Accuracy achieve of 98%. Besides word level accuracy, we use two new metrics to evaluate this task.

CBIR is one of the most widely used approaches for detecting images from an extensive image database. Now a day, numerous approaches have been developed to enhance the CBIR performance. The CBIR have a tendency to retrieve images depending on their visual content. CBIR evades several issues which are linked to the current ways of retrieving images by keywords. Most existing CBIR systems are based on color, text documents, informative charts, and shape of the pictures. A CBIR system takes the input query image and retrieves the similar images. The proposed approach here involves an efficient statistical feature extraction and further classification of the images by these features using Artificial Neural Network (ANN), Naïve Bayes Classifier and Fuzzy Neural-Network. The classifiers help to categorize the images according to the data set. The Precision and Error Rate have been calculated and compared according to the

retrieved content of the images from the datasets and the results have been shown.

The exponential growth in image data over the internet has resulted in a growing need for searching images according to our requirements. Content based image retrieval systems extract similar images from databases or the internet for facilitation of their users. A number of different feature sets and classifiers have been used by researchers for content based image retrieval. The goal of this research is to evaluate some common features sets used for classification of images and identify the best features depending upon the user requirement. Some commonly used features have been studies and a set of six feature sets have been selected for evaluation by the Back - Propagation Neural Network (BPNN). The results have been evaluated on the basis of precision and recall and it can be concluded that for natural images none of the feature sets perform well universally on all classes and the selection of optimal feature set depends on the type/class of images.

4. EXISTING SYSTEMS

Landmark retrieval is to return a set of images with their landmarks similar to those of the query images. Existing studies on landmark retrieval focus on exploiting the geometries of landmarks for visual similarity matches. However, the visual content of social images is of large diversity in many landmarks, and also some images share common patterns over different landmarks. On the other side, it has been observed that social images usually contain multimodal contents, i.e., visual content and text tags, and each landmark has the unique characteristic of both visual content and text content. Therefore, the approaches based on similarity matching may not be effective in this environment. Investigate whether the geographical correlation among the visual content and the text content could be exploited for landmark retrieval. In particular, we propose an effective multimodal landmark classification paradigm to leverage the multimodal contents of social image for landmark retrieval, which integrates feature refinement and landmark classifier with multimodal contents by a joint model. The geo-tagged images are automatically labeled for classifier learning. Visual features are refined based on low rank matrix recovery, and multimodal classification combined with group sparse is learned from the automatically labeled images. Finally, candidate images are ranked by combining classification result and semantic consistence measuring between the visual content and text content. Experiments on real-world datasets demonstrate the superiority of the proposed approach as compared to existing methods.

4.1.1 DISADVANTAGES

- Ignores the geographical correlation between images.
- When the value of neighborhood keeps increasing, the performance will degrade gradually

4.2 PROPOSED SYSTEM

A novel method is proposed to identify the language of the text using SVM classifiers. Once the language is identified, classes based on the frequency of their co – occurrence with other words are considered for further processes. Content Based Image Retrieval is used to display the images that are tagged for the words. While tagging, the language of individual words is identified using language models and dictionaries. When a word is displayed, the proposed method searches for the alphabets available in the dataset dictionary. It is followed by separation of the word into alphabets. Then, the tagged images for the appropriate alphabets will be displayed from the dataset. The proposed method is done for four languages viz., English, Tamil, Hindi and Malayalam. A dataset context is incorporated to improve the performance of the proposed method. This method is implemented and the results show that the proposed method is more efficient than the existing methods.

4.2.1 ADVANTAGES

- Image Re-ranking is included in this technique which helps the user to obtain the image result that is in the highest priority and in the least priority.
- This tool also provides a descriptive note of the image that is produced as the output which describes its surrounding, texture etc...

5. SYSTEM MODULES

- Input
- Word Separation
- Letter Separation
- Image Conversion
- Image Comparison

5.1 INPUT

- The system is designed to get a multilingual document as an input in text or word format. For instance, "hello.txt" is given as input.

5.2 WORD SEPARATION

In this module, the words of the input document are separated. For instance, the separated words of the above document are as follows**:**

Hello வணக்கம் चोर नमस्कारം

5.3 LETTER SEPARATION

The letters of each word are separated individually in text format respectively. For instance, the word Hello is separated.

H E L L O

5.4 IMAGE CONVERSION

The letters of each word are separated and converted into individual images respectively. For instance, the word Hello is converted into images as follows:

5.5 IMAGE COMPARISON AND OUTPUT

- This a computation approaches, that the system should have a prior data for language identification.
- The data used here is the alphabets' images of each language.
- The input images are compared pixel-by-pixel with the images in the data set.
- If all the images of the letters of a word matches the data set of a particular language, then that is the desired language.
- Repeat the above process for all the words in the document.

6. CONCLUSION & FUTURE WORK

The proposed model is used to identify the words of any of these four languages: Tamil, English, Hindi and Malayalam. The related images that are displayed give some idea of the letters of the unknown languages. In future, this method can extend for other Indian languages. The proposed method is planned to execute using a large dataset.

6.1 REFERENCES

[1] Abdelmalek Amine, Zakaria Elberrichi, "Automatic Language Identification: An alternative unsupervised approach using a new Hybrid algorithm", International Journal of Computer and Applications, Techno-Mathematics Research Foundation, Vol. 7, No. 1, pp 94-107, 2010.

[2] Archana Garg, Vishal Gupta, Manish Jindal, "A survey of Language Identification Techniques and Applications", A Journal of Emerging Technologies in Web Intelligence, Vol 6, no 4, pp 388, November 2014.

[3] Bruno M. Schulze, "Automatic Language Identification using both N-gram and word information", International Conference on Statistical Analysis of Textual Data (JADT), Vol.2, pp 263-268, December 23, 1998.

[4] Carlos Ramisch, "N-gram models for Language identification", December 21, 2008.

[5] Carter.S, Weerkamp.W, Tsagkias.M, "Microblog Language Identification: Overcoming the limitations of short, unedited and idiomatic text, Lang Resources and Evaluation", 47, pp.195-215, 2013.

[6] Aditya Koli, Diksha Khurana, Kiran Khatter, Sukhdev Singh, "Natural language processing: state of the art, current trends and challenges", Multimedia Tools and Applications (2023) 82:3713–3744.