Automatic Language Identification Using Multivariate Analysis

Vinosh Babu James J. and Baskaran S.

AU-KBC Research Centre
{vinosbabu, baskaran}@au-kbc.org

Abstract. Identifying the language of an e-text is complicated by the existence of a number of character sets for a single language. We present a language identification system that uses the Multivariate Analysis (MVA) for dimensionality reduction and classification. We compare its performance with existing schemes viz., the N-grams and compression.

1 Introduction

The rapid growth of the lesser-known languages in the Internet has created a need of language identification for applications like multilingual information retrieval, machine translation, spell checking etc. This task is complicated by three factors viz., the varying sizes of the character sets used to encode different languages, the usage of a variety of character sets for a single language and the same script being shared by more than a language.

The predominant technique used in written language identification is that of identifying short sequences of letters, characterizing a language (N-grams) [6]. These sequences may roughly be thought of as encoding the common, yet unique, character sets of a language. The Canvar's algorithm [6], chiefly aimed for text categorization and language identification is viewed as a task of text categorization with the different languages corresponding to different domains. In this method the accuracy directly varies with the increase in the number of N-gram statistics considered.

Dunning [5] uses Markov models in combination with Bayesian decision rules to develop language models for each language in the training data. These language models are then used to determine the likelihood of a test data generated by a particular model.

Benedetto et. al. [2] developed Shannon's ideas on the entropy of a language by suggesting that the compressibility of a given text depends on the source language. This method uses Lempel-Ziv [3] compression algorithm for identification.

We[1] used a combination of techniques for language identification with the N-grams (using Bayes' rule), Compression and the MVA using Principal Component Analysis (PCA). To our knowledge, this is the first work to use MVA technique for language identification. We show that the MVA method consistently outperformed the N-gram and the compression method in the penultimate section of this paper.

[1] The authors acknowledge contributions of Ramesh Kumar and Viswanathan.