

# Comparing Neural Network Approach With N-Gram Approach For Text Categorization

A. Suresh Babu  
Department of CSE

JNTUCEA  
Anantapur, INDIA  
asureshbabujntu@gmail.com

P.N.V.S.Pavan Kumar

Department of CSE  
JNTUCEA

Anantapur, INDIA  
pegatraj@gmail.com

**Abstract**— This paper compares Neural network Approach with N-gram approach, for text categorization, and demonstrates that Neural Network approach is similar to the N-gram approach but with much less judging time. Both methods demonstrated here are aimed at language identification. The presence of particular characters, words and the statistical information of word lengths are used as a feature vector. In an identification experiment with Asian languages the neural network approach achieved 98% correct classification rate with 500 bytes, but it is five times faster than n-gram based approach.

**Keywords**-N-Gram, Neural Network, Language Identification, Text categorization

## I. INTRODUCTION

As Internet services supersizing increasing in popularity, more and more languages are able to make their way online. In such a trend, a need exists for the rapid organizing of ever-expanding electronic documents. A well-trained librarian can easily identify the language of a book or a document, but it is not so easy when it presents online: there are so many documents in so many languages, and most of them cannot be passed immediately with a glance eye. Thus an automatically language identification system is needed to be built to take this task. Because of the sheer volume of documents to be handled, the categorization must be efficient, consuming as small storage and little processing time as possible.

Text classification addresses the problem of assigning a given passage of text (or a document) to one or more predefined classes. This is an important area of information retrieval research that has been heavily investigated.

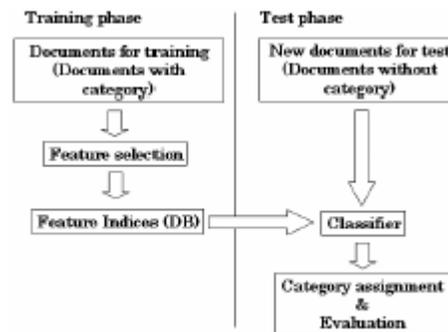
A segmentation-based approach was compared with the non-segmentation-based approach. N-gram based approach is the most widely accepted one and it is proved to have a good performance. As crucial as accuracy, the speed of classification is also a crucial factor for a classifier in a huge volume of categorization environment. However, most of the authors did not provide any information on the speed of classification.

In the following sections, the comparison of the performances of categorization algorithms using Neural networks and N-gram based approach are discussed. It is demonstrated that the identification rate of Neural networks is similar to the corresponding Ngram approach but with much less judging time.

## II. TEXT CATEGORIZATION

The goal of text categorization is to classify the given new documents into a fixed number of pre-defined categories. Fig.1 shows a flow diagram of the text categorization task.

The procedure for automatic text categorization is divided into two phases, the training phase and the test phase, as shown in Fig.1. In the training phase, we input the training documents along with a category. Next, we extract the feature term via a feature selection process and produce an indices database, referred to herein as DB, which is later used for the test phase. In the test phase, several new documents to be classified are input one after another, and one category is allocated in these documents.



**Figure.1.Data flow diagram of Text Categorization**

A variety of features can be used for language identification. In particular the following data are extracted for comparison purpose: the presence of particular characters, words and the word length distribution. An intuitive assumption is: some special words or short sequences are

unique to a particular language, and they have significantly different probabilities in different languages.

III. THE ARCHITECTURE OF NEURAL NETWORK

Neural networks learn to recognize groups of similar input vectors in such a way that neurons physically near to each other in the neuron layer responding to similar input vectors. The class information is used to fine-tune the reconstruction vectors in a Voronoi quantizer [9] so as to improve the quality of the classifier decision regions. A Neural network has two layers: a competitive layer and a linear layer. The competitive layer can learn subclasses. These, in turn, are combined by the linear layer to form target classes. Both the competitive and linear layers have one neuron for each class. The process of combining subclasses to form target classes allows Neural network to create more complex decision boundaries. The underlying philosophy of using a Neural network is based on the assumption that the variation due to different languages and different themes in same language cannot be ignored. In other words, it is assumed that training vectors from each language build an individual class in which there are some sub-classes formed by the inconsistent themes or domains in specific language. Each subclass is represented by a code vector to capture specific statistical characteristics of the text. The corpus is built to be a representative sample of interested population. Especially for balanced corpora, they put together so as to give each subtype of text a share of the corpus that is proportional to some predetermined criterion of importance. Since the text from different domains even in same language has discrepant statistical characteristic, for example, one would expect the entropy of poetry to be higher than that of other written text since poetry can flaunt semantic expectations and even grammar. Generally, the distance between two sub-classes within the same language is much shorter than that between different languages. The statistic discrepancy caused by different domains is often obtained from different segments of same language corpus, they can be simply ignored by the Neural network. Hence the linear layer will merge these subclasses into a class, which indicates they are in the same language. Fig. 2 shows the schematic diagram of our system, which includes the feature extraction part and the Neural network. It is assuming that 16 subclasses in competitive layer, refers each target language class containing two subclasses.

The choice of learning rate forces a trade-off between the speed of learning and the stability of the final weight vector. Considering the variation between different segment texts, the training patterns may have a huge difference. To achieve stable weight vectors, the learning rate is set with 0.1. The “conscience” technique [10] was used for avoiding the dead neuron problem plaguing Neural networks. Neurons which are too far from input vectors to ever win the competition can be given a chance by using adaptive biases that get more negative each time a neuron wins the competition. The result is every neuron has a chance to win. The learning rule for the bias for neuron  $i$  is

$$b_i^{new} = \begin{cases} 0.9 b_i^{old}, & i \neq *i \\ b_i^{old} - 0.2, & i = *i \end{cases}$$

Our corpora are collected in 8 Roman alphabet languages, i.e. English, German, French, Italian, Spanish, Swedish, Czech, and Catalan. They are drawn from the online newspapers. All of these data are encoded in Unicode, which contains accented characters and additional symbols. We divide the corpora into two parts: training set and test set. The training data is selected randomly from the sub corpus for each language and it is no overlapping with the test set. Both training set and test set are 100kB for each language. For each time of training, 500 characters of training data were used. For each ten times of training, we hold a test of 1000 random selections of a training set, and then record the accuracy to check if the Neural network has the best performance. The initial weights and biases are simply defined as zero. After 870 iteration, the learning algorithm has converged, since the accuracy attained the maximum of 97.6% for 500 characters. We maintain this set of weights and biases for the following experiments of performance based on the test set.

IV. RESULTS

Test data, which are constrained to not overlap with the training data, are also selected at random. Results presented here reflect the averages of 1000 random selections of a test set for each of the language involved. The length of test text can be 10 to 1000. Weights and bias of Neural network is set in the pattern introduced before, in comparison, rank-order and Cumulative Frequency Addition (CFA) algorithm based on bigram and trigram are involved. Taking the size of the weight matrix of Neural network as a standard, which only requires 2k bytes storage, we fix the size of N-gram profile that used by rank-order and CFA also to be 2k bytes. An obvious statement can be made from Figure 3 that the N-gram based classify algorithm performs better in short string, but when a string of 500 bytes is examined, rank-order classifier and Neural network have the similar good performance, that Neural network can achieve about correct rate of 97.6%, while 98.8% for rank-order. However, the CFA approach leads a relatively limited accuracy of 82.1%. This phenomenon is mainly due to the fact that, the

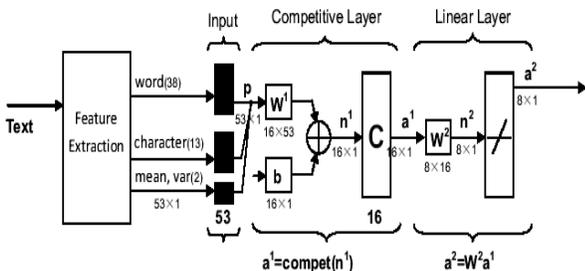


Figure.2. Schematic diagram of Neural network system

short-length samples usually lose some features to be used to train the Neural network. Also the features that are extracted from a sample may not have the statistical properties in line with the characteristics of training, all these result in a misjudgment in Neural network. At the same time, N-gram based algorithm takes full use of the statistics information of each word in a sentence, surprising that it usually has a higher accuracy level. Neural network presets an improved classifying capability as the length grows, and it performs quite satisfaction when the sample-length is long enough to reflect most of the features. Another concerned issue is: many researchers report that N-gram based classifier can achieve a very high accuracy with short string [3,4,5,7]. For strings of 100 bytes, the rank-order and CFA attained an accuracy of 98.97% [5]. However, it cannot be validated in our test, as shown in Figure 3, the CFA approach cannot even reach 90% in accuracy though the sample is longer than 900 bytes while the N-gram profiles in our experiment have been fixed in 2KB. A major reason is: without exception, N-gram based approaches need a table which stored each language in all N-grams, which is obviously a great need of storage space. When we fix the N-gram table to 2k bytes by removing unnecessary N-grams, there is somewhat a lost in statistical profiles, thus, the decreasing accuracy is natural.

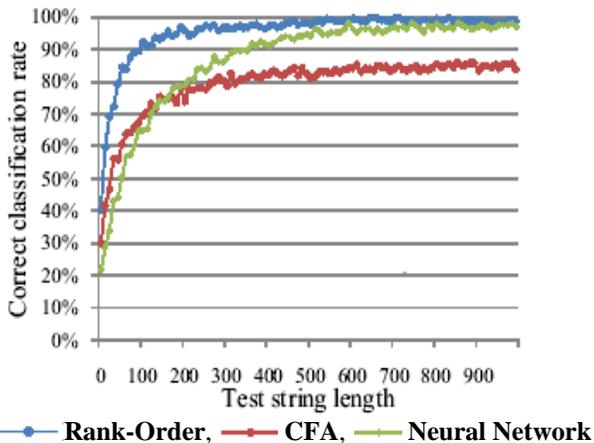


Figure.3. Percent correct classification of Rank-Order, CFA, Neural Network.

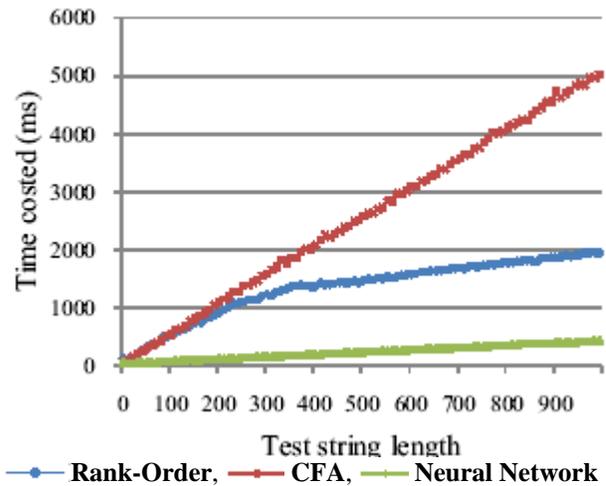


Figure.4. Speed of classification for Rank-Order, CFA, Neural Network.

In dealing with the processing speed, from the Fig.4, it can be seen that an Neural network has an inherent ability to classify more rapidly than the corresponding N-gram with rank-order and CFA algorithm. For a test paragraph of 1000 characters, the Neural network spends 476ms, while rank-order requires 2002 ms, and CFA costs 5023ms. An explanation to the speed advantage of Neural network is that it only needs a simple multiplication by feature vector and weight matrix to receive the verdict, while the rank order approach needs a sorted N-gram and makes a contrast in distance against the resulting N-gram of training. Obviously, sorting is a resource sensitive and time consuming operation. In CFA, it eliminates the sorting operation, it also needs to query repeatedly in the N-gram list. Thus, a test of a hundreds-word segment may need hundreds times of query. Even with Hash-table, this querying process takes a considerable time. Ahmed claimed that CFA is 3-10 times faster than the rank-order statistics method [5], without consideration of their the engine of Microsoft Access has optimized query, conclusion is based on an un-strict experiment. Given the fact, the text to be classified is usually quite a long one, such as news, essays and articles. In the view of situation, the high accuracy and efficiency of Neural networks achieve a great advantage.

V. CONCLUSIONS

A method for identifying the 8 Roman alphabet language using Neural networks is proposed in this paper. The performance comparison with other two Ngram-based approaches has been presented. Although varied researches over several decades showed that the N-gram based approach has an excellent performance on short strings. However, they did not provide any information on the size of N-gram profiles and the speed of classification. As we have shown in this paper, once the training finished, the proposed system requires small storage of 2KB only for the

weights matrix and biases vector. The Neural network based on the proposed design of feature vectors can be further distinguished by its high efficiency and accuracy of classification. The speed of classification of the proposed approach can be particularly useful when classifying text longer length on Internet.

#### REFERENCES

- [1] **“Machine learning for Asian language text classification”**, Journal of Documentation Vol. 63 No. 3, 2007 pp.378-397©Emerald Group Publishing Limited 0022-0418 DOI 10.1108/00220410710743306.
- [2] Makoto SUZUKI, *Member, IEEE*, and Shigeichi HIRASAWA, *Fellow, IEEE*,” **Text Categorization Based on the Ratio of Word Frequency in Each Category**”,1-4244-0991- 8/07/\$25.00 ©2007 IEEE, 3535-3540.
- [3] Juha Hakkinen and Jilei Tian,“**N-Gram and decision tree based language Identification for written words**”,0-7803-7343-X/02/\$17.00 ©2002 IEEE pp.335-338.
- [4] Tomáš ÖLVECKÝ,*Slovak*,” **N-Gram Based Statistics Aimed at Language Identification**”, M. Bieliková (Ed.), IIT.SRC 2005, April 27, 2005, pp. 1-7.
- [5] Bashir Ahmed, Sung-Hyuk Cha, and Charles Tappert, “**Language Identification from Text Using N-gram Based Cumulative Frequency Addition**”, Proceedings of Student/Faculty Research Day, CSIS, Pace University, May 7th, 2004, pp. 12.1 –12.8.
- [6] Dou Shen<sup>1</sup>, Jian-Tao Sun<sup>2</sup>, Qiang Yang<sup>1</sup>, Hui Zhao<sup>1</sup>, Zheng Chen<sup>2</sup> “**Text Classification Improved through Automatically Extracted Sequences**”, 8-7695-2570-9/06 \$20.00 © 2006 IEEE.
- [7] Munirul Mansur, Naushad UzZaman and Mumit Khan,” **Analysis of N-Gram Based Text Categorization for Bangla in a Newspaper Corpus**”, *Center for Research on Bangla Language Processing, BRAC University, Dhaka, Bangladesh*.
- [8] D.D. Lewis and M. Ringuette, “**A comparison of two learning algorithms for text categorization**”, *Proc. 3rd Annual Sympo. on Document Analysis and Information Retrieval (SDAIR)*, pp.81-93, 1994.
- [9] T.Kohonen,”**Self-Organizing Maps**,” *Springer*, Berlin, 1995.
- [10] DeSieno, D., “**Adding a conscience to competitive learning Neural Networks**,”*IEEE International Conference on Volume*, Issue, 24-27 Jul 1988 Page(s):117 - 124 vol.1, 1988.