

## **Abstract**

Semitic languages such as Farsi, Hebrew and Arabic forces special challenges for developing effective natural language processing applications. For example in Farsi most vowels are not explicitly marked in written language, a word written as “Srd” can be pronounced as “Sard”, “Sord” or “Serd” etc. This paper investigates letter grouping approach for morphological and phonological disambiguation in Farsi. A model suggested in this paper investigated the effect of letter ordering on predicting the correct vowel state. The suggested technique of using letter level statistics is not limited to Semitic languages but it is also generalisable in all natural language processing applications and it does not require any handcrafted linguistic knowledge.

A lexicon of 858 words containing an ambiguous Farsi (CVC) syllable was trained and tested through an Artificial Neural Network. In order to assess the significance of letter ordering a random letter ordering data set was also generated and tested through same model. Each letter in a word has a correlation with its adjacent letters; this value will be obtained through the learning sample available from an Artificial Neural Network (ANN). Hence every letter can be indexed with its correlation values and get mapped to the similar queries. When a new word is inserted to the system it will calculate the correlations between the letters and predict the vowel state of that syllable.

Support Vector Machine was also applied as an alternative method for training the data set. The results suggest that the letter grouping can effectively increase the accuracy of word disambiguation system which leads to better linguistic information retrieval. Furthermore support vector machine gave better results in longer processing time. NeuralPower and NeuroSolution were used for training and testing MLP and only NeuroSolution for training and testing with SVM. SPSS package version 12 was used to perform the statistical analysis of the data.