NEPALI UNICODE CONVERTER WITH PERSONALIZED TEXT PREDICTOR USING N-GRAM MODEL

Abstract**—**Nepali Unicode Converter is the simplest and easiest way to type in Nepali Unicode. It automatically converts Roman Nepali text into Nepali Unicode in any of the desktop applications supporting Unicode fonts. This Nepali Unicode is widely used in any media, machine, or browser to support various languages. This can be used in chatting, emailing, messaging, and many other applications. Furthermore with the growth in technologies and the internet, socializing has become much easier. People around the world spend more and more time on their electronic devices like PCs, laptops, mobiles for social networking, email, banking and a variety of other activities. Due to fast paced nature of such conversation saving as much as time possible while typing is necessary. Hence an application that predicts the next possible words is necessary. Predicting the most probable word for immediate selection is useful technique for enhancing the communication experience. The objective of this work is to design and implement a word predictor algorithm that suggests Nepali words that are being used more in combination with other words of the users, with a lower load for system and significantly reduce the amount of keystrokes required by users. The predictor uses methodology of the N-grams for text prediction. This research uses Maximum Likelihood Estimation method for making prediction table of most probable words after each N-gram. Stupid back-off method is used for prediction if Out of Vocabulary sequences encountered. The training data was scraped from various news portals and mixed into final training data by random sampling. About 80% of the total sentences were used for training and remaining 20% were used for testing the model. Vocabulary Size was 46,000. Accuracy of the model was about 48% for 4-Gram model. Perplexity of the 4-Gram model reaches down to 237.

Index Terms—Roman Nepali, Nepali Unicode, Nepali Text Predictor, N-grams, Stupid-Backoff