Application of Witten-Bell Discounting Techniques for Smoothing in Part of Speech Tagging Algorithm for Sinhala Language

Manoj Prasad Jayaweera[1], N.G.J. Dias[2]

The spare data problem is a major issue in part of speech tagging process with standard N-gram models. Since systems are trained from corpus and any particular corpus is finite, some perfectly acceptable N-grams are bound to be missing from the corpus. This is becoming a bigger problem when calculating transition probability and Maximum Likelihood Estimation (MLE) in a Hidden Markov Model based tagging approach.

But there are some techniques we can use to assign a non-zero probability to these "zero probability bigrams". This task of re-evaluating some of these zero-probabilities of N-grams and assigning them non-zero values, is called smoothing.

Witten-Bell discounting is one technique that can be used to handle spare data problem in N-gram algorithms that was introduced by Witten and Bell in 1991. Witten-Bell discounting is based on intuition about zero-frequency events. An unseen word is one that has not seen yet, so Zero-frequency N-gram is one that has not happened yet, when it does happen, it will be the first time we see this new N-gram. So the probability of seeing a zero-frequency N-gram can be modelled by the probability of seeing an N-gram for the first time. So the concept of Witten-Bell is the use of count of things we have seen once to help estimate the count of things never seen. Using this technique, we eliminated getting zero probability values for transition probability and Maximum Likelihood Estimation for sequence of words (N-gram) that is seen first time in our algorithm.

So with applying smoothing techniques in tagging algorithm for unseen word sequences, zero probability transitions can be eliminated and can assign non-zero probabilities, which enables tagging sentences with word sequences that is seen first time. The accuracy of the tagger was improved by eliminating zero probability occurrences. Hence, our tagger shows 91% of overall accuracy, with a considerable improvement compared with the previous work carried out for Sinhala language, since previous results have shown an accuracy around 60%.

Key words: Natural Language Processing, Part of Speech tagging, Witten-Bell discounting

[1] Virtusa (Private) Limited, No 752, Dr. Danister De Silva Mawatha, Colombo 9, Sri Lanka. mjayaweera@gmail.com

[2] Department of Statistics & Computer Science, University of Kelaniya, Sri Lanka. ngjdias@kln.ac.lk