ABSTRACT

Part of Speech (POS) tagging is important in various areas of Natural Language Processing. There are different approaches or methods that have been developed to automate the problem of assigning each words or tokens of a text with part-of-speech tag for general English and Western languages. Therefore, this research conducted to investigate POS tagging performance based on stochastic approach, such as N-Gram tagger (Unigram, Bigram and Trigram tagger) applied in clinical-text domain. In other hand, it was believed a supervised POS tagging requires a large amount of annotated training corpus to tag properly. To investigate it, this research performed experiments to understand how POS tagging performance could also affected with limited resources of available corpora which was implemented in different scenario of data optimization following with: Intra-institution, Inter-institution, and Mix-Institution.

This research utilize annotated training corpus consisting of ten clinical notes from Beth Medical Center and equal size of Partners HealthCare that were manually annotated the training corpus without knowledgeable expert from English linguist and Medical expert. Although, the results from all N-gram taggers applied in different scenario shows Mix-institution with 10 intra-institution cross-validation and also combined tagger that allows chain all N-gram taggers together does best tagging in both clinical-text corpora by estimated an accuracy of 84 % tested on Beth and Partners corpus that could benefitted the training process of POS taggers.

Keywords: Natural Language Processing, POS tagging, N-Gram Tagger, clinical-text