World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:12, No:10, 2018

Unsupervised Part-of-Speech Tagging for Amharic Using K-Means Clustering

Authors : Zelalem Fantahun

Abstract : Part-of-speech tagging is the process of assigning a part-of-speech or other lexical class marker to each word into naturally occurring text. Part-of-speech tagging is the most fundamental and basic task almost in all natural language processing. In natural language processing, the problem of providing large amount of manually annotated data is a knowledge acquisition bottleneck. Since, Amharic is one of under-resourced language, the availability of tagged corpus is the bottleneck problem for natural language processing especially for POS tagging. A promising direction to tackle this problem is to provide a system that does not require manually tagged data. In unsupervised learning, the learner is not provided with classifications. Unsupervised algorithms seek out similarity between pieces of data in order to determine whether they can be characterized as forming a group. This paper explicates the development of unsupervised part-of-speech tagger using K-Means clustering for Amharic language since large amount of data is produced in day-to-day activities. In the development of the tagger, the following procedures are followed. First, the unlabeled data (raw text) is divided into 10 folds and tokenization phase takes place; at this level, the raw text is chunked at sentence level and then into words. The second phase is feature extraction which includes word frequency, syntactic and morphological features of a word. The third phase is clustering. Among different clustering algorithms, K-means is selected and implemented in this study that brings group of similar words together. The fourth phase is mapping, which deals with looking at each cluster carefully and the most common tag is assigned to a group. This study finds out two features that are capable of distinguishing one part-of-speech from others these are morphological feature and positional information and show that it is possible to use unsupervised learning for Amharic POS tagging. In order to increase performance of the unsupervised part-of-speech tagger, there is a need to incorporate other features that are not included in this study, such as semantic related information. Finally, based on experimental result, the performance of the system achieves a maximum of 81% accuracy.

Keywords : POS tagging, Amharic, unsupervised learning, k-means
Conference Title : ICCL 2018 : International Conference on Computational Linguistics
Conference Location : Los Angeles, United States
Conference Dates : October 30-31, 2018