CS671
Unsupervised POS Tagging for Hindi using SVD2 Approach

Abhra Dasgupta, Saurabh Srivastava

Indian Institute of Technology, Kanpur

{abhra, ssri}@iitk.ac.in

Project Proposal
21st October, 2013

Abstract

In this report we provide an outline for building a **POS Tagger** for Hindi Language, using **Unsupervised** Clustering. The approach used here will involve the usage of Singular Value Decomposition, along with *k*-means Clustering Algorithm. The approach is explored by Lamar et al.[1] in their paper *SVD and Clustering for Unsupervised POS Tagging*. The authors have applied the approach on English Language Corpora, and have compared the performance with other Unsupervised approaches for POS Tagging involving Hidden Markov Model. We aim to apply the approach on some Hindi Corpora, and compare the efficiency of the Clustering with respect to a pre-tagged corpus.

1 Introduction

Many approaches have been taken to counter the problem of POS Tagging. Most of these approaches however, are Supervised in nature and require a pre-tagged corpora, often needed to be tagged manually by linguists. Such approaches, although provide good results, but are not directly applicable to languages which lack a considerably large tagged corpus. There have been HMM based approaches, applied to the task of Unsupervised Learning in the field. However, these models are not deterministic in nature, and hence, the output depends on probabilities and random numbers.

The SVD based approach was first tried by Schütze[2]. The paper by Lamar et al.[1] takes up from there, and modify the approach, by application of SVD twice (called *SVD2*), coupled with some other mathematical tricks, to improve the overall efficiency of the process. Since the Indian Languages still lack large tagged corpora, an unsupervised approach to POS Tagging such as SVD2 can be of help.

1.1 Problem Statement

Apply the SVD2 Approach to a Hindi Corpus and:

1. Find a set of Clusters, representing different possible Part-of-Speech Tags in Hindi.
2. Assign Labels to these Clusters, using a pre-chosen Hindi Tagset, using an *M-to-1 mapping*.
3. If available, compare the efficiency of the Clustering algorithm on a pre-tagged Hindi Corpus.

1.2 Related Work

There are two major attempts at Unsupervised Learning for POS Tagging and clustering. The earlier attempt was by Schütze[2], who showed that the Singular Value Decoposition process can identify principle dimensions (or words in this case) around which, other words can be clustered. The paper by Lamar et al.[1] introduced Unit Scaling of Vectors, and another iteration of SVD (over the clusters produced in the first iteration) to produce an even finer level of clustering. Although the algorithm is one that does not disambiguate between different instances of the same word (i.e. different instances of the same word are given the same tag), the accuracies achieved by the approach surpasses it's peers.

References

[1] Lamar, Michael and Maron, Yariv and Johnson, Mark and Bienenstock, Elie, *SVD and clustering for unsupervised POS tagging*. Proceedings of the ACL 2010 Conference Short Papers 2010: 215-219

[2] Hinrich Schütze, *Distributional Part-of-Speech Tagging*. 1995