# CS4740 Project 2 Proposal

Rahul Arora(ra487), Rudhir Gupta(rg495)

For the WSD, we intend to first use a supervised learning approach in which we would be using a machine-learning algorithm to capture the probable word sense.

**Machine Learning Algorithm**

The machine learning algorithm we would be using is Support Vector Machine (SVM) as it finds the maximal-margin hyper-plane that separates the training data into classes and could easily be applied to solve even the non-linearly separable data using kernel. We would be using the libSVM implementation and RBF kernel for the problem.

**Features**

Features to be used are:

- Part of speech tags of the Ngrams around the target word.

- Immediate words around the target word.

- Using dependency graph, use parent headword’s features (POS tag, word-name)

- also can use ‘gloss’ for the target word from the dictionary provided.

- local collocation

**Approach**

Parse training data -> Getfeatures, which we have described -> Apply SVM on features to build a classification model

Parse test data -> Get features for every word sense that we intend to disambiguate -> Apply test data On classification model -> Get Results

**Evaluation approach**

For the evaluation, we are going to segment the provided training data set into training and validation set, and would be evaluating the performance over it. Further, we intend to experiment the system on some selected Wikipedia articles on ambiguous topics like Cricket and Bank.