Automatically Selecting Best Features During Semantic Similarity Measure

**Lasang Jimba Tamang**

University of Memphis, Memphis, Tennessee

ljtamang@memphis.edu

**Abstract.** Feature selection during semantic similarity measure is one of the most important task. The performance of model depends on both the selection of features and appropriate number of features. In this project, we select best Fi number of features for i <= total number of features using chi square feature selection, build model, and evaluate the model, and finally the feature Fi that yields best performance.

1 Introduction

Semantic similarity measure uses different classification and regression machine line approach. One of the most tedious but most important step is selecting the best features to be used. The accuracy of model depends on how well the feature are selected.

In this project proposal, I like to automatically select best features that gives the best result for semantic similarity measure. In doing so, I will be use chi-square feature selection for selecting different number of best features, evaluate model with those different best features set, and report the features that gives best accuracy as best features.

2 Project Plan

First of all, the dataset will be processed so that they can be feed to chi-square method for feature selection and used as classification problem as well. Then we will select different possible best feature set using chi-square method. In third step, we will form model using each of those possible best feature set. Finally, we will evaluate the model accuracy and report the feature set as best feature set which gives the best model.

2.1 Data Set Formation

Our data set consists of 4514 data with 50 different features, their score and gold score for similarity score of two sentence. This data is generated by **deeptutor lab**, Institute of Intelligent System at University of Memphis, during SemEval-2017 International workshop on semantic evaluation. We call this **dataset 1** and we form **dataset 2** by just converting its continuous gold score value into discrete value so that we can use as input to chi-squared feature selection as well treat them as classification problem.

2.2 Feature Set Formation

We use dataset 2 and perform feature selection for i<= 50 times. At each i iteration, we ask the chi-square method to output i best features and called it Fi. In this way, we will have 50 different feature set where each Fi feature set consists of i best features.

2.3 Model Formation

We use 50 different features set from step 2.2 during the model formation. Using each feature set Fi for i<=50, we form both classification and regression model CMi and RMi respectively. Classification model will be use dataset 2 and regression model will use dataset 1.

2.4 Best Features Selection

Each CMi and RMi model will be evaluated using 10-fold cross validation. The feature set Fi will be selected as best feature if average accuracy of CMi and RMI is best.

3 Project Timeline

Estimated project timeline can be found [here](https://livememphis-my.sharepoint.com/personal/ljtamang_memphis_edu/_layouts/15/guestaccess.aspx?guestaccesstoken=BqYGCcgKQe2mvNRP8xI+O/SfbFPO0n/yaVOG+SvqTM4=&docid=2_114b47eafa07540ee88a75ba489b1dd8c&rev=1).