Automatic Labelling of Semantic Roles of English Sentences

Abstract:

In this paper, we present a system that given a sentence and the predicate (verb) of the sentence , we identify and assign semantic roles to the constituents.The system is based on the training of the SVM and incorporating features in strings as well ,in collaboration with numeric features to achieve better results.The training was done using the PropBank corpus ,from where various lexical and syntactic features were extracted from each training sentence. The testing sets were created from the corpus itself,which were not used in training. And ~85% accuracy was achieved in assigning semantic roles to pre-segmented constituents from the testing sets of the corpus.
The trained SVM was modelled to assign semantic roles automatically to any generic sentence passed in English . Proper User interface is built which takes a sentence and predicate position , and identifies the various constituents and assign semantic roles to the constituents with respect to the predicate.

Introduction

Semantic Role labelling is relatively quite new a field and not much significant work has been done in this regard . With parsing,part-of-speech tagging,named entity recognition etc. fields taking more shape and increasing accuracy , the semantic role labelling has also surfaced a lot and presents opportunities to increase accuracy and become domain independent.

The goal of semantic role labeling (SRL) is to identify predicates and arguments and label their semantic contribution in a sentence. Such labeling defines who did what to whom, when, where and how. For example, in the sentence “The actors are playing cricket.”, playing assigns a role to actors to denote that they are the players; and a role to cricket to denote that it is the game being played.Models of SRL have increased dependencies on the NLP tools available these days. As each of these tools are on the base of large corpus available and richly annotated data available in treebanks. But since these corpus are increasing and various domain data are coming with each passing day, the accuracy have increased many folds.

The model we proposed and developed was also based on the latest NLP tools provided in the NLTK.The propbank corpus and treeBank were used from the NLTK ,and also Stanford parser was used in the development and parsing of trees of newly inputted sentences.

In the first section , we have discussed the basics of semantic role labelling and how we approached the problem of assigning semantic roles to general sentences in English.Further, we discussed the features that are being used in the training of SVM, and how we are justified in dropping some features and accepting some, based on the contribution of them in the accuracy Thereafter, we discussed the parameters of SVM used for training the data.After that, the results of training and testing were discussed and in the subsequent section , the model developed to take any new instance and label it automatically is being discussed.Discussion and conclusion marks the end of the paper, where scope of future work is discussed.