Predicting Verb Relations between Subjects and Objects based on NELL Beliefs

Guo Chen GUOC@CS.CMU.EDU
Yilun Cui YILUNC@CS.CMU.EDU
Very Large Information Systems - SCS. Carnegie Mellon University, Pittsburgh PA.

Abstract

Relation prediction given the subject and the object is a new topic in Never Ending Language Learning project. Our prediction is based on a large amount of SVO tuple corpus. We predict the verb based on a Naive Bayes classifier. We propose strategies about stemming, prior and label sizes. We also propose a method to introduce category information into the Naive Bayes model. Experiments results comparing these strategies show our category approach significantly improve the accuracy of the Naive Bayes model.

1. Introduction

NELL

2. Related Works

3. Dataset, Features, Labels

The goal of our work is to find the verb relationship between a subject entity and an object entity. Hence to turn this into a classification task, verbs become labels. Subject, objects and their related attributes become features for classification.

We used two different datasets for this classification task. Our ultimate goal is to classify the belief instances that we find in the **NELL SVO** dataset. This dataset contains 220 million *Subject Verb Object Count* tuples that have been mined by NELL. However we use another dataset: **NELL RTW** formalized dataset to give us metadata over the subjects and objects in the **NELL SVO** dataset. This dataset provides similar information in the form of *Entity Relation Value* however, it provides hierarchical categories for each of the entities and values.

The same belief in the two different datasets are represented differently:
In the SVO dataset:

cat eat fish
In the RTW dataset:
concept:animal:cat   *concept:animalseatfood*
concept:food:fish

We plan to use the categorical data from the **NELL RTW** dataset as supplementary features for subject and object, giving additional information for classification.
We can project the entities in the **NELL SVO** dataset onto the **NELL RTW** dataset to find what the categories for each individual subject and object are.

4. Method

4.1. Naive Bayes

4.2. Stemming

4.3. Smoothing

4.4. Category Smoothing

4.5. Label Size

4.6. Metric

5. Experiment

5.1. Result Comparison

5.2. Prior

5.3. Smoothing Strategies

5.4. Label Set

6. Conclusion