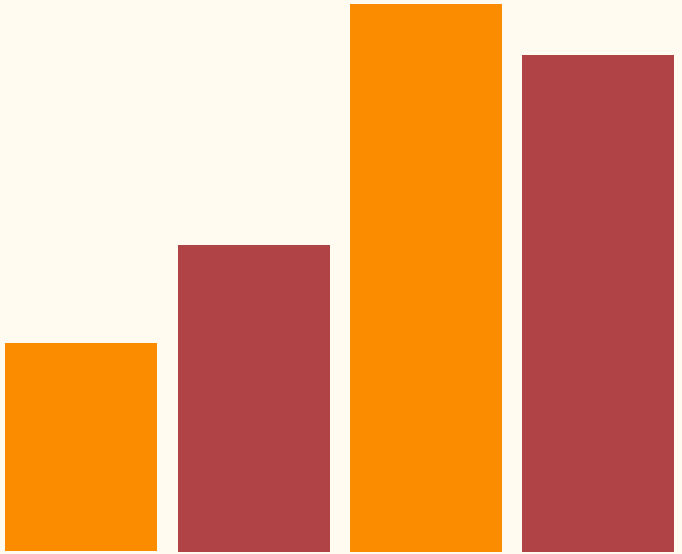
Statistical Methods in Artificial Intelligence

Team 26 - BOLTS

Nishanth Reddy
Sravya Varma
Poojitha Nandigam

The problem

Discriminating Similar Language (DSL),to predict the language of sentences written in similar languages



Definition

The discrimination of similar languages can be defined as the subtask of the Language identification problem. Language Identification is a fundamental task in the area of natural language processing.

Unlike well-separated languages, multilingualism, dialects of language can seriously degrade the quality of Language Identification. Discrimination of Similar Languages, noisy data, non well-formatted text, short sentences, mixed language are other examples of challenging problems in this field.

Dataset

The data sets we used were part of the DSL-Shared Task of 2015. The task provided participants with training (17000 examples per language), development (2000 examples per language) and testing sets (1000 examples per languages). The sets consisted of individually labeled sentences extracted from the journalistic corpora in 13 different languages. The languages were part of 6 language groups.

Method

Step1

Determine language group by word frequency method

Step 2

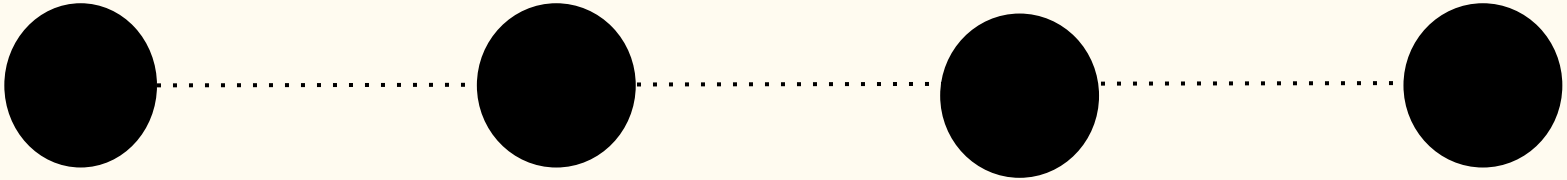
Train group-specific SVMs with character and word n-grams using training dataset.

Step 3

Rank features with tf-idf scoring and tune parameters using devel dataset

Step 4

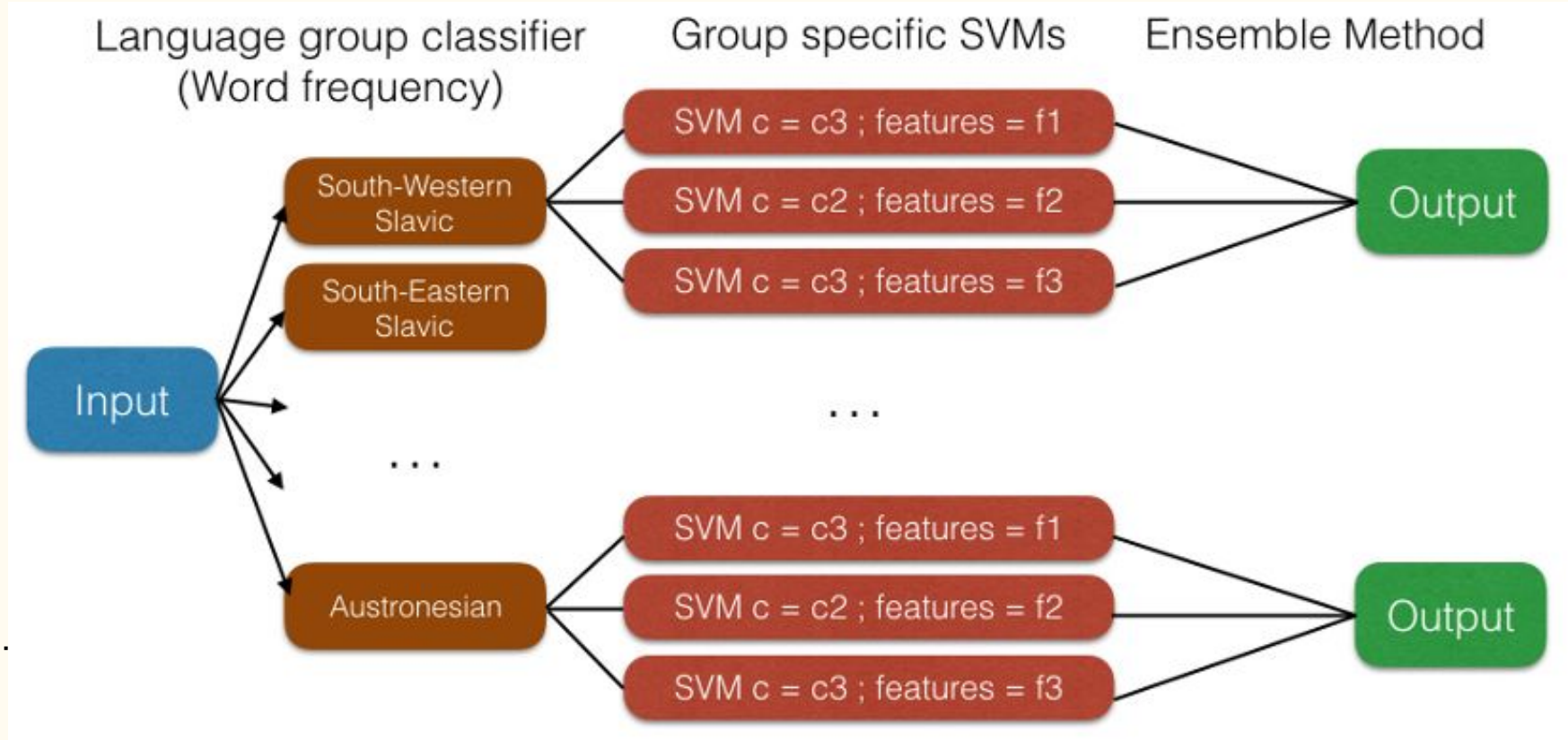
Combine decisions to predict test dataset using ensemble methods to reduce variance



Libraries

- LIBLINEAR SVM to implement multi-class SVM

The Solution



Scope of the project

- DSL Shared Task of 2015 made available the required corpus.
- Now, word frequency method is applied to group the data into 6.
- To differentiate between the similar languages present in the group , Multiclass SVM is used .
- To avoid curse of dimensionality, tf-idf scoring is used for feature scoring, and only high ranked features are considered for ensemble.
- Ensemble combines various SVMs with different hyperparameters. Ensemble method used is Mean Confidence.

Interim Eval II

- Implementation of
  - Word - Frequency
  - Multi class -SVM
  - tf-idf
  - Ensemble (Tentative)

Final Submission

- Implementation of the above mentioned method and 2nd method (From other paper);

Thank You