CS 388: Natural Language Processing Spring 2016

Project Report — Original Language Detection

Students: M. Denend, J. Hoffmann, L.Prakash *Prof. Ray Mooney*

1 Project description

(a) Description

The goal of this project is to be able to find the original language a text was written in, from the translated English version. We believe that even with completely bilingual translators, languages have a specific structure that will influence the translator in the way he creates the translation. We hope to recover this structure during our task.

(b) Methodology

From Project Gutenberg [], we built a dataset of books written in different languages, and translated into English. We then cut these books into slices of 500 sentences, and compute some features on these slices. After feature extraction, we feed these features into a classifier and test it on a completely different set of books.

2 Building the dataset

(a) Pitfalls

Since we built it ourselves, our dataset is not big enough for us to

(b) Description of the dataset

3 Features

(a) Lexical Feature

(b) Parse Features

(c) Homemade Features

4 Classification task

(a) Features Extraction

(b) Comparisons of different algorithm

5 Experiments

(a) Description of different experiments

(b) Results

6 Conclusion