CS446 Class Project

Determining Language of Origin in Hand-Translated Texts from the 19th Century

Ben Seefeldt (seefldt2)
Sean Wilner (swilner2)

December 3, 2013

1 Statement of the Task

With an increasingly ubiquitous global society, language barriers force large scale translation of text bodies into a myriad of languages. In order to better track the flow of information in such a mobile system, it would be beneficial to be able to determine the language of origin of a translated text. Specifically, we would like to be able to classify the original language in which a text was written given a translation of the text into English. As an example case, we will tackle binary classification between French and German languages of origin. These languages were chosen due to the high availability of hand translated texts and their dissimilarity despite both being Indo-European languages. These classication tasks will be completed over a variety of features pulled from the source text focusing on length of sentences, frequency of words, and part of speech ordering.

This cuts the project into two discrete tasks:

- Collect data, clean it, and distill feature sets from the cleaned data
- Multi-class classication over the distilled features

2 Outline and Discussion of Solution

As our task has two parts, our solution is best understood in it's respective components:

2.1 Data

In order to amass a large text body, scanned pages from a variety of translated books were taken and run through object character recognition. The generated text was then sanitized by hand for OCR errors.

The sanitized text was run through *The Stanford Natural Language Processing Group*'s CoreNLP parsing engine and part of speech sequences were extracted. The text was also sampled for all word counts and sentence count length to provide our learning algorithms with features.

2.2 Classification

In order to predict the language of origin, we plan on using two main learning algorithms and compairing their efficacy. First, we will train two HMMs on texts from each language separately using the tag sequences as an initialization. The resulting HMMs will provide us with likelihood estimates that a given test text's part of speech sequence would be generated given either of the underlying languages as a language of origin. Second, we will train a binary classifier using AdaBoost on the bag-of-words frequency counts. Our experiments will be run using 5-fold cross-validation, and the resulting accuracies of all 10 classifiers will be used to create one classifier using a weighted bagging algorithm. This algorithm will then be tested on a reserved test-set to obtain final accuracies.

3 Experiment and Results

Our testing set-up has some issues with the libraries used for syntactic text analysis. We are working on ironing those out and hope to have results shortly.