Chris Nakovski

Ling 472 project proposal

04/24/2015

Task: I propose to build a program that takes text from several different languages as training data, and is then able to read a sample of text and predict what language it's from. This is useful from the perspective of translating a language, where recognizing what language something is written in is the first step to translating it. This will be a simple form of machine learning that takes text from multiple languages, and calculates word frequency to store and later use to recognize whether other texts are written in the same language. I plan to use the frequency a word occurs in training data to weight the occurrence of that word in test data when evaluating likelihoods. If possible I would also like to recognize frequently occurring sequences of words to further fine tune language recognition.

Data: I plan to use texts written in various languages from the online eBook library Project Gutenberg at www.gutenberg.com as training data. I also plan to get my test data from there as well, but not so that I'm testing on training data.

Baseline: A baseline likelihood estimate would recognize the language of a text correctly with probability 1/n where n is the number of languages the system has trained on. In other words, it has to do better than guessing a language at random.

Evaluation: I plan to evaluate the program's effectiveness based on how high of a score it assigns a correct language sample. Ideally, a test dataset that matches some language should receive a high probability score, while one that doesn't should receive a low one.

In this case, precision would measure how much/little junk data interferes with a probability score. In high precision systems, junk data (like character sequences that are common to different languages) would be thrown out and wouldn't affect the analysis of some sample. Without precision, junk data interference could assign high scores for a language match that isn't correct. I propose to measure precision by calculating the percentage match for an incorrect language test. The lower the match percentage, the higher the precision will be.

Recall would be measured as the extent to which training data and test data are compared. In a low recall system, only small parts of the data would be compared (like one word). With high recall, the entire data should be examined and compared, as not to miss anything. I propose to measure recall by calculating the fraction of words recognized in the test data over the total words.