Machine Learning, Language Rules, and Statistical Strategies for Language Translation TJHSST Senior Research Project Proposal Computer Systems Lab 2009-2010

Andrew Runge

October 22, 2009

Abstract

Development of language translators, spoken or written, has most often used either rule-based or statistical strategies. In addition, machine learning is becoming one of the most efficient and effective methods for interpreting and deciphering text. Through the use of machine learning, the less common rule-based strategies may be implemented to greater effect. This project aims to use machine learning strategies to combine these two strategies to create an effective and efficient Latin translator. The project will be tested on several samples of Latin, including original Latin prose and poetry sections. The results will be studied for correct grammar, as well as compared to human translation of the same lines. The program will be done using python and the IDLE interface.

Keywords: Machine Learning, Statistical Translation, N-Grams

1 Introduction Research

Language translators have often been developed using two different strategies. Rule-based strategies are use to translate words properly so that their purpose in the sentence can be accurately defined. In addition, rule-based strategies have been used to try to properly put words in an intelligent order so that the resulting sentence makes sense. However, rule-based strategies

are often not proficient in assigning word order. Another problem with rule-based strategies is that they are relatively inefficient in the code, and take up a lot of time to simply discern the role of a single word in a sentence. However, by using machine learning strategies, such as n-gram generators, it can become quite simple and quick to generate information on each word in a sentence. This n-gram method of word tagging has been employed very frequently in this field. One such example is in the article Structural Tags, Annealing and Automatic Word Classification. This article detailed a method for determining the role of words in a sentence based on their context and similarities that they shared with other words. Using n-grams and machine learning, it is possible to quickly and easily translate a sentence, but there still remains the problem of getting the word order in the sentence correct. This is where the second method, statistical analysis, comes into play. Strategies demonstrated in the article Two-Stage Hypothesis Generation for Spoken Language Translation show the effectiveness of statistical generation of sentence structure. This project will combine the strategies outlined in that article, as well as rule-based strategies for word translation in order to create an effective and efficient language translator.

2 Goal

The goal of this project will be to create a Latin translator which is able to accurately and efficiently translate Latin sentences from all types of material.

3 Design Criteria Procedure

The first thing that my program must be able to do is correctly identify all the important characteristics about each word. For nouns, this would mean case, number and gender. For verbs, this would mean person, number, tense. After that, the next step will be ensuring that it can translate rudimentary sentences. To do this, I will attempt to implement machine learning strategies via n-gram generators which can be used to identify important things about the context of words. From there, the next step will be applying statistical translation strategies on more complicated sentences in order to properly sort out the word order. For this stage, I will use sections of Latin prose from famouse authors. These sentences tend to have complicated word order and

will provide a good test for my statistical alogrithm. The final step will be attempting to teach the to make assumptions about words in the sentence in an effort to have the sentence make sense. For this stage, I will test it on Latin poetry, which often has words left out, or uses some words to mean other things.

4 Scope

My goal for the middle of the second quarter will be to generate the dictionary of words that will be used by the translator. In addition I would like to start working on the machine learning side of the rule-based translation by beginning to create the n-gram generators which will identify characteristics of the words both based on their own endings as well as the context of the words. During the second half of the second quarter and third quarter, my focus will mainly be on statistical translation for word order. My goal is to have the translator working fully on Latin prose and be able to get word order largely correct in those kinds of sentences. If I have time, at that point I will begin programming the computer to be able to translate Latin poetry sections as well.

5 Expected Results

I will present the results using several examples of ttranslated lines, as well as comparisons with actual human translation of the lines.

It will give them a good view of statistical language translation, so that they could improve upon its design and attempt to continue from where my project leaves off.

I expect that my main goals will take me all the way up through the third quarter.