Machine Learning, Language Rules, and Statistical Strategies for Language Translation

Andrew Runge
Computer Systems Lab 2009-2010

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.

Background

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. Machine learning methods, such as use of n-grams for tagging words greatly improve the efficiency of these rule-based methods. One such example used by McMahon and Smith was a method for determining the role of words in a sentence based on their context and similarities that they shared with other words. Then the second method, statistical analysis, comes into play. Chen et al. demonstrate the effectiveness of statistical generation of sentence structure with their project using n-grams to create possible sentences. Attempts have been made to find more efficient methods than bigrams for determining word roles in sentences, such as an experiment done by Pla et al. However, their attempts were unsuccessful at creating a more efficient and accurate method, reinforcing my decision to use bigrams in my own machine translation program. Another experiment by Bowden discussed a method of tagging the words for every possible set of characteristics they could have, and then systematically narrowing down the possibilities until you can more easily order the words based on their characteristics in Latin. This is a similar method to one that I am employing.

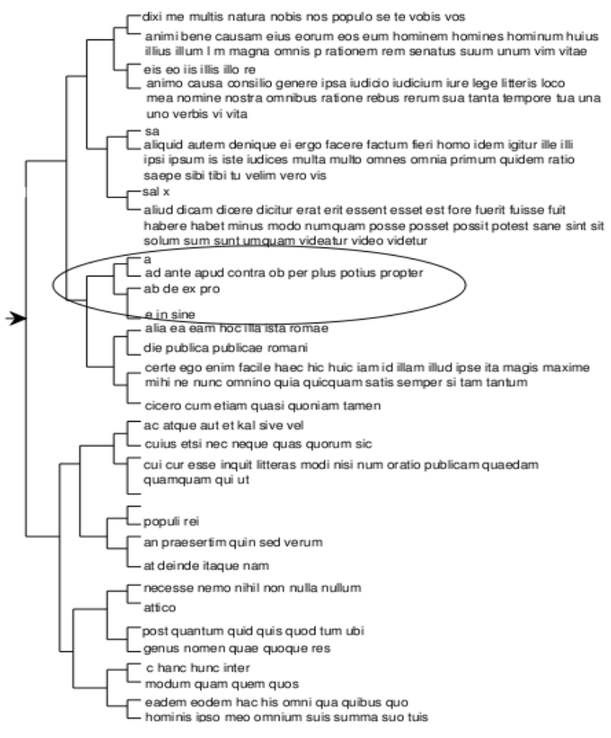

Figure 1: Tree of words sorted by sentence role from the assorted works of Cicero generated by the methods of McMahon and Smith

Original hypotheses	1. it's 5 minutes on foot . 2. it is 5 minutes on foot . 3. it's about 5 minutes' to walk . 4. i walk 5 minutes .
n-grams	it's 5 minutes, 5 minutes on, …… on foot ., about 5 minutes …… 5 minutes .

Fig. 2. Example of original hypotheses and 3-grams extracted from them.

partial hypothesis	it's	about	5	minutes	
n-gram expansion	+		5	minutes	on
new partial hypothesis	it's	about	5	minutes	on

Fig. 3. Expansion of a partial hypothesis through a matching *n*-gram.

New hypotheses	it's about 5 minutes on foot . it's 5 minutes . i walk 5 minutes on foot . ……
Reference	it's about five minutes on foot .

Figure 2: Demonstration of n-gram generation for determining word order in a sentence. Generated by Chen et al.

Discussion

I have finished the development of the dictionary which stores the words for later access for the purposes of defining and translating the words in the sentence. In addition, I've begun the tagging procedure to allow my program to tag nouns with all the possible sets of case, number and declension that the word could belong to.

Design

The next step for my project will be to improve the tagging process for nouns, as well as extend it to work for verbs and adjectives. At that point, I will begin the actual translation of the sentences. To do this, I will first find each word's actual meaning and then sort these meanings into an intelligible order in order to find the meaning of the sentence. After that, I will extend the functionality of my program to include more grammatical types.

Results and Conclusions

The goal for this project will be to create an efficient Latin translator, which will both be able to identify key characteristics of words, as well as organize them into a sensible English sentence. The project will be tested on various forms of Latin prose and evaluated compared to human translations of the same lines. As of now, the program is able to perform much of the initial work needed to translate Latin sentences, such as tagging the words and creating an easily accessible dictionary to reference.