Statistical Machine Translation

(Spanish to English)

Raghav Bashyal

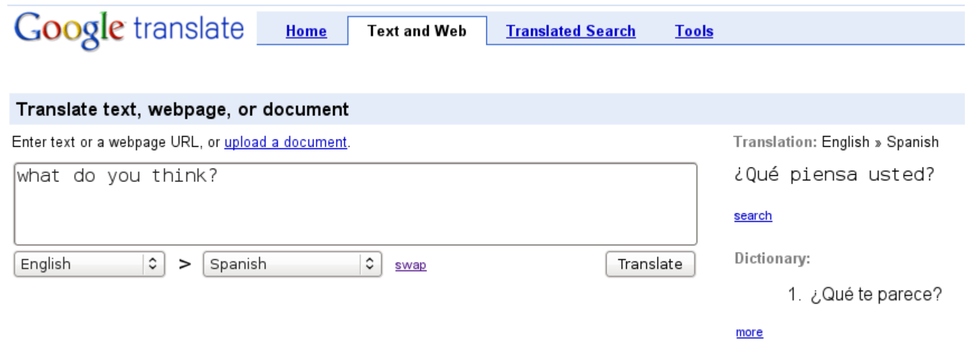
Abstract

Statistical Machine Translation (SMT) aims to learn a language much the same way a human would naturally by comparing a translation to its original text and attempting to associate words between the two. This project aims to build such a program. Although SMT implementations usually are capable of translating to and from any language, this study will focus on Spanish and English. It would then adjust the programming as well as the input to test the effectiveness of new and existing techniques.

Development

The Natural Language Toolkit and its auxiliary packages will compose this project. In addition to the functions that it provides, it has a system that allows mass amounts of data - texts, in this case - to be input in blocks called corpuses. I will be using the provided tools to translate and the corpuses for testing. The testing is fairly simple since the only thing that needs to be done is to compare the results to the available translations or checked manually for accuracy.

Currently my procedure is to study the NLTK book and practice with the problems in it. After I reach a point from which I can jump off into my real work, I will begin writing actual pieces of my program.



Google Translate - A fine example of Statistical Machine Translation

Expected Results

This project should be able to translate text from Spanish to English accurately, and also able to learn continuously form input data. The analysis and effectiveness can be presented by displaying sample translating with highlighted errors and with simple charts that show the frequency of such errors. The program should be able to identify some of its own errors in translation by using a reference-only database. Adjustments in the program, such as hard-coded components of the translation process or an algorithm meant to simply a procedure will be tested to see if they yield better translation results.