French/English Translation

by Sharon Ulery
TJHSST Computer Systems Lab 2006-2007

Abstract

This project uses a combination of hard-coding and statistical techniques in the field of computational linguistics to translate French to English and English to French well enough to be understandable to someone who knows only the output language.

Introduction

Starting with a word-for-word translation from French to English and vice versa, I will give the program grammar rules so that it correctly translates increasingly complex grammar structures. This project can change in size as needed throughout the year. At a minimum, the program should be able to deal with "subject verb object" type sentences with a wide vocabulary range in all tenses. At a maximum, the program will be able to translate all grammatically correct, non-idiomatic sentences in both languages with correct agreement of number and gender and context-specific translation of words with multiple definitions.

Purpose

This project uses computational linguistics to serve students of French or English as a non-native language as well as those who know only one of these languages. The output should be understandable even to someone who knows only the output language. Even a less than perfect translation is useful for surfing the web, reading texts in a foreign language, and communication with someone from another country. It can also be used for students to check their writing by translating back to their native tongue. They can check mechanics and make sure the writing is comprehensible by checking these areas of the translation.

Expected Results

This project will determine methods to best understand the French and English languages and to best translate between the two. Comparing the project with web-based translation programs will provide a standard to compare my program with. It will contrast the method of translation in which grammar rules are hard-coded into a program, which is the method I will primarily use, and the method in which the translation program "learns" grammar rules by using a few basic rules and going over a large corpus of written material in both languages (most current sophisticated programs).

At the end of the project, I will invite users to submit several sentences in either language to be translated. These inputs and corresponding outputs will be presented together to allow for analysis of the results. These outputs can be compared to those of several other, freely available automatic translators and determine which translations have more correct meanings and which are more grammatically correct.

Procedures and Methodology

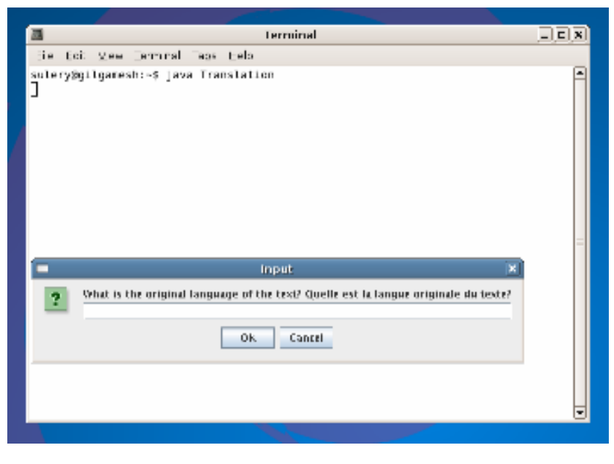
Input: The user will work through a GUI to first specify the original language of the text. Then, still through a GUI, the user will input the phrase to be translated.
Output: Output will be in the terminal. For grammatically correct input, output will be the grammatically correct, equivalent phrase of the input in the non-input language.
Requirements: The program is in Java. A large bilingual dictionary or a large, bilingual corpus (the Hansard corpus) from which to create one will be needed. A technique for inputing and outputing accent marks will also be needed.
Presentation of results is relatively simple. A list of a variety of sample inputs and corresponding outputs will allow clear evaluation of the results.
Testing is also relatively simple. As I increase capability, I will use specific structural and functional testing of the new grammatical structures the program is expected to include. From time to time, including at the end of the project, I will use dynamic testing to make sure that there are no hidden bugs that I hadn't thought to test.

Background

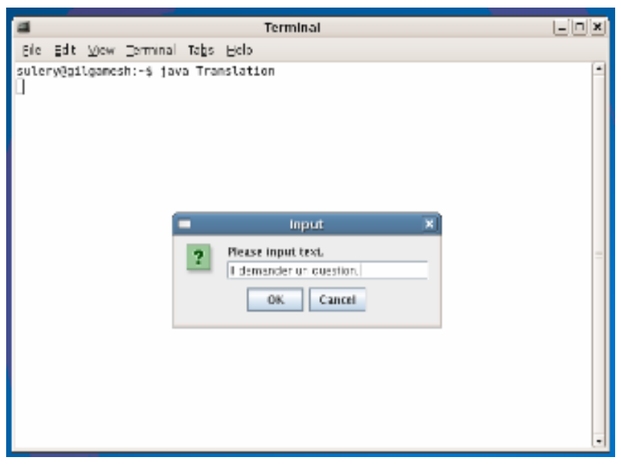
Most current in this area is far above the introductory level of this research project. I read *Foundations of Statistical Natural Language Processing: Chapter 1* by Manning and Schutze. Manning and Schutze believe that a division of language into "grammatical" and "ungrammatical" statements is an artificial and unsuccessful way of looking at it; rather grammatical structures should be seen as more or less commonly used. They use a method such that the program learns the parts of speech of words and common syntactical structures by training it on a large body of input text from a wide variety of fields, because this approach is much more robust than hardwiring all knowledge into the program at the beginning. I also read “Words & Transducers: Morphology, Tokenization, Spelling” and “Machine Translation” from *Speech and Language Processing, 2nd Ed.* 3 by Daniel Jurafsky and James M. Martin. The former source is a good introduction to some of the main problems in artificially "understanding" natural language. It talks about parsing words into morphemes so that a dictionary can be a reasonable size and begins talking about understanding words in context. The latter discusses the problems involved in machine translation and gives an overview of how to address many of these issues. I read Kevin Knight's "Automating Knowledge Acquisition for Machine Translation". This source is an overview of word-for-word statistics-based translation, syntax-based translation and semantics-based translation. It compared the usefulness of each and discussed high-level implementation techniques. Unfortunately, using the full statistical techniques recommended by these sources is far beyond the scope of a year-long, high school project.

Screen Shots:
Three stages of the Program

1)



2)



3)