TJHSST Computer Systems Lab Senior Research Project French/English Translation 2006-2007

Sharon Ulery

October 26, 2006

Abstract

This project uses computational linguistics to serve students of French or English as a second language as well as those who know only one of these languages. The program will translate French to English and English to French well enough to be understandable 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.

Keywords: computational linguistics, computer translation

1 Introduction - Elaboration on the problem statement, purpose, and project scope

1.1 Scope of Study

Starting with a word-for-word translation from French to English and vice versa, I will hard code grammar rules into the program 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.

1.2 Expected results

This project will determine methods to best understand French and languages and to best translate between the two. Comparing the project with web-based translation programs will provide a contrast between the method of translation in which grammar rules are hard-coded into a program (my program) 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 web-based programs).

Personally, I hope to learn more about computational linguistics and how computers can be said to "understand" natural language. I also want to get a deeper understanding of French and English grammar and, possibly, how humans understand language (from trying to imitate these processes artificially).

1.3 Type of research

This project is use-inspired basic research. I want to develop a useful translation device, but I hope that in doing so I will gain an understanding of the basics of computational linguistics.

2 Background and review of current literature and research

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. In this work, Manning and Schutze are considering the problem of having a computer "understand" natural language. They believe that language cannot be divided into "grammatical"

and "ungrammatical" statements; rather there are more and less commonly used structures. 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. Obviously, their methods do not produce perfect results, but this more modern approach is much more robust than the older approach of hardwiring all knowledge into the program at the beginning. It can be made to expand much more easily if the software grows by reading more text than if the programmers must write further grammar rules directly into the code. In fact, most current computer translation uses this type of statistical technique. Unfortunately, this approach is too sophisticated to learn and implement in a year-long project.

3   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 through a GUI. For grammatically correct input, output will be the grammatically correct, equivalent phrase of the input in the other non-input language.

Requirements: The program is in Java. An extensive English to French dictionary and French to English dictionary, each with translations and grammar information about each work, will be necessary for the program to be useful. 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.

4 Expected Results

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.