TJHSST Computer Systems Lab Senior Research Project
Statistical Machine Translation (Spanish to English)
2009-2010

Raghav Bashyal

October 29, 2009

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.

Keywords: Statistical Machine Translation, Spanish, English

1 Introduction

1.1 Scope of Study

The purpose of this project is to successfully implement Statistical Machine Translation techniques to translate from Spanish to English and to test the effectiveness of new techniques such as hard-coded syntax.

1.2 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.

1.3 Type of research

My research is in part applied research, as it focuses a lot on the implementation of the translation system. It does have components of use-inspired research as I will be testing various techniques to improve the translations made, which could be used for real-life research.

2 Background and review of literature

This project requires me to become familiar with the Natural Language Toolkit, a free tool commonly used for projects involving Natural Language Processing. to become familiar with the field and with the tool, I have read Statistical Machine Translation by Adam Lopez and Getting Started with Natural Language Processing with Python by Nitin Madnani. These pieces gave me an idea of what areas the field incorporated. Christina Wallin implemented the tool last year in her research paper: Naive Bayes Classication, where she tested a a new technique for classifying news into categories. In Improving English-Spanish translation, Preslav Nakov implemented techniques like paraphrasing and expanding recasing and tokenization with improved translations, although smaller text turned out worse than larger text. I am also working with the NLTK book to gather the knowledge required to implement my ideas.

3 Procedures and Methodology

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.

4 Expected Results

The results from this project would be shown in highlighted errors in translated text, as well as charts that plot the frequency of errors. The results would also show whether or not the variations implemented in the techniques provided positive results.

This project could be improved by implementing algorithms not implemented and changing things so that they may produce better results.