TJHSST Senior Research Project
Statistical Machine Translation
2009-2010

Raghav Bashyal

October 27, 2009

1 Purpose

The purpose of this project is (1) to successfully implement statistical machine techniques to translate from Spanish to English (and/or English to Spanish and (2) to test the effectiveness of a new or existing technique.

2 Background

This project requires me to become familiar with the Natural Language Toolkit, afree tool commonly used for projects involving Natural Language Processing. to become famililar 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 Classification, where she tested a a new technique for classifying news into categories. I am working with the NLTK book to gather the knowledge required to implement my ideas.

3 Procedure

The Natural Language Toolkit and its auxilary 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 weill 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 availbable translations or checked manually for accuracy.

4 Expected Results

I expect that this project will be able to translate texts from Spanish to English farily accurately, and to be able to learn from input data. The analysis of the effectiveness can be presented by displaying sample translations with highlightedd errors and with simple graphs that show the number of errors present. If the project were completed, i would imagine it gathering data from corpuses and coming up with a translation, and maybe even counting up the errors.

References

[1] A. Lopez, "Statistical Machine Translation", http://doi.acm.org/10.1145/1380584.1380586 , 2008

[2] N. Madnani, "Getting Started on Natural Language Processing with Python", http://www.umiacs.umd.edu/ nmadnani/pdf/crossroads.pdf , 2007

[3] C. Wallin, "Naive Bayes Classification", http://www.tjhsst.edu/ rlatimer/techlab09/WallinPaperQ4-09.pdf , 2009