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

Raghav Bashyal

June 16, 2010

**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. Using the Natural Language Toolkit and Python, this project began by creating a translator as describe by Dr. Kevin Knight. After an investigation of the vast and complex field of Statistical Machine Translation, this project aims to exercise this knowledge and spread it through the creation of a basic translator with a new, simple algorithm.

**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 in a simple Spanish-to-English translation algorithm designed to exercise the knowledge gained through a thorough investigation of the topic.

## 1.2 Type of Research

This project is, for the most part, applied research, as it focuses a lot on the implementation of the translation system. It does have components of use-inspired research as it does test a completely new algorithm for translation through Statistical Machine Translation techniques.

# 2 Background and Review of Literature

This project requires familiarity 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 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.

The purpose of this project is to investigate the field of Statistical Machine Translation and to successfully translate from simple Spanish to English. Dr. Kevin Knight, one of the renowned names in the field, in his guide to building a Statistical Machine Translation system, uses techniques that are complex and demand a lot of time and resources, both human and computational. To avoid this issue, to better understand the field, and to add to it, a new, simple algorithm was created and tested. There are two main parts to translating using Statistical Machine Translation: matching (aligning) words and checking that they are likely to exist. Dr. Knight uses a fairly straightforward and accepted technique to deal with the second part ' n-grams. An n-gram of a phrase can calculate likelihood by providing a number, which is derived from the frequency at which the components of the phrase appear in order in a certain area of text. For example, if a bi-gram (an n-gram that looks at two words at a time) were to be calculated for 'frankly bedazzle,' it would be done by counting the number of times 'frankly' is found starting a sentence in the English language, the number of times 'frankly bedazzle' is found, and the number of times 'bedazzle' is found ending a sentence. This bi-gram would not be very high.

After calculating the bi-gram, the probability of 'frankly bedazzle' can be calculated. Through the application of the Bayes rule, P(B—F) - the probability that 'bedazzle' is found given 'frankly' - the bi-gram of 'bedazzle frankly' is divided by 'frankly' to remove redundancy. This count - and the calculation of n-grams as well - depend upon corpuses. A corpus is a mass of text that is used to gather data and analyze it during translation. A good English corpus would contain as much documentation of English as possible, from different genres of writing. Through the use of n-grams and corpuses, Knight establishes the second step in Statistical Machine Translation, part of the heavily used 'IBM Model 3.'

Although the second step is fairly straightforward, the first step is dense and complicated, completed only by college students and professors in teams taking months to do so. Knight fails to explain in a straightforward manner, and, by looking at similar projects, it is clear that this method is much too difficult to do alone and with little time or expertise.

This new algorithm goes like this:

1. Match

a. Take small Spanish input

b. Look through the corpus to find instances of the input

c. Collect the Spanish sentences in which this input was found, as well as the English translation right below each sentence

d. Compare the English sentences to discover similar words

e. Find the most common similar words and find permutations of them

2. Check

a. Gather bi-gram values for each permutation using the bigram calculator

b. Calculate the probabilities for each permutation with Knight's formula

c. Return the most probable permutation as the most likely simple translation

This new, simple algorithm, by applying just a little of what Knight outlines, seeks to teach the basic principle of aligning words with each other in Statistical Machine Translation and to present a new idea that could be incorporated into a project or expanded upon to fit larger and more complex data.

# 3 Procedures and Methodology

The Natural Language Toolkit (NLTK) and its auxiliary packages helped compose this project. In addition to the functions that it provides, NLTK has a system that allows mass amounts of data - texts, in this case - to be input in blocks called corpora. The provided tools were used to access corpora and simplify the tokenization (separation into usable blocks) of text.

The sources of input for the program are the user and two corpora. The user provides a simple Spanish phrase (in this case, not longer than two words) that is then run through the first corpora, which is a parallel corpus that holds Spanish sentences and their English counterparts one after the other. After the word has been found, the translations of the found sentences are taken and analyzed to find words that match within. These words undergo a probability test - a bi-gram calculator uses the second all-Spanish corpus to determine the most-likely-to-occur translation.

There was really only one way to test if the algorithm worked - if it made a successful translation. Thus, at every step of the algorithm, the program was tested until the simple translator output its basic, English translation.

# 4 Results

The algorithm that was implemented was a success, and, despite its simplicity, may be useful in teaching someone about SMT or could be elaborated upon or included in a bigger picture.The actual test of the implementation of the algorithm was conducted with two simple corpuses, 'cosas' and 'monkey,' and the translating task being 'el mono' ('the monkey'). The code was able to go through at every step and return the right answer.

# 5 Conclusion

This investigation of Statistical Machine Translation yielded a deeper understanding of translation programs like Google translate - a tool that has analyzes a world of data and uses statistics to decipher the natural meanings of words. By developing my own, simple algorithm, I hope I have contributed to the study of SMT and that it will be useful for future projects.

# 6 Bibliography

Charniak et. Al. "Syntax-based Language Models for Statistical Machine Translation." Department of Computer Science, Brown University; Information Sciences Institute, University of Southern California.

Knight, Kevin. "A Statistical MT Tutorial Workbook." April 1999. JHU Summer Workshops.

Fonollosa, J. and Khalilov, Maxim. "N-gram-based Statistical Machine Translation versus Syntax Augmented Machine Translation: comparison and system combination." Proceedings of the 12th Conference of the European Chapter of the ACL, pages 424-432, Athens, Greece, 30 March - 3 April 2009. c 2009 Association for Computational Linguistic.

Melamed, Dan. "Algorithms for Syntax-Aware Statistical Machine Translation." Computer Science Department. New York University.

Palmer, Martha and Wu, Zhibiao. "Verb Semantics and Lexical Selection." National University of Singapore; University of Pennsylvania