Statistical Machine Translation

(Spanish to English)

Raghav Bashyal
Computer Systems 2009-2010

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.

Development

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.

Statistical Machine Translation Algorithm:

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

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

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.

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.