Statistical Machine Translation

(Spanish to English)

Raghav Bashyal

Abstract

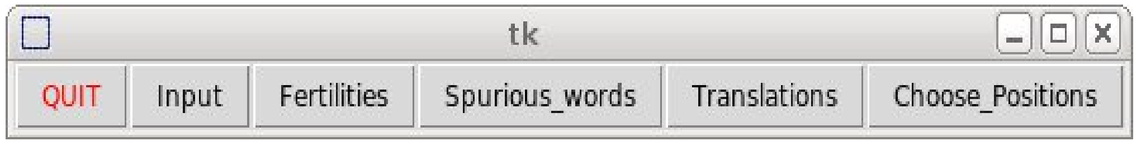
Statistical Machine Translation (SMT) aims to learn a language much the same way a human would naturally by comparing a translation to its original text and attempting to associate words between the two. This project aims to build such a program. Although SMT implementations usually are capable of translating 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. It would also test the effectiveness of combining syntax based translation, which is translation facilitated by hard-coded rules, with SMT.

Development

NLTK - The Natural Language Toolkit and Python are the building blocks of this project. NLTK provides all of the necessary functions needed for Natural Language processing. The program accepts input in the form of corpora – large masses of text. So far, the program accepts text input either from the NLTK database or with self-made corpuses.The ideal output comes in translation, which is in text and fairly simple. It will be analyzed by hand for accuracy, which will determine the effectiveness of the program. Currently, I am studying the NLTK book and learning Statistical Machine Translation through a worksheet. As I discover tools and concepts that can be implemented, I code them down. The program, as of now, can take the input form the corpus and tokenize it by sentences for clean strings. The program is also able to calculate the probability of words present in a certain amount of text. Although this is done pretty accurately, the range of the program is limited.

SMT Process

Using Kevin Knight's “A Statistical Translation Workbook,” I am learning the theory behind Statistical Machine Translation and the process to implement it. The goal, as shown in the figure, is to find the group of words that are most likely to be correct translations of the input text. The process begins with the calculation of basic probabilities, such as the probability of a certain word occurring in a certain amount of text. The process builds upon these probabilities with conditional probabilities, such as the probability of encountering a word after a certain word, which requires the use of the Bayes rule. These probabilities are expanded upon with N-grams, which describe the probability of a word occurring after a pair of words. N-grams are then smoothed with coefficients that use machine translation. Like this, the process eventually ends with coming up with the group of words with the highest probability of being the correct translation.



Translation GUI

Main Algorithm for Translation – Model 3

1. For each English word ei indexed by i = 1, 2, ..., l, choose fertility phi-i with probability n(phi-i | ei)
2. Choose the number phi-0 of "spurious" French words to be generated from e0 = NULL, using probability p1 and the sum of fertilities from step 1
3. Let m be the sum of fertilities for all words, including NULL
4. For each i = 0, 1, 2, ..., l, and each k = 1, 2, ..., phi-i, choose a French word tau-ik with probability t(tau-ik | ei)
5. For each i = 1, 2, ..., l, and each k = 1, 2, ..., phi-i, choose target French position pi-ik with probability d(pi-ik | i, l, m)
6. For each k = 1, 2, ..., phi-0, choose a position pi-0k from the phi-0 - k + 1 remaining vacant positions in 1, 2, ...m, for a total probability of 1/phi-0!
7. Output the French sentence with words tau-ik in positions pi-ik (0<=i<=l, 1<=k<phi-i)

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.