Using Genetic Algorithm for decoding in Statistical Machine Translation

By

Ali Reza Ebadat

arezae (at) gmail.com

University of Saarland, Germany
University of Nancy 2, France

Decoding algorithm is a main part of Statistical Machine Translation. Decoder algorithm is searching among all possible translation for given source sentence. The search space is so huge because of different possible translation for each word (phrase) with different ordering in sentence. We can consider decoding as a NP-complete algorithm. It is not possible to design an algorithm to search among all possible translation to find best translation. Researchers are looking for an algorithm to find optimum solution for decoding problem.

Different decoding algorithm proposed for Statistical Machine Translation which most of them are based on partial sentence evaluation. It means, these algorithms have access only to partial sentence and have limitation to decide about the translation quality. Beam search algorithm, Greedy decoder, stack decoding algorithm … are some examples.

In this thesis we proposed a Genetic Algorithm for Decode in Statistical Machine Translation. By using Genetic Algorithm we have the possibility of having access to complete sentence. In each step, the algorithm chooses some translations among translation candidates based on their fitness (by considering Translation Table and 3Gram Language Model). The algorithm proposed here is the first draft and it is the first Genetic Algorithm for Decoder so far. I use Language Model and Translation Table to calculate fitness value for each sentence.

To have a first evaluation, the proposed decoder used to translate 100 sentences and then the result compared with the Pharaoh output (using the same data) and also with Google translator. The results show that the algorithm can find good translation for words (it is word based decoder) but have some difficulty to find the best word ordering which is depend on Translation Model used. As I used only Translation Table (word translation) with 3-gram Language Model, the results were not good enough as Pharaoh (it uses Distortion Model) and Google (it uses larger Language Model and complete Translation Model). I have suggested using more statistical data (including distortion model, fertility model …) in order to have better fitness function. It can solve some translation problems.