NATIONAL INSTITUTE OF TECHNOLOGY, HAMIRPUR (H.P)

Innovative Research Incubation Club

Innovative Research Project Proposal Format

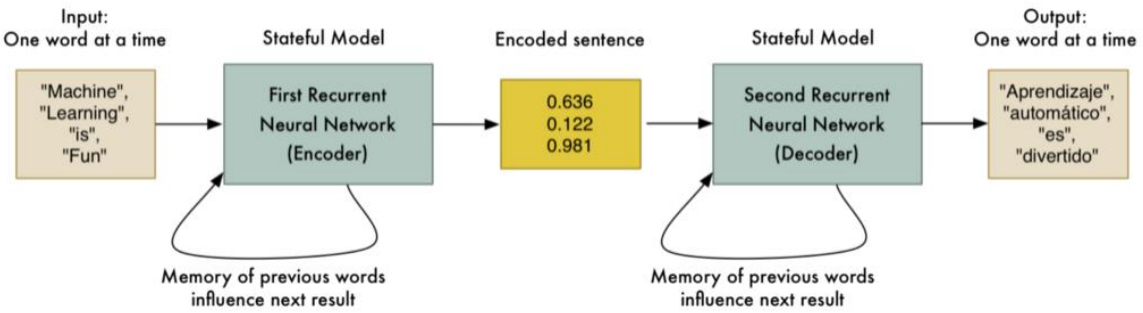
Research Project Title

Sanskrit to English Machine Translator

Project Description

Sanskrit, a very ancient language encomposes a very rich tradition of poetry and drama as well as scientific, technical, philosophical and religious texts. Huge volumes of texts are written in sanskrit are of interest to scholars of almost every field including mathematics, astronomy, history, political science, linguistics etc. To mention a few ancient text which are of academic importance - Panini Grammar (known to model any language in a computational way), Ayurveda (important to medical science), Kautilya Arthashastra (well known masterpiece of Political Science), Vedic maths, Vedic science, etc. It is tedious as well very time consuming to translate these huge texts manually in detail but with advancement in computational linguistics, it is now possible to do this task automatically with good accuracy without losing any detail and in no time. The aim of this project is to build a machine translator that would convert Sanskrit text to English.

Today the state of the art way to build a translator is using Deep Learning models which we plan to use in this project. Researches invented this only 3 years ago but it is found to be outperforming the traditional models based on statistics and probability. With Statistical Machine Translation, humans are needed to build and tweak the multi-step statistical models but deep learning models takes only training data and on its own uses it to learn how to translate between those two languages without any human intervention. This approach doesn't depend on knowing any rules about human language. The algorithm figures out those rules itself. Following is a general example how we might do it if we use RNN (Recurrent Neural Network):



The only disadvantage of this model is is that it is mostly limited by the amount of training data and computational power we have. So, the more training data and computational power we have, better will be our solution.

Already a lot of work has been done by scholars around the world on Sanskrit which has lead to the development of Sanskrit dictionaries, taxonomies, lexicons, etc which would be very helpful in our project. For example INRIA (The French Institute for Research in Computer Science and Automation) has made the Sanskrit Heritage Site (http://sanskrit.inria.fr/) that provide linguistic services for the Sanskrit language. Further, digitized version of many Sanskrit texts have been done by some institutes and made available online for public use. One such example is of http://sanskritlibrary.org/ which provides access to digitized texts and manuscript images, lexical resources, linguistics software, and computerized research and study tools that analyze and maximize the utility of digitized Sanskrit materials. This creates a very idle environment to make this project a success as we have the required training data and supporting tools which we can reuse and don't have to build everything from scratch.

Project Completion Time

One year

Student Skills Required

B.Tech. students of computer science and engineering with design, Analytical and problem solving capabilities. Familiar with the computer programming and text processing.

- Strong knowledge of Machine Learning, Artificial Intelligence and NLP
- Strong knowledge of Deep Learning
- Knowledge of linguists
- Good programming skills
- Sanskrit Grammar

Project Completion time:

Mid Term (Up to one Year)

Number of Students Required (UG/PG) for the Project:

One UG: Lokesh Sharma (CSE 3rd Year Student)

Name of Faculty Members: Dr. Lokesh Chouhan (Assistant Professor),
Garima Sharma (Lecturer).
Department: Computer Science & Engineering (CSE) Department, N. I. T. Hamirpur
Email ID: lokesh@nith.ac.in, garimasharma2007@gmail.com
Contact Nos.: 9827235155, +91-1972-254462, 9816845668