

Neural Machine Translation

| Deep Learning Project Phase-1

| 200968108 DSE-A 27 Batch-1

Problem Statement :

The ability to communicate with one another is a fundamental part of our daily life. There are nearly 7,000 different languages worldwide. As our world becomes increasingly connected, language translation provides a critical cultural and economic bridge between people from different countries and ethnic groups. Some of the more obvious use-cases include:

- **Business:** international trade, investment, contracts, finance
- **commerce:** travel, purchase of foreign goods and services, customer support
- **media:** accessing information via search, sharing information via social networks, localization of content and advertising
- **education:** sharing of ideas, collaboration, translation of research papers
- **government:** foreign relations, negotiation

To meet these needs, technology companies are investing heavily in machine translation. This investment and recent advancements in deep learning have yielded major improvements in translation quality.

According to Google, switching to deep learning produced a 60% increase in translation accuracy compared to the phrase-based approach previously used in Google Translate. Today, Google and Microsoft can translate over 100 different languages and are approaching human-level accuracy for many of them.

However, while machine translation has made lots of progress, it's still not perfect. 😊

Dataset :

The dataset used in this project for the task of converting/translating one language to another language is taken from the below website.

Tab-delimited Bilingual Sentence Pairs from the Tatoeba Project (Good for Anki and Similar Flashcard Applications)

If you don't already use Anki, visit the website at <http://ankisrs.net/> to download this free application for Macintosh, Windows or Linux. Any flashcard program that can import tab-delimited text files, such as Anki (free) can use these files. Warning! There are errors in the Tatoeba Corpus.

 <http://www.manythings.org/anki/>

We downloaded an English-German dataset that consists of bilingual sentence pairs from the Tatoeba Project. In this project, text in **English** language is being translated into **German** Language.

Metadata of the Dataset :


Each line in the dataset is a tab-delimited text consisting of an English text sequence, the translated German text sequence and some information regarding attribution. Note that each text sequence can be just one sentence, or a paragraph of multiple sentences. In this machine translation problem where English is translated into German, English is called the *source language* and German is called the *target language*.

Exploratory Analysis :

The text data available at the above mentioned website is present in the raw form(consists of punctuations, symbols, letters in both lower & upper case). So, pre-processing has been performed on the downloaded dataset and has been made ready! to be used for implementing various Deep Learning models and architectures.

There are 2 columns. One column has English words/sentences and the other one has German words/sentences. And this dataset can now be used for language translation task.

Google Colaboratory

 https://colab.research.google.com/drive/1fAlkxthovv399UqXnTu3l9AbupVFb_IH?usp=sharing



Preprocessing Pipeline :

Below are a few pre-processing steps implemented on the downloaded dataset:

1. Load & examine the data.
2. Cleaning the data. (includes the following)
 - Splitting each sample/text into English-German pairs.
 - Converting the data into an array for easy implementation.
 - Reducing the size of dataset to save the computation cost.(Only in this case)
 - Removing irrelevant text like attribution details
 - Removing punctuations.
 - Converting the text to lower case.
3. Tokenizing & vectorizing the text into numerical sequences.
4. Padding those sequences with 0's to bring them to same length.

Objectives :

The goal is to have machines translate content well enough for human translators to understand its meaning and easily improve upon the text.