Sarvam Fellowship Project Documentation

Aim:

This project discusses the Procrustes method to translate a language from one to another. (English to Hindi).

Packages Used:

- Gensim
- Numpy
- Scikit-learn
- Scipy
- fastText

Data Preprocessing (Extraction and Transforming of Data):

Whenever we try to work with text data, it is good practice to remove unnecessary data to gather better insights.For the given problem statement, we are obtaining data from a pre-processed framework(fastText embeddings), that reduces the amount of work done in the Extraction and Transformation phase of the project.

Loading Data:

The text embeddings are downloaded from fastText and uploaded to Google Colab. We then load them into our code using Gensim as word2vec text embeddings.

Training the Model:

To perform translation of the text given, we have to follow certain steps beforehand:

1. Organizing Data into Clusters
2. Dictionary Creation
3. Applying Procrustes Method and map text

Organizing Data into Clusters:

From the loaded text vectors, we map the source and target text values of our text to perform the necessary transformations to translate text.

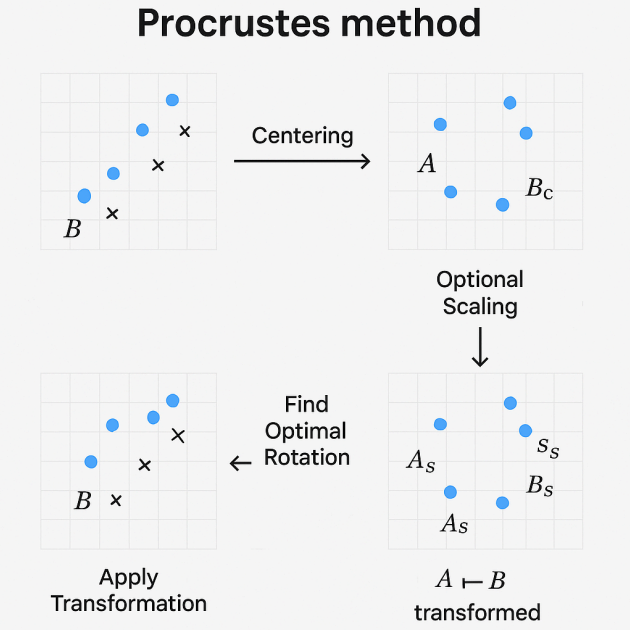
Dictionary Creation:

Now, we need a dictionary to train our model to perform translation. To do this, we download a dictionary from the MUSE dataset. As per the requirement, we restrict the number of words present in the vector embedding to the top 1,00,000 words.

Applying Procrustes Method and map text:

The Procrsutes method is used to align two feature vectors to find out simlarities using a similarity metric. This method can be used to find out cross-lingual similarities between two languages. For our purpose, we define dictionaries and utilize the respect word embedddings of English and Hindi to perform translation. We then use the obtained mapping to translate text from one language to another.

Sarvam Fellowship Project Documentation



Results and Observation:

The results obtained from running the Procrustes Method to perform Translation is given below. From the results shown in Table 1, we observe that an increase in Dictionary Size will be beneficial as it provides more information on the text that we are working on.

Demo of translation:

cat -> ['बिल्ली', 'कुत्ता', 'कुत्ते', 'पालतू', 'कुत्तों']
house -> ['मकान', 'फ़्लैट', 'घर', 'कमरे', 'हाउस']
river -> ['नदी', 'किनारे', 'गंगा', 'चनाब', 'नहर']

Dictionary Size	Precision@ 1	Precision@ 5
5000	0.6667	1.0000
10000	0.6667	0.6667
20000	0.6667	1.0000

Table 1: Results of Ablation Study

Sarvam Fellowship Project Documentation

Precision Values:
Precision@1: 0.6667
Precision@5: 1.0000

Conclusion:

By performing the given experiment, we understand that the Procrustes Method is an efficient method that can employed with minimal resources to perform complex tasks such as translation.

Resources:

- ChatGPT (Deep Analysis on Procrustes Method and Image Generation for Procrustes Method).
- MUSE dataset and pre-trained embeddings: https://github.com/facebookresearch/MUSE
- FastText: https://fasttext.cc/
- Procrustes alignment method: Described in "Word Translation Without Parallel Data" by Conneau et al. (2017)