Team Name : **Na Bien Na Mal**
Team Members : **Akshay Goindani (20171108), Zubair Abid (20171076)**

Reference paper: https://www.aclweb.org/anthology/P18-1073.pdf

Outline:

The Problem

Multilingual word embeddings are hard, but are needed if we plan to implement NLP solutions in the real world. Particularly in the Indian context. Indian social media, or any form of communication over the internet amongst Indians is not of the type that'll work with pre-implemented monolingual solutions, as there is a lot of code mixing and switching involved, often with more than two languages. The issue is aggravated when we realise that Indian languages are low resource, so supervised learning of embeddings becomes a challenge. We need a way to train such embeddings in a completely unsupervised way, or at the very least with minimal supervision.

Reference Paper

https://www.aclweb.org/anthology/P18-1073.pdf.

The paper proposes a completely unsupervised way to train multilingual embeddings by " exploit(ing) the structural similarity of the embeddings, and a robust self-learning algo-rithm that iteratively improves this solution."

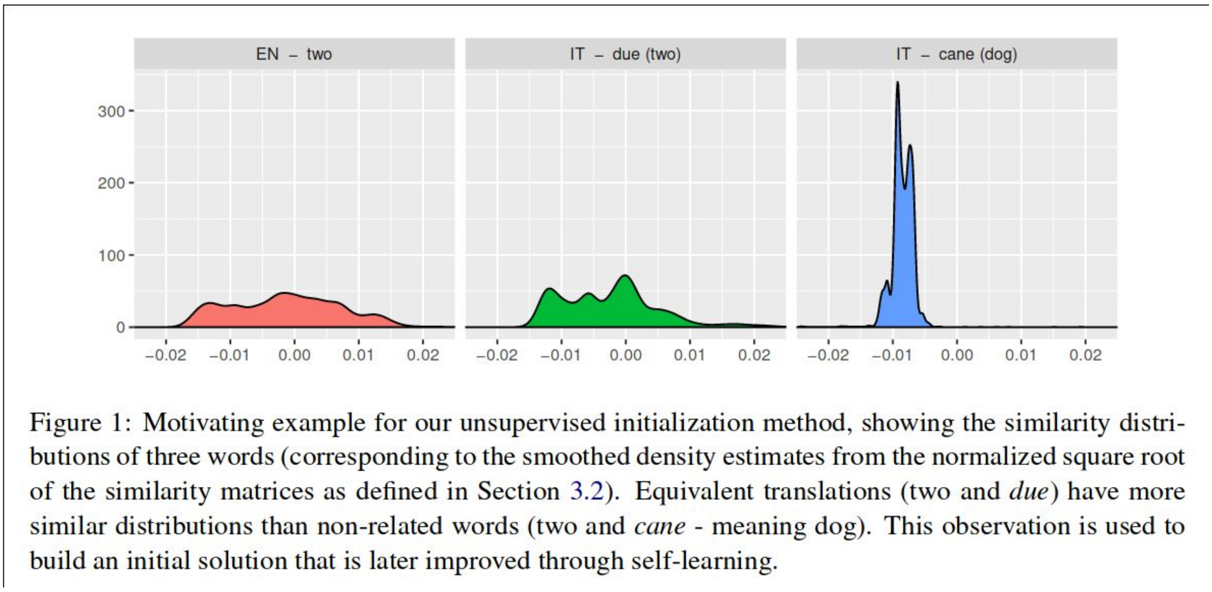

Figure 1: Motivating example for our unsupervised initialization method, showing the similarity distributions of three words (corresponding to the smoothed density estimates from the normalized square root of the similarity matrices as defined in Section 3.2). Equivalent translations (two and *due*) have more similar distributions than non-related words (two and *cane* - meaning dog). This observation is used to build an initial solution that is later improved through self-learning.

Fig 1: Assumption by the authors that they empirically verified in the paper

The paper rests on the assumption that words that mean the same thing in different languages are similarly distributed. It verifies this assumption for English and Italian, and then proceeds to suggest a completely unsupervised Initialisation scheme for the training, and then run their proposed algorithms on the data to get results. The paper claims to have bettered the SOTA in the field, comparing against several other methods previously proposed.

Problem Breakdown: Tasks

The end goal is to try and replicate such results, or similar, on an English-Indian Language pair. It can be divided into several largely sequential tasks.

1. **Verifying that the assumption holds for Indian languages and English** The paper used Italian, which English has borrowed a fair bit from and which exhibits significant similarity to English, more so than some Indian language would. The task would be to verify whether or not the method would work anyway. In case of failure, we shall go ahead anyway to see how off the results can get.

2. **Pre-processing to normalize the embeddings** A step from the method proposed

3. **Fully unsupervised initialization scheme that creates an initial solution** This is an integral part of the solution, and a lot will rest on implementation for Indian language-English pairs.

4. **A self learning procedure to improve the solution** A step from the method proposed, and the final one required for a result.

5. **A final refinement step that further improves the resulting mapping through symmetric re-weighting (Optional)** This is the final step suggested by the paper, but might be skipped due to constraints of time

6. **Verifying results against other methods (Optional)** This , too, might be skipped due to constraints of time

Problem Breakdown: Resources and Parameters

- Language pair chosen : Hindi - English

- Link to the dataset : https://github.com/joshua-decoder/indian-parallel-corpora/tree/master/hi-en

Faculty Signature