NLP Lab 1

Sub-topic: Can we map from one embedding space to another where the embeddings are trained on different data?

Darshan Singh
2441307

Bwateketa Kabengele
3046857

***Index Terms*—Natural Language Processing, Word2Vec, Word Embeddings.**

I. Introduction

The ability of Word2Vec models to represent words as dense vectors in a low-dimensional space has been a significant advancement in NLP. These embeddings capture fine-grained semantic and syntactic regularities, such that words with similar meanings are located close to each other in the vector space.

II. Methodology

A. Data Preparation

Two separate corpora were created by splitting the full text of *Harry Potter and the Goblet of Fire*. Corpus A consists of the first half of the book (561080 characters), and Corpus B consists of the second half (561080 characters). For each corpus, the text was converted to lowercase, and all punctuation and stop words were removed.

From the cleaned and tokenised text, a vocabulary was constructed for each corpus, limited to the 10000 most frequent words to maintain computational feasibility. Each vocabulary was used to create a set of training pairs based on the skip-gram model with a context window size of 2.

B. Word2Vec Model Architecture

For each corpus, an independent Word2Vec model was implemented in PyTorch. The model architecture consists of:

1) An **Embedding Layer** that takes a word's integer index as input and outputs a 50-dimensional dense vector.
2) A **Linear Layer** that maps the hidden layer representation to an output vector with a size equal to the vocabulary.
3) A **Log-Softmax** activation function on the output layer.

The models were trained for 50 epochs using the Adam optimiser with a learning rate of 0.025 and a batch size of 1024.

C. Learning the Linear Transformation

After training, we obtained two embedding matrices, one for Corpus A and one for Corpus B. To learn the mapping from space A to space B, we first identified the 4527 words shared between the two vocabularies. We then constructed two matrices, X_A and X_B, which contain embeddings for only these shared words.

The goal is to find a transformation matrix T such that $X_A T \approx X_B$. This is a linear least-squares problem, which was solved for T using the Moore-Penrose pseudo-inverse.

III. Results

The data preparation and model training resulted in two distinct embedding spaces. The key statistics for each corpus are summarised in Table I. The models were trained for 50 epochs, and while the loss decreased, as shown in Fig. 1, it remained high, indicating that the models had not fully converged on the complex data.

The primary quantitative result of our experiment is the average cosine similarity between the mapped vectors from space A and the original target vectors in space B. The final evaluation yielded:

- **Average Cosine Similarity: 0.2102**

IV. Discussion and Conclusion

The similarity result of 0.2102 is not an indication of experimental failure, but rather a meaningful insight into the challenges of aligning embedding spaces. We can say that the low score demonstrates a mapping is possible, but a simple linear transformation cannot fully reconcile the geometric structures of two spaces trained on texts with significant contextual and semantic shifts.

A primary reason for this is the unique narrative structure of the source text, *Harry Potter and the Goblet of Fire*. The book can be seen as two semantically different stories, with the turning point being Harry's selection by the Goblet of Fire. The first half of the book focuses on the setup of the Triwizard Tournament, introducing characters and establishing rules. The second half shifts dramatically to the tournament's dangerous tasks, the rise of Voldemort, and the story's climax. This creates a significant divergence in the contexts in which words appear. For example, words related to the different tournament stages or specific magical creatures would be unique to each half, making a 1:1 mapping impossible on those terms. Another fact is that the Goblet of Fire is the second longest Harry Potter book. So, while having a low cross-entropy loss 1 could occur for looking at only a two similar paragraphs and their similarities would be very close, it would be a misleading result with such small datasets. So,

yes - it is a common practice that requires a sufficient number of shared words, or anchor points, to effectively align and map embedding spaces trained on different data.

CONTRIBUTION STATEMENT

- Darshan Singh: data preparation, quantitative similarity results, domain knowledge on discussion and report writing, gpu optimisations/parallelisation for model training.
- Bwateketa Kabengele: refining initial data preparation, model architecture and implementation, and report writing.

V. REFERENCE

Mikolov, T., Chen, K., Corrado, G. and Dean, J., 2013. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

TABLE I
CORPUS STATISTICS

Metric	Corpus A (First Half)	Corpus B (Second Half)
Vocabulary Size	7,860	7,230
Training Pairs	214,610	212,238
Shared Words	4,527	

TABLE II
EXAMPLE WORD-LEVEL SIMILARITIES

Word	Mapped Cosine Similarity
'aback'	0.4152
'ability'	0.3319
'able'	0.2659
'abruptly'	0.3836
'absolutely'	0.3604
'accept'	0.3679

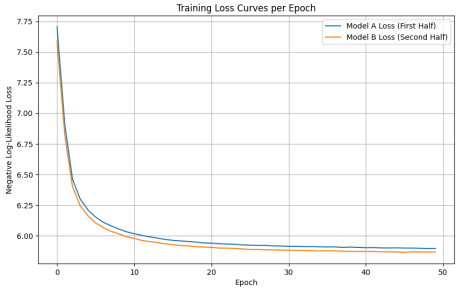

Fig. 1. Training Loss Curves for Model A and Model B.