Exploring Word2Vec Skip-gram Model: An Experimental Approach

Muhammed Muaaz Dawood (2425639)
Altaaf Ally (2424551)
Sayfullah Jumoorty (2430888)
Natural Language Processing
COMS4054A

I. Introduction

Word2Vec is a popular technique in natural language processing that generates vector representations of words, capturing semantic relationships between them. This lab involves implementing the Skip-gram model using a small dataset extracted from a text file.

II. Contribution of each team member

1) Data Preprocessing: Altaaf Ally
2) Context Extraction: Muhammed Muaaz Dawood
3) Model Training: Sayfullah Jumoorty
4) Experimentation: Muhammed Muaaz Dawood

III. Methodology

The methodology involves the following steps:

1) Data Preprocessing: Read and clean text data, extract unique words, and generate one-hot encodings.
2) Context Extraction: For each word, extract the context words within a specified window size.
3) Model Training: Train a neural network with a linear hidden layer to predict context words from a given word.
4) Experimentation: Investigated the effect of removing stop words.

IV. Results

A. Baseline Model

- **Number of words in text:** 1187
- **Vocabulary size:** 435
- **Number of training samples:** 4742

Sampled words: [’and’,’dursley’,’of’,’number’,’four’,’privet’, ’drive’,’were’,’proud’,’to’]
Sampled vocabulary: [’a’,’about’,’across’,’after’,’afternoon’, ’again’,’all’,’almost’,’also’,’although’]

B. Model with Stop Words Removed

- **Number of words in text:** 594
- **Vocabulary size:** 370
- **Number of training samples:** 2370

Sampled words: [’dursley’,’number’,’four’,’privet’,’drive’, ’proud’,’say’,’perfectly’,’normal’,’thank’]
Sampled vocabulary: [’across’,’afternoon’,’almost’,’also’, ’although’,’always’,’amount’,’angrily’,’another’,’anyone’]

TABLE I
Training Loss over Epochs for Baseline Model

Epoch	Loss
10	0.005784
20	0.005471
30	0.005200
40	0.004961
50	0.004745
60	0.004556
70	0.004383
80	0.004228
90	0.004089
100	0.003963

TABLE II
Training Loss over Epochs for Model with Stop Words Removed

Epoch	Loss
10	0.006372
20	0.006174
30	0.006000
40	0.005832
50	0.005681
60	0.005531
70	0.005399
80	0.005266
90	0.005145
100	0.005034

V. Discussion

The results demonstrate that removing stop words resulted in a reduced vocabulary size and fewer training samples, accompanied by a slight increase in training loss, as shown in Table II. Despite this increase, the model’s convergence was comparable to the baseline (Table I), indicating that while stop words provide some contextual structure, they are not essential for the model to learn effective word embeddings.

To assess the effectiveness of the learned embeddings, we employed a novel approach involving a logistic regression classifier trained on a synthetic dataset. The dataset consisted of sentences labeled as either ”magic” or ”muggle,” with examples such as *”harry casts a powerful spell”* for ”magic” and *”dursley family is non-magical”* for ”muggle.” Each sentence was represented by averaging the embeddings of its words, learned using the Word2Vec model. This method allowed us to directly evaluate how well the embeddings captured the semantic distinctions between the two categories.

The performance of the classifier, visualized in Figures 1 and 2, showed a marked improvement in accuracy from 0.4444 when stop words were included to 0.8889 when they were removed. This significant increase underscores the hypothesis that stop word removal enhances the semantic richness of the embeddings. By focusing the model on content-bearing words, we reduced the noise in the input data, which led to more precise embeddings and better classification performance.

Unlike traditional models where performance might be evaluated using metrics like the F1 score, our approach focused on how well the embeddings preserved semantic information in a straightforward binary classification task. The dramatic improvement in accuracy suggests that removing stop words allows the model to capture the underlying semantic relationships more effectively, which is critical in tasks where understanding the nuances of language is essential.

These findings suggest that while stop words play a role in maintaining the syntactic structure of sentences, their removal can significantly improve the quality of word embeddings in specific contexts. However, the impact of stop word removal might vary depending on the nature of the task and the model used. Future work could explore alternative methods, such as adjusting the weighting of stop words rather than removing them entirely, and further investigate the effects on different natural language processing tasks.

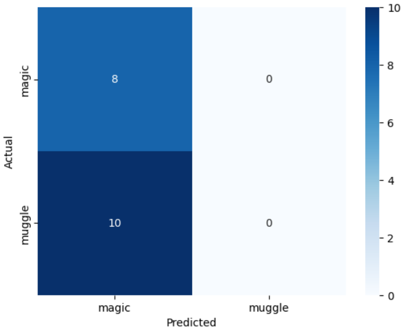

Fig. 1. Confusion Matrix with stop words

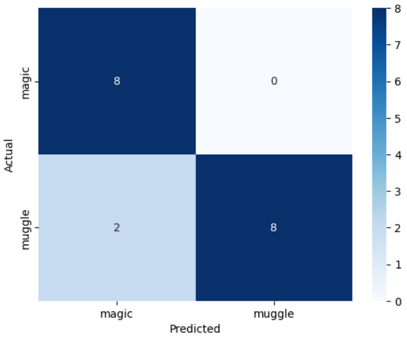

Fig. 2. Confusion Matrix with stop words removed

VI. Conclusion

In conclusion, removing stop words led to more effective word embeddings, as demonstrated by the improved classification performance in our synthetic dataset. This approach offers valuable insights into how preprocessing choices can influence the quality of word representations, with implications for a wide range of NLP applications.