Natural Language Processing - CS6120
Assignment 3 Report

PART II - ANALOGY DATASET

Introduction

We are using the following pre - trained word embeddings.
- GloVe [Pennington et al., 2014]: http://nlp.stanford.edu/projects/glove/
- Word2vec [Mikolov et al., 2013]: https://code.google.com/archive/p/word2vec/

The dataset used for testing is a *subset* of Mikolov;'s analogy dataset which includes 4 semantic relations. In the test files, each line represents one analogy question, in the form of four words < *a*, *b*, *c*, *d* >. We are working with the 8 groups as specified in the assignment.

1. ANALOGY TEST

Each model was assessed based on its ability to correctly identify the target word given three other words in the analogy format (a : b :: c : ?). The missing word is represented as "d" .

2. RESULTS

The following results were obtained:

	Glove	w2v
capital-world	0.84	0.80
currency	0.19	0.48
city-in-state	0.36	0.82
family	0.90	0.90
gram1-adjective-to-adverb	0.04	0.15
gram2-opposite	0.12	0.41
gram3-comparative	0.80	0.93
gram6-nationality-adjective	0.79	0.79

2.1. Analysis of Results

We can see from the table that word2vec model performed better in several categories and especially in 'currency' (0.48 vs 0.19) and "city-in-state" (0.82 vs 0.36) . This tells us word2vec has more understanding of relationships between words in these categories.

Both models excelled in "family" category achieving an accuracy of 0.90.

GloVe struggled with "currency" and "gram1-adjective-to-adverb", achieving accuracies of only 0.19 and 0.04. Word2Vec model showed some improvement in these areas but still exhibited lower performance in "gram1-adjective-to-adverb" (0.15)

While both models exhibit strengths and weaknesses across different categories, Word2Vec consistently outperformed GloVe in several key areas, like in analogy tasks involving semantic relationships.

3. ANTONYMS

One known problem with word embeddings is that antonyms (words with meanings considered to be opposites) often have similar embeddings. We verified this by searching for the top 10 most similar words to a few verbs like increase and enter that have clear antonyms (e.g., decrease and exit, respectively) using the cosine similarity.

To answer - why do embeddings have these tendency, we have some reasoning as below :

- Training Goal : We usually focus on training the model for predicting the similar words, rather than antonyms.
- Polysemy (Same word, multiple related meanings) Example: The term "cold" in a clinical setting. It could refer to a common viral infection (common cold). It might also describe the temperature sensation (the patient feels cold)
- Synonymy (Different words with similar meanings): Terms like "myocardial infarction," "heart attack," and "cardiac arrest."
- Training Data : text corpus used for training the embeddings can also impact the results.