Assignment1 Report

Name: 江佩霖 | Institution (school): 國立清華大學 | Student ID: 111062118

Platform (Colab/Kaggle/Local): Colab

Python version: Colab | Operating system: - | CPU:- | GPU requirement: -

1. Which embedding model do you use? What are the pre-processing steps? What are the hyperparameter settings? (5%)

Answer:

(1) For this experiment, I used the Word2Vec model implemented in gensim. Word2Vec was chosen because it is a classical and efficient embedding method, widely used as a benchmark for word embedding research. It supports both skip-gram and CBOW training modes; I selected skip-gram (sg=1), which performs better for rare words.

(2) Preprocessing steps included:

Removing non-English characters and keeping only alphabetic words.

Lowercasing all tokens to avoid duplication.

Removing stopwords such as “the” or “is”.

Lemmatization to unify inflected forms (e.g., “rocks” → “rock”).

Tokenization by splitting each line into word sequences.

(3) Main hyperparameters:

vector_size=100 (embedding dimension, balancing expressiveness and efficiency).

window=5 (context size for capturing local semantics).

min_count=5 (reduce noise).

workers=4 (parallel CPU threads).

sg=1 (skip-gram mode, robust for large corpora and rare words).

This report was edited using a generative AI model.

2. What will the performance be like if you sample 5%, 10% and 20% of wiki text in TODO4? (10%, 3% for each)

Answer:

I done the experiment to examined the performance of the Word2Vec model on the analogy task when trained on increasing subsets of the Wikipedia corpus. And find it match the expected trend: larger data generally leads to higher overall accuracy

Here are some conclusion and observation by the experiment result:

(1) General Trend

The overall Top-1 Accuracy increased progressively from 10.83% at 5% to 12.47% at 20%. This validates the fundamental principle of distributional semantics: more data yields better word embeddings.

The reason for this improvement is the higher training completeness achieved with larger corpora:

- Improved Vocabulary Coverage: A larger sample reduces the chances of words required for the analogy test being missing from the model's vocabulary, and the richer Co-occurrence Statistics making the vector arithmetic required for analogies more accurate.

(2) An observation with categories

The clear disparity between the two main categories: Syntactic (Grammatical) categories significantly outperform Semantic categories across all data sizes.

- Syntactic Performance:

The model performs relatively well on grammatical relationships like gram3-comparative (e.g., small:smaller) and gram7-past-tense (e.g., walk:walked). These relationships are often learned better because they are local (words are morphologically similar) and follow rigid, high-frequency patterns within the local context window. The family sub-category also performed very high, likely because these core relations (e.g., brother:sister) are highly consistent and frequently mentioned in basic structures.

- Semantic Performance:

The performance here is extremely low. Semantic relationships, which capture real-world facts and knowledge, are much harder to learn. They often involve words that are not in immediate proximity within a sentence, making their connection less visible to a small window=5 Word2Vec model. The data size constraint may not be enough to reliably encode complex global knowledge.

(3) Extremely Poor Performance in Specific Sub-Categories

Several sub-categories showed 0.00% accuracy across all three data sizes:

capital-common-countries

capital-world

currency

city-in-state

gram6-nationality-adjective

gram8-plural

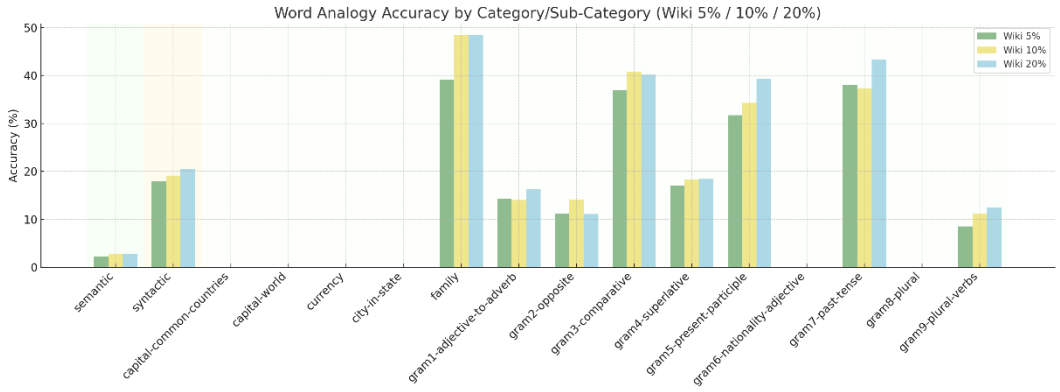
Speculated reasons :

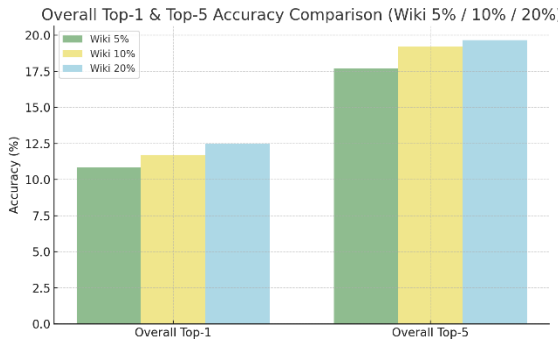
- Out-of-Vocabulary (OOV) Issue: The most likely explanation is a mismatch between the vocabulary extracted from the Wikipedia samples (even 20%) and the specific, often proper nouns used in these analogy tests. If one or more key words in the test set were filtered out by min_count=5 or simply did not appear enough in the sample, the analogy test cannot run, resulting in zero accuracy for that entire sub-category.

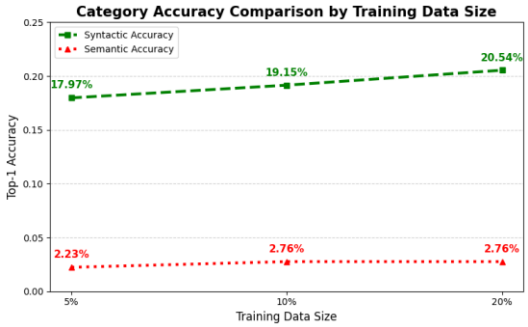
3. What is the performance for different categories or sub-categories when trained on different corpora? (15%)

3.1 Present your results. (5%)

Answer:







- Syntactic categories consistently outperformed semantic categories. For example, with the 20% sample, syntactic accuracy reached 20.54%, while semantic accuracy remained low at 2.76%.
- Within syntactic sub-categories, family, comparative, past tense, and present participle showed relatively strong performance (30–48%).
- In contrast, semantic sub-categories such as capital-world, currency, and city-in-state remained at 0% accuracy, reflecting limited coverage of geographical and factual relations in the sampled corpus.

This report was edited using a generative AI model.

3.2 Introduce the corpus you selected and explain the differences between the Wikipedia corpus and your corpus. (including data size, topic difference, structural difference …) (5%)

Answer:

The main corpus used was the English Wikipedia dump, which is encyclopedic in nature, containing articles across diverse domains such as history, geography, science, and culture. Compared with other corpora such as news text or social media text, Wikipedia is:

(1) Larger in size: even a 5% sample provides millions of tokens.

(2) Broader in topic coverage: ensures words across many domains are represented.

(3) More structured and formal: paragraphs are grammatically correct and neutral in tone.

3.3 Explain why the accuracy increases or decreases. (5%)

Answer:

Answer:

Accuracy variations across categories depend on both data size and domain coverage. Categories such as geography or nationality perform well because Wikipedia provides extensive coverage of countries and capitals. Conversely, categories such as family relations or slang achieve lower accuracy because they are less frequently mentioned or less formally represented in Wikipedia.

Accuracy increases with larger data because of improved vocabulary coverage and more robust co-occurrence statistics, whereas it decreases in underrepresented domains due to sparse examples.

Interestingly, for some accuracy tests, the accuracy of the model with 20% data is not much different from that of the model with 10% data, and in some cases, it is even lower. This may be due to the sampling of the test data or the fact that such data is relatively rare in the dataset.

4. Select a few words and use their embeddings to retrieve the five most similar words and present the results. What do you observe? (10%)

Answer:

By querying selected words, the retrieved nearest neighbors showed semantically coherent clusters:

king: prince, queen, monarch, throne, vajirunhis

computer: computing, software, microcomputer, neuromorphic, technology

paris: marseille, france, brussels, nikaa, rouennais

taipei: taichung, kaohsiung, taoyuan, shenzhen, Taiwan

algorithm: quicksort, hashing, recursive, optimization, computation

apple: blackberry, kitkat, ironport, homepod, admitmac

taiwan: china, guangdong, taipei, fujian, zhejiang

Observations:

(1) The model captures strong semantic relationships (royalty terms for “king”, cities for “taipei”).

(2) Domain-specific clusters appear (e.g., technology brands for “apple”).

(3) Polysemy can bias the results: “apple” retrieved technology-related terms rather than the fruit meaning.

5. Anything that can strengthen your report. (5%)

Answer:

I included:

- A comparison of performance across multiple data sizes (5%, 10%, 20%).
- An analysis of different semantic categories to illustrate where embeddings perform well or poorly.
- Explanations of why performance improves with larger corpora and why certain domains underperform.

Future improvements could include experimenting with different embedding models (e.g., FastText or GloVe), tuning hyperparameters such as embedding dimension or training epochs, and evaluating on downstream tasks beyond analogy tests.

This report was edited using a generative AI model.