DS-GA-1012: Natural Language Understanding and Computational Semantics, Spring 2025

Intrinsic Evaluation of Word Embeddings

Aditeya Baral
N19186654

Problem 1: Load Embeddings

Problem 1c: Extra Credit

The discrepancy occurs because of the way `__getitem__()` has been defined and implemented. The function's `words: Iterable[str]` argument expects an iterable of strings (type `str`) and then **extracts the corresponding vector using its index for each element in the iterable.** In both code snippets, the input provided is an iterable. In the first example, the argument `"the", "of"` is internally compiled by Python into an iterable `tuple` of the form `("the", "of")` and passed to the function, which processes each of the **tokens** independently to return a vector for each token. However, in the second example, only a single token `"the"` is passed. Since the function processes each element of the passed iterable and **strings in Python are also iterables with each character in the string being of type** `str`, the function instead processes each **character** of the input string. This leads it to individually process the characters `'t'`, `'h'` and `'e'` sequentially. This results in an error since the character `'h'` is not present in the saved embeddings file, and thus leads to a `KeyError` when we try looking up its index.

Problem 4: Interpretation of Results

Problem 4a: Syntactic vs Semantic Relation Types

Embedding Space	Semantic	Syntactic	Overall
CBOW 300	0.155	0.531	0.361
Skip-Gram 300	0.50	0.559	0.533
GloVe 50	0.400	0.276	0.332
GloVe 100	0.445	0.278	0.354
GloVe 200	0.317	0.217	0.262

Table 1: Analogy question accuracy for different GloVe embedding dimensions (`k=1`). Cells marked in green and red indicate the highest and lowest accuracies across each relation type respectively.

1. **Comparison with Word2Vec**
   - We observe that the **Skip-Gram model significantly outperforms both CBOW and GloVe models across all categories**. GloVe embeddings achieve better semantic accuracy than CBOW but fall short compared to Skip-Gram. For syntactic relations, both Skip-Gram and CBOW outperform GloVe embeddings by a large margin.
   - The GloVe 200d model achieved 26.2% overall accuracy, which is lower than both CBOW (36.1%) and Skip-Gram (53.3%) 300d models. It is also lower than the GloVe 50d and 100d models which achieved a 33.2% and 35.4% overall accuracy respectively.
   - **GloVe outperforms CBOW in semantic accuracy** across all dimensions but falls short compared to Skip-Gram, which achieves a significantly higher semantic accuracy of 50.0%.
     - The best semantic accuracy for GloVe is observed with 100-dimensional embeddings (44.48%), showing a clear advantage over CBOW.
   - **CBOW and Skip-Gram outperform GloVe significantly in syntactic accuracy**, with Skip-Gram achieving 55.9% and CBOW achieving 53.1%, compared to GloVe's best syntactic accuracy of 27.78% at 100 dimensions.
     - This is likely due to its **reliance on global co-occurrence** for representing words rather than local context windows, which Word2Vec captures effectively.
   - **Skip-Gram achieves the highest overall accuracy** (53.3%) followed by CBOW (36.1%).
     - GloVe's overall accuracy peaks at 35.36% with the 100-dimensional embeddings, slightly below CBOW but far from Skip-Gram's performance.
2. **Effect of Dimensionality**
   - **Increasing dimensionality from 50 to 100 dimensions improves semantic, syntactic and overall accuracy**, indicating that higher dimensions capture richer information about words and can represent intricate and nuanced relationships.
     - **Semantic tasks benefit more from increased dimensionality than syntactic tasks**, suggesting that higher dimensions help capture more nuanced semantic relationships.
   - However, **when dimensionality increases further to 200 dimensions, there is a significant drop in both semantic (31.67%) and syntactic (21.72%) accuracies**, as well as overall accuracy (26.24%). This suggests that **increasing the number of dimensions does not always improve the quality of learnt embeddings**. Additionally, overly high dimensions may lead to overfitting, inefficiencies in capturing meaningful relationships with larger embedding spaces and reduced generalizability in smaller datasets.

  - In contrast, **Word2Vec models (CBOW and Skip-Gram) perform better at higher dimensions** (300d), highlighting their ability to leverage larger embedding spaces effectively.
- Overall, **Skip-Gram remains the most robust model across both semantic and syntactic tasks**, while GloVe provides competitive performance in semantic tasks at lower computational cost compared to Word2Vec models like CBOW and Skip-Gram.

Problem 4b: Effect of Lenience

Embedding Space	Semantic	Syntactic	Overall
CBOW 300	0.155	0.531	0.361
Skip-Gram 300	0.50	0.559	0.533
GloVe 50	0.566	0.536	0.550
GloVe 100	0.665	0.659	0.662
GloVe 200	0.705	0.672	0.687

Table 2: Analogy question accuracy for different GloVe embedding dimensions (`k=2`). Cells marked in green and red indicate the highest and lowest accuracies across each relation type respectively.

1. **Comparison with Word2Vec**
   - We observe that **GloVe embeddings consistently outperform both Word2Vec models**, with GloVe (200d) achieving the best overall accuracy of 68.7%, compared to Skip-Gram's 53.3% when the lenience factor `k` is increased to 2.
   - **For semantic analogies, GloVe embeddings achieve significantly higher accuracy than both Word2Vec models**, demonstrating their strength in capturing semantic relationships. At 200 dimensions, GloVe achieves a semantic accuracy of 70.5%, compared to Skip-Gram's 50% and CBOW's much lower performance of 15.5%.
     - This suggests that **GloVe's global co-occurrence-based training method is particularly effective at capturing semantic relationships**.
     - The CBOW model lags significantly behind both Skip-Gram and GloVe in semantic tasks, likely because it averages context words during training, which may dilute meaningful semantic relationships.
   - **For syntactic analogies, the Skip-Gram model performs slightly better than GloVe embeddings at lower dimensions** (e.g., CBOW: 53.1% and Skip-Gram: 55.9%), but GloVe embeddings surpass them at higher dimensions (e.g., 200d).
     - This indicates that while GloVe is strong in semantic tasks, it also **performs well in syntactic tasks as dimensionality increases**.
     - The Skip-Gram model remains strong in syntactic analogies due to its focus on local context prediction, but GloVe embeddings surpass it when considering both semantic and syntactic tasks together.
   - **GloVe embeddings outperform both Word2Vec models in overall analogy question accuracy for all tested dimensions**. At 200 dimensions, GloVe achieves an overall accuracy of 68.7%, compared to Skip-Gram's best performance of 53.3%.
2. **Effect of Dimensionality**
   - **Increasing the dimensionality of GloVe embeddings improves accuracy across all categories** (semantic, syntactic, and overall).
     - Semantic accuracy improves significantly as dimensionality increases, from 56.6% (50d) to 70.5% (200d).
     - Syntactic accuracy also improves steadily with dimensionality, from 53.6% (50d) to 67.2% (200d).
     - Overall accuracy follows a similar trend, peaking at higher dimensions, improving from 55% (50d) to 68.7% (200d).

- This shows that an increased dimensionality suggests that larger embedding spaces help capture more nuanced relationships between words.

3. **Effect of Lenience** `k`.
   An increased lenience of `k=2` allows for more **flexibility in evaluating correct answers** and thus higher accuracy scores across all categories because the correct word does not need to be ranked first but can be within the top `k` closest words.

Problem 4c: Qualitative Evaluation

Analogy Question	Gold Answer	GloVe 50	GloVe 100	GloVe 200
france : paris :: italy : x	rome	rome	rome	rome
france : paris :: japan : x	tokyo	tokyo	tokyo	tokyo
france : paris :: florida : x	tallahassee	miami	florida	florida
big : bigger :: small : x	smaller	larger	larger	smaller
big : bigger :: cold : x	colder	cold	cold	cold
big : bigger :: quick : x	quicker	quick	quick	quick

Table 3: Qualitative results of analogy questions for different GloVe embedding dimensions. Words marked in gold indicate the gold, green indicate correct and red indicate incorrect answers respectively.

1. **Performance of GloVe Embeddings**
   - We observe that for semantic analogies like "france : paris :: italy : rome" and "france : paris :: japan : tokyo", all GloVe embeddings correctly predicted the gold answers.
     - However, for "france : paris :: florida : tallahassee", none of the GloVe embeddings returned the correct answer. Instead, GloVe50 predicted "miami", and GloVe100 and Glove200 predicted "florida".
     - This suggests that while GloVe embeddings are strong at capturing well-known capital-city relationships, they **struggle with less prominent associations and cannot capture less common semantic relationships**.
   - For syntactic analogies like "small", only the 200-dimensional embedding correctly predicted "smaller". The other dimensions predicted "larger", which is incorrect. For "cold" and "quick", all embeddings failed to predict the correct answers ("colder" and "quicker"), instead returning the base words ("cold" and "quick").
     - This indicates that GloVe embeddings **struggle with syntactic analogies, particularly in forming comparative relationships** and forms.
2. **Comparison with Mikolov et al. (2013)**
   - The Skip-Gram model from Mikolov et al. (2013) achieves strong performance on semantic analogies (e.g., italy:rome, japan:tokyo) due to its ability to capture global relationships effectively.
     - Although GloVe embeddings perform similarly well for prominent semantic relationships, they **fail for less common associations**. This suggests that while both models excel at capturing high-frequency relationships, **Skip-Gram may generalize better to low-frequency cases**.
   - Similarly, the Skip-Gram model demonstrates strong performance on syntactic analogies (e.g., big:bigger::small:smaller) while GloVe embeddings struggle with syntactic relationships across all dimensions.
     - This shows that **GloVe's global co-occurrence-based training method cannot capture fine-grained syntactic patterns** effectively.
3. **Effect of Dimensionality**
   - **Increasing dimensionality improves performance for some analogies** (eg., "small"), however, higher dimensions do not guarantee consistent improvement across all analogy types (eg., "cold" and "quick").
   - This suggests that while higher-dimensional embeddings capture richer information and more nuanced relationships, they may still **lack robustness in certain syntactic tasks**.