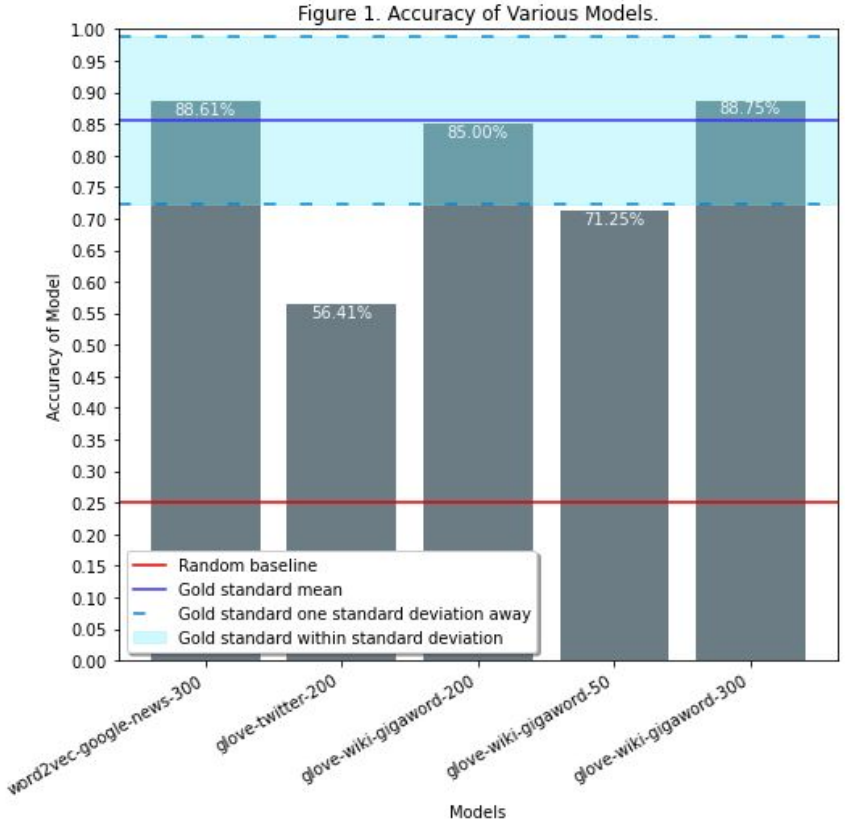
COMP 472 AI Mini-Project 3

AI Tech:
Kevin Rao 40095427
Lydia Fodouop 40132543
Suthan Sinnathurai 40086318
Due December 13th, 2021

Analysis Output

Model Name	Vocabulary Size	Correct Answers	NumOfValids	Accuracy
word2vec-google-news-300	3000000	70	79	0.886076
glove-twitter-200	1193514	44	78	0.564103
glove-wiki-gigaword-200	400000	68	80	0.85
glove-wiki-gigaword-50	400000	57	80	0.7125
glove-wiki-gigaword-300	400000	71	80	0.8875
Standards				
Human gold-standard	--	--	--	0.8573±0.1318
Random Baseline	--	--	--	0.25

Graph of Performance



Comparing models with same corpus but different embedding sizes

Comparison between glove-wiki-gigaword-50 and glove-wiki-gigaword-300:

- Both models made the same number guesses in the test set (zero).
- The accuracy of the model with the embedding size of 300 was higher than the lower embedding size model.

Model Name	Correct Answers	NumOfValids	Accuracy
glove-wiki-gigaword-50	57	80	0.7125
glove-wiki-gigaword-300	71	80	0.8875

Comparing models with different corpus but same embedding sizes

Accuracy difference between the models:

- glove-wiki-gigaword-200 vs glove-twitter-200:
  28.59%points.
- glove-wiki-gigaword-300 vs word2vec-google-news-300:
  0.14%points.

The corpus used for the model wildly affects its accuracy in our similarity tests. Some corpora have similar performance, while others have significant differences in performance.

Comparing all models with the human gold-standard and random baseline

Models with accuracy greater than the human gold-standard mean (85.57%) (but still within std.):

- word2vec-google-news-300 (88.61%)
- glove-wiki-gigaword-300 (85.73%)

Models with accuracy significantly lower than the human gold-standard mean (85.57%):

- glove-twitter-200 (56.41%)
- glove-wiki-gigaword-50 (71.25%)

Accuracy of the models ranked with the human gold standard and random baseline:

- glove-wiki-gigaword-300 > word2vec-google-news-300 > human gold standard > glove-wiki-gigaword-200 > glove-wiki-gigaword-50 > glove-twitter-200 > random baseline

Overall, the accuracy of the models is highly dependent on the pre-trained model's corpus as well as on its embedding size. Theoretically, the chosen test set could also have an effect on the accuracy of the models.

Ways to Improve the Models.

1. Have a larger embedding size.
    - As the analysis shows, models with larger embedding size perform better.
    - Eg. gigaword-50 to gigaword-300: accuracy from 71% to 89%.
    - Larger embedding size requires larger memory to perform, and much more processing to gather.

2. Selection of a corpus.
    - As the analysis shows, models with certain corpora perform better than others.
    - Eg. twitter-200 to gigaword-200: accuracy from 56% to 85%.
    - Though, there’s no way to pre-emptively know which corpus is best. Trial & error required.

Problems encountered while testing some models with the synonyms dataset

Problem #1 : When the question word is not part of the vocabulary of the pre-trained models.

Solution: Added an if condition that checks whether the model has an index for the question word. If it does, then we proceed as normal. Else, we perform a random guess out of the list of options.

Problem#3 : When none of the option words were part of the vocabulary of the model.

Solution: Randomly pick from all the original options.

Problem#2: When one word in the options are not part of the vocabulary of the pre-trained model.

Solution: Consider only the options that are part of the vocabulary of the pre-trained model.
If the question word is invalid, then the random choice will be among these filtered choices.

Team Contributions

Everyone participated.

Code is written together and the outputs were compared with each other.

Code were compared with, then merged together.

Slides were built together.