## 1 Question we decided to investigate

How different context vectors generation methods can help to reduce the effect caused by word frequency in word similarity task and correct/incorrect spelling task?

## 2 Preliminary Task

The words required to be tested were 'cat', 'dog', 'mouse', 'computer', '@justinbieber'. Table 1 shows the ranked list of those words and their corresponding cosine similarity.

Table 1: Ranked List

| Score | Word Pair | Occurrence | Occurrence |
|---|---|---|---|
| 0.36 | cat vs dog | 169733 | 287114 |
| 0.17 | comput vs mous | 160828 | 22265 |
| 0.12 | cat vs mous | 169733 | 22265 |
| 0.09 | mous vs dog | 22265 | 287114 |
| 0.07 | cat vs comput | 169733 | 160828 |
| 0.06 | comput vs dog | 160828 | 287114 |
| 0.02 | @justinbieber vs dog | 703307 | 287114 |
| 0.01 | cat vs @justinbieber | 169733 | 703307 |
| 0.01 | @justinbieber vs comput | 703307 | 160828 |
| 0.01 | @justinbieber vs mous | 703307 | 22265 |

## 3 Method

### Methods description

We used PPMI (Positive-Pointwise Mutual Information), AS-PPMI (Alpha Smoothed PPMI) in later and t-test to create context vectors and combined with four similarity methods, cosine similarity, Jaccard Index and Dice coefficient.

PPMI is the basic way to generate word vecters, and we used PPMI as our baseline. However, PPMI has some disadvantages, such as bias in low frequency words and undefined PPMI when one word never occurs. Therefore, as introduced by , we used AS-PPMI, to reduce this effect from word frequency. In AS-PPMI, we used the values 0.75, which is suggested from [1]. Moreover, we tried t-test as our third vector generation method, to test the difference of reducing word frequency effect between these two methods [1] . Here we defined t-test as: $t-test = \frac{P(x,y)-P(x)P(y)}{\sqrt{P(x)P(y)}} = \frac{c(x,y)-N*c(x)c(y))}{\sqrt{c(x)c(y)}}$.

For the similarity methods, we implemented three methods for comparison: Cosine similarity, Jaccard Index and Dice coefficient. Cosine similarity is pretty straight forward to calculate the similarity of two words by calculating the cosine values of angle between two context vectors. Jaccard Index method (we used Tanimoto-Jaccard here) calculates the intersection of two vectors over the union of the same two vectors, to represent the similarity of the pair words. Dice coefficient is similar to Jaccard Index, but Dice calculates twice of the intersection and divide it by the sum of the cardinalities of the vectors.

### Word pairs choosing idea and word pairs

Our first idea was to choose similar words from British English vs American English (BA), since it's easier to choose similar words and all pairs were used as a "golden standard" in evaluation. Our second idea was to choose some words with correct spelling and incorrect spelling (C/I), where incorrect spellings would normally have low frequency, and the pairs were also used as "golden standard". Details in Table 2.

## 4 Results

As we can see from **Table 3**, the AS-PPMI has the closest values to zero in all cases, which means in all similarity methods with AS-PPMI, we had the lowest correlation between similarity and word frequency. The

Table 2: Chosen Words

| Task | Category | WP1 | WP2 | WP3 | WP4 | WP5 | WP6 |
|---|---|---|---|---|---|---|---|
| **British English vs American English (BA)** | British English | subway | underground | underground | children | prawn | holiday |
| | American English | metro | subway | metro | kid | shrimp | vacation |
| **Correct Spelling vs Incorrect Spelling (C/I)** | Correct | calendar | committe | forty | separate | address | achieve |
| | Incorrect | calender | commitee | fourty | seperate | adress | acheive |

t-test reduced the correlation in cosine, but not in Jaccard or Dice. Therefore, we believed using AS-PPMI can best reduce the effect of word frequency to the similarity in the similarity methods we chose.

Table 3: Spearman's Rank Correlation Coefficient

| Method | British vs American | Correct vs Incorrect |
|---|---|---|
| Cosine (PPMI) | 0.4214 | 0.7698 |
| Cosine (AS-PPMI) | 0.079 | 0.5679 |
| Cosine (t-test) | 0.3726 | 0.6337 |
| Jaccard (PPMI) | 0.4145 | 0.8596 |
| Jaccard (AS-PPMI) | 0.1356 | 0.5814 |
| Jaccard (t-test) | 0.4824 | 0.7977 |
| Dice Coefficient(PPMI) | 0.4145 | 0.8596 |
| Dice Coefficient (AS-PPMI) | 0.1356 | 0.5814 |
| Dice Coefficient (t-test) | 0.4824 | 0.7977 |

We compared the top 6 ranking in similarity results with our two golden standards (6 pairs in each golden standard, so we chose top 6 ranking pairs). We defined the precision score as: if a pair in golden standard shows up in the top 6 ranking of our similarity results, regardless of the ranking order, we count once. The precision scores can be found in the Table 4.

When we looked at the precision score of **BA**, we could found that in this task, similarity methods with PPMI had the best scores, and other two context vector methods did not help to improve performance no matters word frequency was high or low. As we looked at the detail similarity results, we found many "incorrect" such as "subway&prawn" (due to the Subway sandwiches), "kid&vacation", which were not synonyms, and we guessed this could because these pairs were relative words and showing up together frequently in real life.

However in the task of **C/I**, all similarity methods with AS-PPMI had better precision scores than other methods. Generally, the incorrect spellings would have low frequency which would lead to a lower similarity results if we didn't do some frequency corrections (here AS-PPMI did frequency corrections). From both precision score and similarity results, we believed AS-PPMI helped to reduce the effect of low frequency and return better results in this task. Detail results can be found in the Section 6 (Appendices).

Table 4: The Number of Correct Answers in Top 6

| Task | Context Vector Method | Cosine Similarity | Jaccard Index | Dice Measure |
|---|---|---|---|---|
| **BA** | PPMI | 5 | 5 | 5 |
| | AS-PPMI | 4 | 4 | 4 |
| | T-Test | 4 | 5 | 5 |
| **C/I** | PPMI | 2 | 0 | 0 |
| | AS-PPMI | 6 | 5 | 5 |
| | T-Test | 6 | 2 | 2 |

## 5 Future direction

Sometimes high similarity score does not mean two words are similar, but can because they often occur together (such as subway &prawn adn kid&holiday) or have same function in a sentence (such as rice and bread). We can try some methods, for example, to create a word list of frequent co-occurrence words and use it to increase the accuracy in synonyms calculation. Also, we could try more word pairs to see the results. We have not discussed if a new word occurs in t-test (this time we defined t-test value as 0 if this situation happens), and we can try smoothing methods in t-test in future.

## References

[1] Daniel Jurafsky and James H Martin. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition, 2008.

## 6 Appendices

### 6.1 BA

| Sort by cosine similarity with PPMI | | | |
|---|---|---|---|
| **Score** | **Word Pair** | **Occurrence** | **Occurrence** |
| 0.43 | shrimp vs prawn | 16339 | 2582 |
| 0.29 | kid vs children | 646599 | 135443 |
| 0.28 | holiday vs vacat | 113007 | 69668 |
| 0.17 | subway vs shrimp | 31084 | 16339 |
| 0.16 | metro vs underground | 21907 | 14054 |
| 0.16 | metro vs subway | 21907 | 31084 |
| 0.15 | subway vs prawn | 31084 | 2582 |
| 0.10 | holiday vs children | 113007 | 135443 |
| 0.10 | metro vs vacat | 21907 | 69668 |
| 0.10 | metro vs holiday | 21907 | 113007 |

| Sort by cosine similarity with AS-PPMI | | | |
|---|---|---|---|
| **Score** | **Word Pair** | **Occurrence** | **Occurrence** |
| 0.44 | shrimp vs prawn | 16339 | 2582 |
| 0.10 | holiday vs vacat | 113007 | 69668 |
| 0.10 | kid vs children | 646599 | 135443 |
| 0.06 | subway vs prawn | 31084 | 2582 |
| 0.06 | subway vs shrimp | 31084 | 16339 |
| 0.05 | metro vs subway | 21907 | 31084 |
| 0.02 | metro vs underground | 21907 | 14054 |
| 0.01 | subway vs underground | 31084 | 14054 |
| 0.00 | kid vs vacat | 646599 | 69668 |
| 0.00 | kid vs prawn | 646599 | 2582 |

| Sort by cosine similarity with t-test | | | |
|---|---|---|---|
| **Score** | **Word Pair** | **Occurrence** | **Occurrence** |
| 0.40 | shrimp vs prawn | 16339 | 2582 |
| 0.25 | kid vs children | 646599 | 135443 |
| 0.23 | holiday vs vacat | 113007 | 69668 |
| 0.06 | subway vs prawn | 31084 | 2582 |
| 0.06 | kid vs holiday | 646599 | 113007 |
| 0.06 | metro vs underground | 21907 | 14054 |
| 0.05 | kid vs vacat | 646599 | 69668 |
| 0.05 | metro vs subway | 21907 | 31084 |
| 0.04 | subway vs shrimp | 31084 | 16339 |
| 0.04 | holiday vs children | 113007 | 135443 |

| Sort by Jaccard Index with PPMI | | | |
|---|---|---|---|
| **Score** | **Word Pair** | **Occurrence** | **Occurrence** |
| 0.20 | shrimp vs prawn | 16339 | 2582 |
| 0.18 | holiday vs vacat | 113007 | 69668 |
| 0.14 | kid vs children | 646599 | 135443 |
| 0.14 | metro vs underground | 21907 | 14054 |
| 0.11 | metro vs subway | 21907 | 31084 |
| 0.11 | subway vs shrimp | 31084 | 16339 |
| 0.08 | holiday vs children | 113007 | 135443 |
| 0.08 | metro vs vacat | 21907 | 69668 |
| 0.07 | metro vs holiday | 21907 | 113007 |
| 0.07 | subway vs prawn | 31084 | 2582 |

| Sort by Jaccard Index with AS-PPMI | | | |
|---|---|---|---|
| **Score** | **Word Pair** | **Occurrence** | **Occurrence** |
| 0.22 | shrimp vs prawn | 16339 | 2582 |
| 0.06 | holiday vs vacat | 113007 | 69668 |
| 0.04 | subway vs shrimp | 31084 | 16339 |
| 0.04 | subway vs prawn | 31084 | 2582 |
| 0.03 | metro vs subway | 21907 | 31084 |
| 0.03 | kid vs children | 646599 | 135443 |
| 0.01 | metro vs underground | 21907 | 14054 |
| 0.01 | subway vs underground | 31084 | 14054 |
| 0.00 | vacat vs prawn | 69668 | 2582 |
| 0.00 | kid vs prawn | 646599 | 2582 |

| Sort by Jaccard Index with t-test | | | |
|---|---|---|---|
| **Score** | **Word Pair** | **Occurrence** | **Occurrence** |
| 0.16 | holiday vs vacat | 113007 | 69668 |
| 0.15 | kid vs children | 646599 | 135443 |
| 0.15 | shrimp vs prawn | 16339 | 2582 |
| 0.08 | metro vs underground | 21907 | 14054 |
| 0.08 | metro vs subway | 21907 | 31084 |
| 0.06 | subway vs shrimp | 31084 | 16339 |
| 0.06 | holiday vs children | 113007 | 135443 |
| 0.06 | metro vs holiday | 21907 | 113007 |
| 0.05 | metro vs vacat | 21907 | 69668 |
| 0.05 | kid vs holiday | 646599 | 113007 |

| Sort by dice coefficient with PPMI | | | |
|---|---|---|---|
| **Score** | **Word Pair** | **Occurrence** | **Occurrence** |
| 0.33 | shrimp vs prawn | 16339 | 2582 |
| 0.30 | holiday vs vacat | 113007 | 69668 |
| 0.25 | kid vs children | 646599 | 135443 |
| 0.24 | metro vs underground | 21907 | 14054 |
| 0.20 | metro vs subway | 21907 | 31084 |
| 0.19 | subway vs shrimp | 31084 | 16339 |
| 0.15 | holiday vs children | 113007 | 135443 |
| 0.15 | metro vs vacat | 21907 | 69668 |
| 0.14 | metro vs holiday | 21907 | 113007 |
| 0.12 | subway vs prawn | 31084 | 2582 |

| Sort by dice coefficient with AS-PPMI | | | |
|---|---|---|---|
| **Score** | **Word Pair** | **Occurrence** | **Occurrence** |
| 0.36 | shrimp vs prawn | 16339 | 2582 |
| 0.10 | holiday vs vacat | 113007 | 69668 |
| 0.07 | subway vs shrimp | 31084 | 16339 |
| 0.07 | subway vs prawn | 31084 | 2582 |
| 0.06 | metro vs subway | 21907 | 31084 |
| 0.06 | kid vs children | 646599 | 135443 |
| 0.03 | metro vs underground | 21907 | 14054 |
| 0.02 | subway vs underground | 31084 | 14054 |
| 0.00 | vacat vs prawn | 69668 | 2582 |
| 0.00 | kid vs prawn | 646599 | 2582 |

| Sort by dice coefficient with t-test | | | |
|---|---|---|---|
| **Score** | **Word Pair** | **Occurrence** | **Occurrence** |
| 0.28 | holiday vs vacat | 113007 | 69668 |
| 0.27 | kid vs children | 646599 | 135443 |
| 0.26 | shrimp vs prawn | 16339 | 2582 |
| 0.15 | metro vs underground | 21907 | 14054 |
| 0.14 | metro vs subway | 21907 | 31084 |
| 0.12 | subway vs shrimp | 31084 | 16339 |
| 0.11 | holiday vs children | 113007 | 135443 |
| 0.10 | metro vs holiday | 21907 | 113007 |
| 0.10 | metro vs vacat | 21907 | 69668 |
| 0.10 | kid vs holiday | 646599 | 113007 |

## 6.2 C/I

| Sort by Cosine similarity with PPMI | | | |
|---|---|---|---|
| **Score** | **Word Pair** | **Occurrence** | **Occurrence** |
| 0.29 | committe vs address | 14689 | 55190 |
| 0.28 | seper vs separ | 4898 | 26431 |
| 0.22 | calendar vs calend | 23072 | 2672 |
| 0.18 | calendar vs address | 23072 | 55190 |
| 0.18 | separ vs address | 26431 | 55190 |
| 0.17 | committe vs separ | 14689 | 26431 |
| 0.17 | address vs adress | 55190 | 1995 |
| 0.16 | committe vs calendar | 14689 | 23072 |
| 0.15 | forti vs fourti | 3906 | 335 |
| 0.12 | separ vs calendar | 26431 | 23072 |

| Sort by Cosine similarity with AS-PPMI | | | |
|---|---|---|---|
| **Score** | **Word Pair** | **Occurrence** | **Occurrence** |
| 0.33 | seper vs separ | 4898 | 26431 |
| 0.27 | address vs adress | 55190 | 1995 |
| 0.25 | calendar vs calend | 23072 | 2672 |
| 0.17 | committe vs commite | 14689 | 236 |
| 0.17 | forti vs fourti | 3906 | 335 |
| 0.08 | achiev vs acheiv | 73716 | 607 |
| 0.04 | committe vs address | 14689 | 55190 |
| 0.03 | calendar vs address | 23072 | 55190 |
| 0.02 | seper vs calend | 4898 | 2672 |
| 0.02 | committe vs calendar | 14689 | 23072 |

| Sort by Cosine similarity with t-test | | | |
|---|---|---|---|
| **Score** | **Word Pair** | **Occurrence** | **Occurrence** |
| 0.56 | calendar vs calend | 23072 | 2672 |
| 0.45 | seper vs separ | 4898 | 26431 |
| 0.44 | address vs adress | 55190 | 1995 |
| 0.21 | committe vs commite | 14689 | 236 |
| 0.21 | forti vs fourti | 3906 | 335 |
| 0.10 | achiev vs acheiv | 73716 | 607 |
| 0.07 | committe vs address | 14689 | 55190 |
| 0.07 | calendar vs address | 23072 | 55190 |
| 0.04 | committe vs calendar | 14689 | 23072 |
| 0.04 | separ vs address | 26431 | 55190 |

| Sort by Jaccard Index with PPMI | | | |
|---|---|---|---|
| **Score** | **Word Pair** | **Occurrence** | **Occurrence** |
| 0.18 | committe vs address | 14689 | 55190 |
| 0.14 | separ vs address | 26431 | 55190 |
| 0.12 | calendar vs address | 23072 | 55190 |
| 0.12 | committe vs separ | 14689 | 26431 |
| 0.11 | committe vs calendar | 14689 | 23072 |
| 0.11 | separ vs calendar | 26431 | 23072 |
| 0.10 | seper vs separ | 4898 | 26431 |
| 0.09 | achiev vs separ | 73716 | 26431 |
| 0.09 | achiev vs address | 73716 | 55190 |
| 0.08 | calendar vs calend | 23072 | 2672 |

| Sort by Jaccard Index with AS-PPMI | | | |
|---|---|---|---|
| **Score** | **Word Pair** | **Occurrence** | **Occurrence** |
| 0.12 | seper vs separ | 4898 | 26431 |
| 0.10 | calendar vs calend | 23072 | 2672 |
| 0.08 | address vs adress | 55190 | 1995 |
| 0.05 | forti vs fourti | 3906 | 335 |
| 0.02 | committe vs address | 14689 | 55190 |
| 0.02 | committe vs commite | 14689 | 236 |
| 0.01 | calendar vs address | 23072 | 55190 |
| 0.01 | seper vs calend | 4898 | 2672 |
| 0.01 | achiev vs acheiv | 73716 | 607 |
| 0.01 | committe vs calendar | 14689 | 23072 |

| Sort by Jaccard Index with t-test | | | |
|---|---|---|---|
| **Score** | **Word Pair** | **Occurrence** | **Occurrence** |
| 0.12 | committe vs address | 14689 | 55190 |
| 0.10 | seper vs separ | 4898 | 26431 |
| 0.10 | calendar vs calend | 23072 | 2672 |
| 0.09 | separ vs address | 26431 | 55190 |
| 0.09 | calendar vs address | 23072 | 55190 |
| 0.07 | committe vs separ | 14689 | 26431 |
| 0.07 | separ vs calendar | 26431 | 23072 |
| 0.06 | committe vs calendar | 14689 | 23072 |
| 0.05 | address vs adress | 55190 | 1995 |
| 0.04 | achiev vs separ | 73716 | 26431 |

| Sort by Dice coefficient with PPMI | | | |
|---|---|---|---|
| **Score** | **Word Pair** | **Occurrence** | **Occurrence** |
| 0.30 | committe vs address | 14689 | 55190 |
| 0.25 | separ vs address | 26431 | 55190 |
| 0.22 | calendar vs address | 23072 | 55190 |
| 0.22 | committe vs separ | 14689 | 26431 |
| 0.20 | committe vs calendar | 14689 | 23072 |
| 0.19 | separ vs calendar | 26431 | 23072 |
| 0.18 | seper vs separ | 4898 | 26431 |
| 0.17 | achiev vs separ | 73716 | 26431 |
| 0.16 | achiev vs address | 73716 | 55190 |
| 0.15 | calendar vs calend | 23072 | 2672 |

| Sort by Dice coefficient with AS-PPMI | | | |
|---|---|---|---|
| **Score** | **Word Pair** | **Occurrence** | **Occurrence** |
| 0.21 | seper vs separ | 4898 | 26431 |
| 0.18 | calendar vs calend | 23072 | 2672 |
| 0.15 | address vs adress | 55190 | 1995 |
| 0.09 | forti vs fourti | 3906 | 335 |
| 0.04 | committe vs address | 14689 | 55190 |
| 0.04 | committe vs commite | 14689 | 236 |
| 0.03 | calendar vs address | 23072 | 55190 |
| 0.02 | seper vs calend | 4898 | 2672 |
| 0.02 | achiev vs acheiv | 73716 | 607 |
| 0.02 | committe vs calendar | 14689 | 23072 |

| Sort by Dice coefficient with t-test | | | |
|---|---|---|---|
| **Score** | **Word Pair** | **Occurrence** | **Occurrence** |
| 0.21 | committe vs address | 14689 | 55190 |
| 0.18 | seper vs separ | 4898 | 26431 |
| 0.18 | calendar vs calend | 23072 | 2672 |
| 0.16 | separ vs address | 26431 | 55190 |
| 0.16 | calendar vs address | 23072 | 55190 |
| 0.13 | committe vs separ | 14689 | 26431 |
| 0.12 | separ vs calendar | 26431 | 23072 |
| 0.12 | committe vs calendar | 14689 | 23072 |
| 0.09 | address vs adress | 55190 | 1995 |
| 0.09 | achiev vs separ | 73716 | 26431 |