A Projection-Based Asymmetric Similarity Measure for Distributional Semantic Models

Ria Das
riadas@mit.edu

Abstract

A popular and archetypal class of distributional semantic models is based on the theory that words can be represented as points in a high-dimensional semantic space. The classic definition of word similarity in these models is some function of the distance between two word vectors. However, such a similarity measure fails to account for many human word associations that are asymmetric or directed. For example, when presented with a word such as *wick*, subjects are quick to think of the word *candle*, but when presented with *candle*, they are more likely to think of words like *flame* and *wax* before *wick*, indicating that *wick* is more strongly associated to *candle* than *candle* is to *wick*. In this paper, we evaluate an asymmetric similarity measure based on word vector projections to improve the predictions of spatial semantic models of directed psychological association. We find that the projection-based metric generates predictions that match human association data to a statistically significant extent, though greater accuracy remains desirable.

1 Introduction

Spatial models of lexical semantics have endured significant criticism over the years, the most potent of which is directed at the fundamental tenets of this category of models. Specifically, representations of words as vectors in multidimensional space have traditionally employed similarity metrics based on the Euclidean distance between vectors [1]. Points that are closer together in space represent concepts that are more similar to each other. Consequently, the similarity relationships encoded by these geometric models must satisfy certain axioms, which were first noted by Tversky (1977): (1) The distance between a vector and itself must equal 0; (2) the distance between two vectors must be *symmetric*; and (3) the distances must satisfy the *triangle inequality* [2]. However, human similarity judgments have been shown to repeatedly violate these axioms, calling into question the validity of spatial approaches in modeling human psychological associations.

For example, when humans hear a word such as *funeral*, they are more likely to think of the word *wake* than they are to think of the word *funeral* after first hearing the word *wake*. The difference in the likelihood that subjects associate one word with the other depending on which word they are presented with first suggests that there is an inherent directionality in human similarity judgments. A symmetric distance-based similarity measure cannot account for these observations.

Many hypotheses have been proposed to try to explain why human similarity judgments are often asymmetric. These explanations include prototype theory, which asserts that some elements in group are more representative of the group than other elements. Another hypothesis is that the degree of generality of a word plays a role, and that words that are more general are associated less with words that are more specific [3]. Despite the large body of experimental evidence against their ability to accurately model human similarity judgments, however, spatial models remain very common. The idea that any human behavior, including free association, is the result of a cognitive process operating on some type of memory structure that is spatially related is very natural and remains popular [1].

In this paper, we experiment with a new approach to defining similarity in spatial semantic models, to amend the documented inability of such models to ac-

count for asymmetric free association data. We construct our new model of similarity based upon projections of one word vector to another, taking advantage of the inherent directionality in the definition of a projection vector. We apply our model to GloVe vector representations (trained on the Common Crawl 42B corpus) of pairs of words taken from two databases on free association norms, and compute the correlation between the human responses and our model's predictions.

The rest of the paper is structured as follows: In Section 2, we introduce our definition of word similarity in a spatial semantic model. In Section 3, we describe our experimental setup, detailing the contents of our reference datasets and how we processed them to generate our plots. In Section 4, we discuss the results of our experiments, and further compare with an earlier approach to integrating an asymmetric similarity measure with a spatial model.

2 Similarity Metric Definition

When thinking about how to geometrically represent a directed relationship between two vectors, the vector projection is a natural answer. If $\vec{v}_1$ and $\vec{v}_2$ are two vectors in an n-dimensional vector space, then $proj_{\vec{v}_1}\vec{v}_2$ represents the component of $\vec{v}_2$ that lies along $\vec{v}_1$, while $proj_{\vec{v}_2}\vec{v}_1$ represents the component of $\vec{v}_1$ that lies along $\vec{v}_2$. An immediate hypothesis about how to interpret this concept of projection in the context of vector representations of words is to let the projection of one word onto another word relate to how similar the first word is to the second word. For example, let the word "China" $= w_1$ and "Korea" $= w_2$, to use an oft-cited instance in the similarity literature. Then, $proj_{\vec{w}_1}\vec{w}_2$ would relate to the similarity of Korea to China, and $proj_{\vec{w}_2}\vec{w}_1$ would relate to the similarity of China to Korea.

We must be careful in reasoning about what types of comparisons between the two directed vector projections would correlate with the degree of asymmetry in the similarity relationship between two words. A first guess at a comparison metric would be to use the ratio between the norms of the two projections, i.e.,

$$\text{sym}(w_1, w_2) \sim \frac{proj_{\vec{w}_2}\vec{w}_1}{proj_{\vec{w}_1}\vec{w}_2},$$

where $\text{sym}(w_1, w_2)$ represents the symmetry of the similarity relationship between two words w_1 and w_2. If $\text{sym}(w_1, w_2) \approx 1$, then word w_1 is roughly as similar to word w_2 as word w_2 is to word w_1. If $\text{sym}(w_1, w_2) >> 1$ or $\text{sym}(w_1, w_2) << 1$, then the similarity relationship between the two words is asymmetric. At first glance, one might think that $\text{sym}(w_1, w_2) >> 1$ means that w_1 is more similar to w_2 than w_2 is to w_1, because it is natural to assume that the norm of the projection of $\vec{w}_1$ onto $\vec{w}_2$ is directly proportional to how similar w_1 is to w_2. However, we debunk this assumption by looking at some sample data points (word pairs with experimentally determined directional similarity) along with intuition about corpus-derived vectorizations of words.

We note that

$$\frac{proj_{\vec{w}_2}\vec{w}_1}{proj_{\vec{w}_1}\vec{w}_2} = \frac{|\vec{w}_1|\cos\theta}{|\vec{w}_2|\cos\theta} = \frac{|\vec{w}_1|}{|\vec{w}_2|},$$

so that our definition $\text{sym}(w_1, w_2)$ represents the ratio between the norms of the two word vectors. Since this definition does not capture information related to the angle between the two vectors, one might think that it could not possibly capture the similarity relationships between two word vectors. Indeed, between words that are very different, which, in terms of semantic space models, implies that the angle between the word vectors is large, then a ratio between the two vector norms likely does not signify much. However, for similar words (e.g. with angles less than $\pi/2$), it can be argued that the shorter vector is likely to be more similar to the longer vector than the longer vector to the shorter vector. The intuition behind this logic is that the longer GloVe vector is likely to be similar to more vectors than the shorter vector, because the greater sum of its squared coordinates indicates that it has stronger relations with other word vectors that are high-valued in other dimensions.

This intution is validated by several examples of word pairs that display asymmetric similarity relations, and in which the word with the shorter vector is more similar to the word with the longer vector. For example, to use the China-Korea example, experiments have shown that subjects tend to believe that Korea is more similar to China than China is to Korea, and the norms of the China and Korea GloVe vectors are 7.711 and 7.108, respectively. To give another example, the word "ace" is more similar to "bandage" than "bandage" is to "ace," and the GloVe vectors for "ace" and "bandage" have norms of 6.103 and 7.267. Thus, we decide to go forward with evaluating this similarity definition in our experiments.

Another similarity definition related to this discussion is the difference between the norms of the projections of the two vectors as opposed to the ratio, or

$$\text{sym}(w_1, w_2) \sim proj_{\vec{w}_2}\vec{w}_1 - proj_{\vec{w}_1}\vec{w}_2,$$

which can be written as

$$|\vec{w_1}| \cos\theta - \vec{w_2}| \cos\theta = (|\vec{w_1}| - |\vec{w_2}|) \cos\theta.$$

This definition is similar to the first in that it compares the norms of the two word vectors, but it also takes into account the angle between the two vectors, which intuitively should also be related to word similarity (though not asymmetry within a similarity relationship). We hence decide to go forward with testing this metric as well, so our two models are

1. $\text{sym}(w_1, w_2) \sim \frac{proj_{\vec{w}_2}\vec{w}_1}{proj_{\vec{w}_1}\vec{w}_2}$
2. $\text{sym}(w_1, w_2) \sim proj_{\vec{w}_2}\vec{w}_1 - proj_{\vec{w}_1}\vec{w}_2$.

In the first model, $\text{sym}(w_1, w_2) >> 1$ means that w_2 is more similar to w_1 than w_1 is to w_2, and in the second model, $\text{sim}(w_1, w_2) >> 0$ means that w_2 is more similar to w_1 than w_1 is to w_2.

The idea of using projections in computing the similarity between two words in a spatial semantic model is not entirely new. Pothos et. al. describe a model of word representation based on Quantum Probability theory, in which concepts are represented by different subspaces (with possibly multidimensional bases) and directional similarity relates to the projection from one subspace to another. However, Pothos et. al. describe this concept only theoretically within the framework of QP theory, without testing it empirically with existing datasets on word similarity [4]. The purpose of this paper is to empirically test these projection-based metrics against human data on the asymmetry of word similarity relationships, and we discuss our experimental setup in the following section.

3 Testing the Metric

We outline the methods by which we implemented our model and generated relevant data and plots, beginning with some more information about the reference data we consulted.

3.1 Reference Data

We evaluate the performance of our metric using two datasets containing the results of free association experiments. The first and smaller dataset comes from the association experiments run by McRae for words representing living and non-living concepts, and the second dataset is from the *University of South Florida Free Association Norms* database.

3.1.1 McRae Dataset

In the McRae association experiments, rougly 725 participants were given a set of 541 words representing living ("bird") and non-living ("chair") concepts, called cues, and were tasked with writing down the first three words related to each cue that came to their minds in order. The resulting dataset consists of 1169 rows of cue-response word pairs, along with the number of participants that wrote the response word to each cue first, second, and third, along with the total number of participants that wrote down that response. Two example rows from the McRae dataset containing the columns relevant to our analysis are shown below:

C	R	#R1	#UT	#WT
dog	bone	12	26	56
bone	dog	34	54	134

Figure 1: *Example from McRae data*

In this representation, C stands for Cue, R for Response, #R1 for number of participants who wrote down this response first, #UT for the unweighted total number of participants who wrote down this response as one of their three words, and #WT for the weighted total number of participants (where writing down first has a higher weight than writing down third).

3.1.2 USF Dataset

In the USF Word Association, Rhyme and Word Fragment Norms study, participants were given cue words and asked to write down a *single* response word that they associated with the cue. A total of 5,019 cue words were used, and each row in the resulting database consists of a cue word C, a response word R, and a number of metrics related to the number of participants who responded to each cue with the particular response. Four of these metrics that are of interest to us are #G, the number of participants who were presented with the particular cue word C; #P, the number of participants who responded to C with R; FSG, or *forward strength*, the ratio of #P to #G; and BSG, or *backwards strength*, the ratio of FSG to BSG from the row where the $(cue, response)$ pair is reversed. Two example rows from the USF database are given below:

C	R	#G	#P	FSG	BSG
aid	first	12	26	.291	.038
mussel	clam	34	54	.300	.014

Figure 2: *Example from USF data*

3.2 Methodology

We describe how we generate our experimental data using the McRae and USF datasets in conjunction with the pre-trained GloVe vectors. In both evaluation datasets, we are interested in the pairs of words (w_1, w_2) that appear in both the form $(w_1, w_2) = (cue, response)$ and $(w_1, w_2) = (response, cue)$. In the USF dataset, these rows are easy to identify, because they are the ones with both FSG > 0 and BSG > 0. In the McRae dataset, we iterate through each row in the dataset and explicitly check if the reversed word pair is also present in the dataset.

For each dataset, once we have collected all of the pairs that appear in both forward and reverse cue-pair relationships, we extract the Common Crawl GloVe vectors for each word represented among the pairs. We then compute the projections $proj_{w_1} w_2$ and $proj_{w_2} w_1$ for every pair of words (w_1, w_2) in our collection. Finally, we plot the human data (first response counts, unweighted total response counts, and weighted total response counts for the McRae dataset, and the response ratios for the USF dataset) against the ratios of the two projection norms (Section 4). We use the plotted data to additionally compute other measures to compare this projection model to previous attempts at modeling asymmetric similarity relationships (Section 5).

Specifically, we use the following forms of the human data in our plots:

1. McRae: Difference of #R1 of (w_1, w_2) to #R1 of (w_2, w_1)

2. McRae: Difference of #UT of (w_1, w_2) to #UT count of (w_2, w_1)

3. McRae: Difference of #WT of (w_1, w_2) to #WT of (w_2, w_1)

4. USF: Ratio of #FSG to #BSG for each (w_1, w_2)

5. USF: Difference between #FSG and #BSG for each (w_1, w_2)

These y-coordinates are plotted against both the ratio $|proj_{w_1} w_2|/|proj_{w_2} w_1|$ and the difference $|proj_{w_1} w_2| - |proj_{w_2} w_1|$ for each (w_1, w_2) in our collection, as discussed in Section 2.

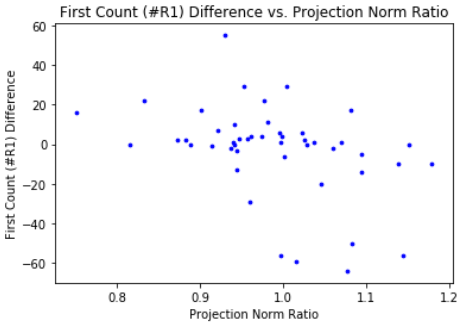
4 Results and Evaluation

We begin by discussing the preliminary plots of the McRae and USF association data against the ratios of and differences between the projection norms. These plots are shown in Figures 3 and 4.

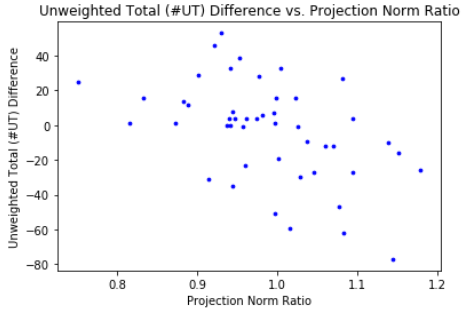
4.1 Preliminary Experiments

We first consider the McRae plots (Figure 3). After eliminating the cue-target pairs that did not also appear in reversed target-cue order, there remained 47 pairs of words in the dataset. Though the size of this sample is not large enough to draw far-reaching conclusions about the effectiveness of our projection measure in generally predicting psychological directed association, it is an interesting first test nonetheless. When we plotted the McRae experimental data against the ratios and differences of the computed projection norms for each of the 47 pairs, we found surprisingly significant Pearson and Spearman correlation coefficients: The average of the Pearson correlation coefficients across all six plots (#R1 vs. norm ratio, #R1 vs. norm difference, #UT vs. norm ratio, #UT vs. norm difference, #WT vs. norm ratio, #WT vs. norm difference) was -0.444, and the average of the Spearman correlation coefficients was -0.45. The negative sign in the correlation coefficients is consistent with our predictions, because we expect projection norm ratios greater than 1 and norm differences greater than 0 to correspond with target words being more strongly associated to cue words than cue words to target words. This is represented by experimental data ratios less than 1 and experimental data differences less than 0 (according to the direction of these ratios and differences in our definition). In addition, the plots with the projection norm ratios on the x-axis produced slightly better average Pearson and Spearman correlation coefficients (-0.449 and -0.474) than the plots with the projection norm differences on the x-axis (-0.439 and -0.443), though not by a significant amount.

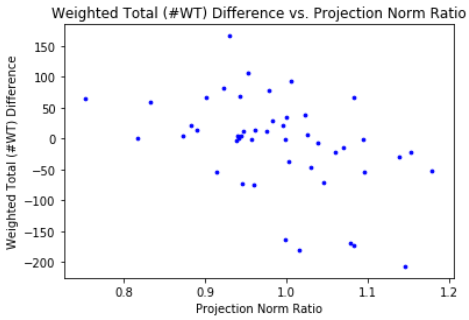
In addition to Pearson and Spearman coefficients, we also calculated the accuracy of the model for each plot, or the proportion of word pairs for which it correctly predicted the direction of association (without regard to magnitude). These values were all high with an average of 0.723, indicating that the model classified the directions of association in the McRae data quite well.



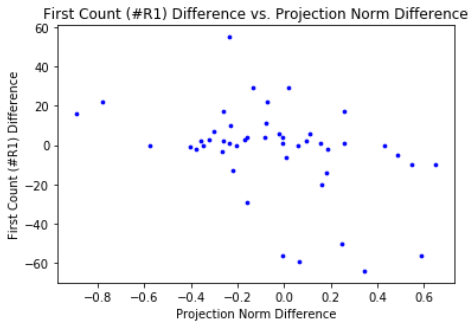
(a) $r = -0.409$, $\rho = -0.434$, $acc. = 0.681$



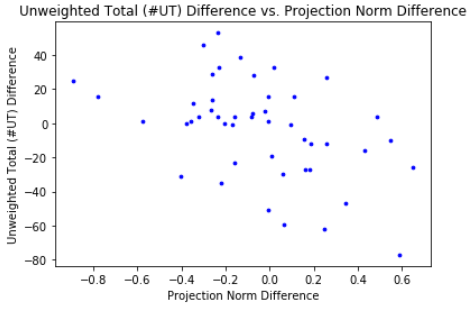
(b) $r = -0.487$, $\rho = -0.500$, $acc. = 0.766$



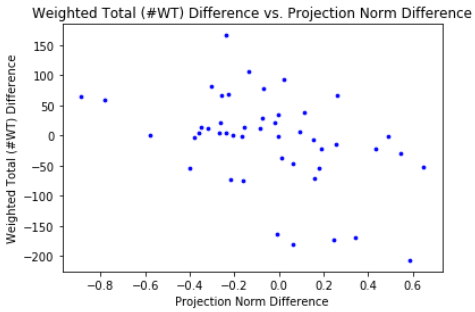
(c) $r = -0.452$, $\rho = -0.489$, $acc. = 0.723$



(d) $r = -0.416$, $\rho = -0.393$, $acc. = 0.681$



(e) $r = -0.460$, $\rho = -0.447$, $acc. = 0.766$

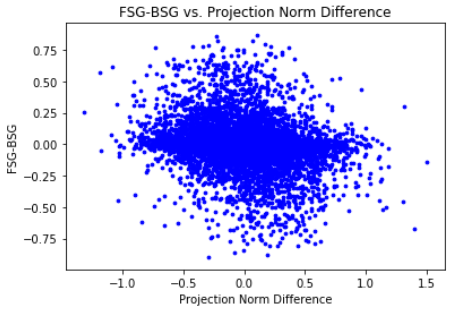


(f) $r = -0.442$, $\rho = -0.437$, $acc. = 0.723$

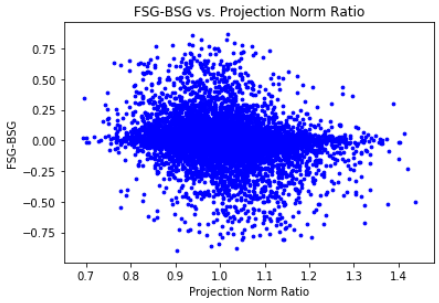
Figure 3: Plots of McRae free association data against GloVe vector projection norm ratios and differences. r represents the Pearson correlation, ρ represents the Spearman correlation, and *acc.* represents the proportion of word pairs for which the metric correctly predicted the direction of association.

We next consider the USF plots (Figure 4). The number of distinct pairs of words with both positive FSG and positive BSG is 8316. When plotting the USF experimental data against the ratio and difference in the projection norms, we decided to apply a log transformation to the FSG/BSG ratios on the y-axis to make any monotonic relations easier to see. The Pearson and Spearman correlation coefficients we obtained from these plots are lower than the those obtained from the McRae data, with an average of $r = -0.210$ and $\rho = -0.211$. The average accuracy of predicting the direction of the asymmetry relation for each pair of words is 0.574, less than the accuracy for the McRae data. This is not unexpected given the greater size of the USF dataset, which makes its correlation coefficients and accuracy less affected by random bias (of both the words selected as cues and from the fact that different populations of respondents produce different responses to cues). The fact that the correlation coefficients are still negative and are of a not insignificant magnitude indicate that the projection measure does capture some aspect of the asymmetric similarity relationships between the words. We build off of this observation of a weaker correlation given a larger dataset in the following section, by looking specifically at the word

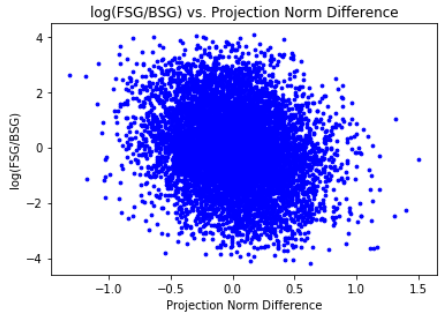
pairs in the USF database that have clearly asymmetric FSG/BSG values as opposed to the entire dataset at once.



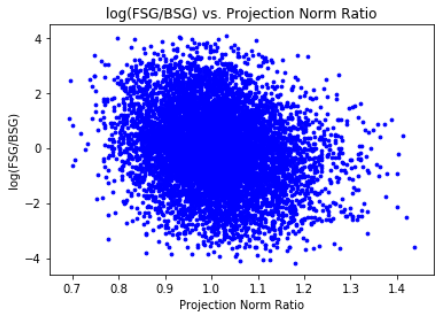
(a) $r = -0.195, \rho = -0.208, acc. = 0.574$



(b) $r = -0.219, \rho = -0.214, acc. = 0.574$



(c) $r = -0.202, \rho = -0.209, acc. = 0.573$



(d) $r = -0.222, \rho = -0.214, acc. = 0.573$

Figure 4: Plots of USF free association data against GloVe vector projection norm ratios and differences.

4.2 Further Analysis of USF Data

We are interested in knowing whether the lower asymmetry direction accuracies in the USF plots are caused by the measure finding it difficult to classify points that are roughly symmetric or are asymmetric. If the lower accuracy is caused by misclassification of roughly symmetric points, this does not indicate a more systematic problem with the model, because it is difficult to determine which word in a pair is more strongly associated to the other when both words have similar association statistics. If the measure definition misclassifies the direction of the more asymmetric pairs as often as it misclassifies the direction of the more symmetric pairs, then it is more likely that there exists an underlying problem with the model. We generate statistics from which this question by calculating the direction accuracy on the subsets of word pairs with $|FSG - BSG|$ above certain thresholds, shown in Figure 5. We note that the accuracy increases steadily as symmetric points are removed, indicating that the model predicts the asymmetry direction of more extreme points well, as desired.

$\lvert FSG - BSG \rvert$	#Points	Direction Accuracy
0	8316	0.574
0.1	2704	0.634
0.25	1092	0.671
0.5	279	0.717

Figure 5: *Measure Accuracy for Asymmetric Points*

4.3 Comparison with Previous Model

We finish our analysis by comparing our results with those achieved by Michelbacher et. al., who developed two competing approaches to measuring asymmetry in similarity relationships. Their first measure was to use conditional probabilities, defining the association of word w_1 to word w_2 as the joint probability of w_1 and w_2 divided by the probability of w_2. Their second measure was a rank measure based on the χ^2 statistical test. One of their statistics comparing the effectiveness of the two models considered the top ten most asymmetric word pairs in the USF

w_1	w_2	$FSG - BSG$	FSG	BSG	$\log(\frac{proj_{\vec{w}_2}\vec{w}_1}{proj_{\vec{w}_1}\vec{w}_2})$	$proj_{\vec{w}_2}\vec{w}_1 - proj_{\vec{w}_1}\vec{w}_2$
cow	moo	0.899	0.061	0.96	-0.100	-0.297
fish	trout	0.877	0.036	0.913	0.037	0.191
cheddar	cheese	0.867	0.922	0.055	0.017	0.102
cry	weep	0.859	0.058	0.917	-0.020	-0.080
halt	stop	0.856	0.906	0.050	-0.063	-0.232
cut	scissors	0.845	0.034	0.879	-0.003	-0.010
assist	help	0.826	0.842	0.016	-0.042	-0.217
exhausted	tired	0.820	0.895	0.075	0.015	0.071
late	tardy	0.811	0.088	0.899	0.134	0.271
baby	crib	0.810	0.032	0.842	0.047	0.207

Figure 6: Measures calculated for ten most asymmetric word pairs in USF dataset.

database, and looked at how the two calculated measures corresponded to the human data. This experiment performed with our projection-based measure is shown in Figure 6. We find that our projection-based measure predicted the correct direction of asymmetry in just five out of the ten example pairs, while the two measures by Michelbacher et. al. predicted the correct direction for ten and seven out of the example pairs. Though this negative result may be a result of the small sample size of just ten word pairs (since the average model accuracy for the top 50% of pairs was calculated to be 72% in Section 4.2), this result highlights an area of refinement for our model that we will focus on in future extensions to this project.

5 Conclusion and Future Work

We introduced and evaluated two measures of asymmetry in word similarity relationships in semantic space models, both based on vector projections. We computed these measures for the GloVe representations of pairs of words used in two free association experiments, that conducted by McRae et. al. and that used to create the *USF Free Association Norms* database.

We found that our computed asymmetry scores correctly predicted the direction of asymmetry for 72% of the relevant word pairs from the McRae dataset, and 57% of the relevant word pairs from the USF dataset. We further found that the prediction accuracy of our approach on the USF data steadily improved as we removed the more symmetric (and hence more directionally ambiguous) pairs: With the bottom 50% of word pairs based on degree of asymmetry removed from the dataset, we found that the prediction accuracy of our model jumped to 72%, indicating that our model is better at classifying asymmetric pairs than symmetric pairs. When looking at the top ten most asymmetric pairs in the USF data, however, we found that our model performed worse than that of a competing model, despite the strong average performance.

Future improvements to this project include performing the same statistical experiments on GloVe vectors trained on a larger corpus, such as the Common Crawl 840B-token corpus, as opposed to the Common Crawl 42B-token corpus. In addition, as indicated by the result in Section 4.3, we would like to refine our constructed model so that we obtain improved performance in predicting the directions of asymmetric pairs. To do this, we should read more of the existing literature on asymmetric similarity modeling in spatial semantic models, and observe what types of decisions about the construction of a measure lead to more accurate predictions. A useful starting point may be Pothos's Quantum Probability model, because the framework described in their paper provides several theoretical details that we did not explore in this first proof-of-concept paper.

I really enjoyed the first-hand experience of working with the idea that observable human behavior can be modeled as arising from cognitive processes operating on a spatial representation of memory structure. It was exciting to see a model as simple as projections of word vectors achieve a degree of success in capturing human psychological judgments. It makes me motivated to continue thinking about this space of problems related to how words/concepts and the relationships between them are represented in the human mind. I am eager to continue working on the model developed in this project, in particular by implementing and further building off of the quantum probability concepts described in the Pothos paper. I wonder if a marriage between the spatial models of lexical semantics and probabilistic models of cognition is possible, and whether such an integration would produce a more accurate model of human memorial representation.

6 Acknowledgements

I would like to thank Idan Blank for providing guidance on this paper as a "Project TA," as well as thank the 6.804 TAs for giving us the opportunity to work with BCS graduate students and postdocs on our final project. It greatly enriched the experience of designing and carrying out the experiments in and writing this paper.

References

[1] Jones, Michael B.; Gruenenfelder, Thomas M.; Recchia, Gabriel (2017). *In Defense of Spatial Models of Lexical Semantics*. New Ideas in Psychology. 50. 10.1016/j.newideapsych.2017.08.001.

[2] Johannesson, Mikael. *Modelling asymmetric similarity with prominence* (2000). British Journal of Mathematical and Statistical Psychology 53: 121–39.

[3] Michelbacher, Lukas; Evert, Stefan; Schütze, Hinrich (2007). *Asymmetric association measures*. Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP), Borovets, Bulgaria.

[4] Pothos, E. M., Busemeyer, J. R., and Trueblood, J. S. (2013). *A quantum geometric model of similarity*. Psychological Review, 120, 679–696.