Querying Word Embeddings for Similarity and Relatedness

Fatemeh Torabi Asr
Simon Fraser University
Buranby, BC, Canada
ftorabia@sfu.ca

Robert Zinkov
University of Oxford
Oxford, UK
zinkov@robots.ox.ac.uk

Michael N. Jones
Indiana University
Bloomington, IN, USA
jonesmn@indiana.edu

Abstract

Word embeddings obtained from neural network models such as Word2Vec Skipgram have become popular representations of word meaning and have been evaluated on a variety of word similarity and relatedness norming data. Skipgram generates a set of word and context embeddings, the latter typically discarded after training. We demonstrate the usefulness of context embeddings in predicting asymmetric association between words from a recently published dataset of production norms (Jouravlev and McRae, 2016). Our findings suggest that humans respond with words closer to the cue within the context embedding space (rather than the word embedding space), when asked to generate thematically related words.

1 Introduction

Modern distributional semantic models such as Word2Vec (Mikolov et al., 2013a,b) and GloVe (Pennington et al., 2014) have been evaluated on a variety of word similarity and relatedness datasets. A considerable amount of attention has been paid to what models and, more recently, what parameter settings and input data produce embedding representations that better reflect similarity/relatedness between words, taking human normative judgments as the gold standard (Baroni et al., 2014; Kiela et al., 2015; Levy et al., 2015; Melamud et al., 2016; Sahlgren and Lenci, 2016).

Similarity between two words is often assumed to be a direction-less measure (e.g., *car* and *truck* are similar due to feature overlap), whereas relatedness is inherently directional (e.g., *broom* and *floor* share a functional relationship). In addition, it is well established in human behavioral data that similarity and relatedness judgments are both asymmetric. For example, humans judge *leopard* to be much more similar to *tiger* than *tiger* is to *leopard* (Tversky and Gati, 1982). A concordant asymmetry is seen in relation tasks: in free association data, *baby* is a much more likely response when cued with *stork* than *stork* would be as a response when cued with *baby* (Nelson et al., 1999). The distinction between similarity and relatedness, and the asymmetry of the judgments have typically been ignored in recent evaluations of popular embedding models.

There is ample experimental evidence in the psycholinguistic literature that similarity and relatedness are both well represented in human behavior (see Hutchison (2003), for a review), and are qualitatively distinct representations or processes. In semantic priming paradigms, a target word is processed more efficiently when briefly preceded by a related or similar word (e.g., *honey-bee* or *wasp-bee*) relative to a neutral or unrelated prime (e.g., chair-bee). Facilitation is seen for word pairs that are purely category coordinates (*lawyer-surgeon*) or purely associates (*scalpel-surgeon*), and pairs that share both types of relations (*nurse-surgeon*) tend to see an additive processing benefit that reflects the privilege of both similarity and relatedness, an effect generally referred to as the "associative boost" (Chiarello et al., 1990; Lucas, 2000). Asymmetries are the norm in semantic priming data, leading to the early theoretical prominence of spreading activation models to account for human data.

Free association data provide complimentary evidence of the qualitative distinction between relatedness and similarity in human memory. In a free association task, participants are provided with a cue word and are asked to rapidly respond with a word that comes to mind first. Huge norms of human responses have been collected over the years; for example, (Nelson et al., 1999) early norms contain three-quarters of a million responses to over 5,000 cue words across 6,000 par-

Proceedings of NAACL-HLT 2018, pages 675–684
New Orleans, Louisiana, June 1 - 6, 2018.

ticipants. More recently, (De Deyne et al., 2016) have more than doubled the size of Nelsons norms in multiple languages by gamifying the task[1] . The majority of responses in free association data are based on thematic relatedness rather than similarity per se (De Deyne and Storms, 2008). As with semantic priming, free association norms are dominated by asymmetric relations: While *stork* has a very high probability of eliciting *baby* as a response across participants, cuing with *baby* brings so many competitors to mind that it is extremely unlikely to respond with *stork* (Hutchison, 2003).

The difficulty of accounting for similarity and relatedness with a single vector representation for each word has led to the suggestion that distinct representations, and perhaps even distinct learning models, are needed for optimal performance on these distinct tasks (Mandera et al., 2017). It may be unrealistic to expect a single vector representation to account for qualitatively distinct similarity and relatedness data. Further, asymmetries in human similarity and relatedness tasks have been used as strong evidence against spatial models of semantics such as word embedding models, and in favor of Bayesian models (Griffiths et al., 2007); but see (Jones et al., 2017). The cosine between two word vectors is inherently symmetric: *leopard-tiger* has the same cosine as *tiger-leopard*.

In order to understand how distributional representation of words reflect similarity and relatedness one should study the algorithms. Each cell of a word vector in a count model indicates the first-order association between the target word and a context word, document, or topic. Dimensionality reduction algorithms are applied to obtain denser representations that can demonstrate second-order relatedness/similarity between words (e.g. applying SVD to PMI matrix). Relative to these classic models, predictive distributional models such as Word2Vec are generally more complicated. Decomposition and interpretation of the neural word embeddings is less straightforward because the final vectors incrementally converge from a predict-and-update process based on a local objective function rather than by global counting or a batch abstraction process. Most evaluative studies of predictive distributional semantics have viewed these models as a black box, considering only at the output vectors. For example, the Word2Vec Skipgram architecture has easily taken the lead and become representative of the predictive distributional semantic models, but little attention has been paid to what statistical information is best represented in the two resulting embedding sets. The Skipgram is a feed-forward network with localist input and output layers, and one hidden layer which determines the dimensionality of the final vectors. It is trained on word-context pairs with an objective function trying to minimize the error of predicting context words within a specific window around the center word. At the end of training, two matrices are produced, one representing **word embeddings** and the other representing **context embeddings** for each and every vocabulary word. While word embeddings have been used as the output of Skipgram in many previous studies, little attention has been paid to the context embeddings and the usefulness of these vectors in performing lexical semantic tasks (Levy et al., 2015; Melamud et al., 2015; Aoki et al., 2017).

Recently, Asr and Jones (2017) used an artificial language to evaluate how hyperparameter settings affected the Skipgrams representation of first- vs. second-order statistical sources. In natural languages, paradigmatic and syntagmatic information sources are non-independent, confounding similarity and relatedness judgments. Words that are more similar tend to also share functional, script, or thematic relations (Hutchison, 2003; Lucas, 2000); e.g., *surgeon-nurse*. Asr and Jones artificial language was engineered to disentangle the two sources of statistical information. Following on suggestions by Levy et al. (2015), Asr and Jones found that averaging context vectors with the word vectors (w+c post-processing) produced optimal organization of the semantic space for both paradigmatic and syntagmatic structure. The goal of the current work is to more systematically explore the integration of word and context vectors in similarity and relatedness data; our two core objectives are:

1. To evaluate the Skipgram model on thematic relatedness production norms, which implicitly manifests asymmetric relations between words compared to the typical evaluation on direction-less similarity/relatedness.
2. To explore novel ways of computing relatedness scores by contributing both word and context embeddings produced by Word2Vecs Skipgram architecture.

[1] https://smallworldofwords.org

2 Similarity vs. Relatedness Data

One of the famous datasets on word similarity/relatedness is Wordsim353 (Finkelstein et al., 2001) including 353 English word pairs, and a revised version (Agirre et al., 2009) splitting similar from related word pairs (WrordSim). These data have been repeatedly used in comparative studies on distributional semantic models. Recently, the division between similarity and relatedness judgments has been highlighted in the literature, resulting in development of new datasets with more specific annotation instructions.

Hill et al. (2015) introduced the SimLex-999 dataset (SimLex) for purified evaluation of word similarity by asking the annotators explicitly not to score based on degree of relatedness. For example, the word pair coast-shore received an average similarity score of 9.00 in SimLex and 9.10 in WordSim, while the related word pair clothes-closet was assigned an average score of 1.96 in SimLex and 8.00 in WordSim. More recently, Jouravlev and McRae (2016) collected pure relatedness data through a production experiment. They presented participants with cue words and instructed them to response only with directly related words and not taxonomically similar words. This database (ProNorm) includes responses to 100 object words, providing us with directional relatedness score for 1,169 word pairs.

The important distinction of SimLex and ProNorm datasets compared to other available similarity/relatedness data is the explicit instruction of participants to pay attention to one aspect of word relations and not the other. The ProNorm dataset, also has an advantage of a more natural setup, where associatively related words were generated by participants, rather than being selected by language experts and only rated by the participants. In this paper, we use ProNorm as the main dataset to investigate how word embedings should be used to measure relatedness between two words and how the free recall experiment can be simulated for the model. The SimLex dataset is used to set a baseline for comparison against the similarity measurement task, which is the most common intrinsic benchmark for evaluation of word embeddings. Finally, we use the WordSim dataset to explore whether the observed differences between vector-based measures of similarity and relatedness come out if the benchmark data is collected in implicit setup, where participants did not know they were rating for similarity or relatedness.

3 Word and Context Embeddings

Word embeddings produced by the Skipgram architecture have been used in many previous studies as the word meaning representation and are the main output of the model. In the original implementation of Word2Vec, the context embeddings (weights on the hidden to output layer of the neural network) were discarded after learning was complete. Inspired by Pennington et al. (2014) in the architecture of the GloVe model, Levy et al. (2015) proposed that the final word embeddings in Word2Vec could be obtained from the average of word and context embeddings. They implemented word + context (w+c) as a useful post processing option for the Word2Vec Skipgram algorithm in their published version of the model[2]. The w+c option allows computation of word similarity based upon both first and second-order co-occurrence information. The cosine similarity between two words based on the dot product of their w+c embeddings, which we call the AA measure (*A* standing for the average of word and context embeddings of a word), includes the following terms:

$$cos(a,b) = \frac{W_aW_b + C_aC_b + W_aC_b + C_aW_a}{2\sqrt{W_aC_a + 1}\sqrt{W_bC_b + 1}} \quad (1)$$

While traditional measures, i.e., WW (cosine similarity of the word embeddings), and AA (cosine similarity of the word+context embeddings) are suitable predictors for words similarity, we hypothesize that the asymmetric measures WC (word embedding of the first word and context embedding of the second) and CW (context embedding of the first word and word embedding of the second) should be better indicators of relatedness. This decomposition of similarity measures is especially useful when asymmetric associations between words are being inferred: the asymmetric measures reserve the direction and the type of relation: WC reflects the likelihood of the second word occurring in the context of the first word, and CW reflects the likelihood of the first word occurring in the context of the second word. These two quantities are different, given that the W and C matrices are obtained from two different layers

[2] https://bitbucket.org/omerlevy/hyperwords

of the neural network, one connected to the input layer and the other to the output layer.

4 Experiments

SimLex and ProNorm provide complementary scores on similarity and relatedness between words. In order to demonstrate and examine how word embeddings should be used in asymmetric relatedness measurement, we designed two experiments. In both experiments word and context embeddings were obtained from Skipgram models trained on a tokenized English Wikipedia dump[3]. We slightly modified the original Word2Vec Skipgram implementation by Levy et al. (2015) to save both word and context vectors.

We tested vector spaces with varying dimensionalities (*dim*=100/200/300) and number of context words (*win*=3/6/10), as well as minimum occurrence cutoff (*min*=1/5), negative samples (*neg*=1/5) and iterations (*iter*=1/5). These variations were tested to ensure the observed patterns reported in the experiments, but we report numerical results only for best performing models. In particular, higher dimensional vectors with *dim*=300 produced consistently better alignment with human scoring data. We also found *min*=1, *neg*=5 and *iter*=5 to be the optimal parameter settings across all experiments.

5 Quantitative Evaluation

Our first experiment follows an established evaluation strategy by computing the Spearman correlation coefficient between the set of similarity measures produced by the word embedding model (WW/CC/WC/CW/AA) and the similarity/relatedness scores taken from the SimLex and ProNorm datasets. As ProNorm score of a word pair (w_1, w_2), we simply use the total number of times a response word w_2 was produced by all subjects given w_1 as a cue word. Interested readers are encouraged to see Jouravlev and McRae (2016) for more details on the data collection procedure.

Our hypothesis is that for taxonomic similarity judgment the classic WW measure, i.e., the cosine of the word vectors of w_1 and w_2 would perform best, especially given the fact that in collection of similarity norms the direction between two words was not a factor. For explicit relatedness judgment, on the other hand, we expect one of the asymmetric measures to be the best predictor. WC, which is the cosine between the word embedding of the cue w_1 and the context embedding of the response w_2 tells us how likely we would see w_2 and similar words in the context of w_1. CW reflects the opposite way relatedness, meaning how likely it is to see w_1 and similar words in the context of w_2. Note that these two quantities are different both mathematically and conceptually, because they are obtained from generalization over word occurrences in many different contexts. We hypothesize that WC should be the best predictor for the ProNorm score of (w_1, w_2) given that production in the constrained setup of the ProNorm experiment was guided by thematic relatedness, making it more like a non-syntactic language modeling task: guessing which other words/concepts might appear within the context of the current word.

SimLex and ProNorm collections have almost the same number of word pairs. However, it is important to note that ordering ProNorm word pairs based on their relatedness scores is probably more difficult than ordering the SimLex list of word pairs. This is because in the ProNorm data collection setup, all word pairs were basically generated based on relatedness, whereas in SimLex, experimental items were pre-designed in a way they covered a wide range of closely similar to totally different word pairs. Ordering SimLex should in turn be harder than ordering words in the old WordSim353 similar and related word pair collections, because each of the latter subsets has a much smaller number of items compared to SimLex collection.

In order to demonstrate the difference between the tasks of ordering words based on similarity vs. relatedness in an explicit setup (SimLex and ProNorm) with an implicit, i.e., a mixed setup we include WordSim353 (Agirre et al., 2009) in our experiment. We hypothesize that the patterns of superiority of one vector-based measure to another in ranking word pairs based on their similarity and relatedness should come out even if people were not explicitly instructed to pay attention to a specific aspect.

5.1 Results

Table 1 displays correlation scores between similarity ratings in SimLex and Skipgram similarity measures introduced in the previous section (all

[3] https://sites.google.com/site/rmyeid/projects/polyglot

Measure	300-3	300-6	300-10
WW	**0.44**	**0.42**	**0.41**
CC	0.40	0.40	0.39
WC	0.34	0.36	0.37
CW	0.32	0.36	0.35
AA	0.42	0.41	**0.41**
AllReg	**0.46**	**0.43**	**0.41**

Table 1: Spearman correlation between human similarity judgments (SimLex) and Skipgram measures.

Measure	300-3	300-6	300-10
WW	0.19	0.22	0.23
CC	0.18	0.19	0.20
WC	**0.24**	**0.25**	**0.26**
CW	0.20	0.20	0.20
AA	0.20	0.22	0.22
AllReg	**0.24**	**0.27**	**0.27**

Table 2: Spearman correlation between forward relatedness scores (ProNorm) and Skipgram measures.

significant at $p < 0.001$). Results on models with *dim*=300 and *win*=3/6/10 are reported (see Appendix for supplementary results). The WW measure exhibits consistently a better alignment with the human rating data compared to the all other measure. This suggests that second-order co-occurrence information plays the main role in similarity between two words. In collection of SimLex, subjects were asked explicitly not to rate similarity based on thematic relatedness. It is likely that the human ratings were affected not only by co-occurrence information encoded in word embeddings but also in context embeddings. As we expected, the best predictors of this data are the symmetric similarity measures, and in particular, WW. The last row of the table includes Spearman correlation between human similarity judgment and a linear regression model using all Skipgram measures as predictors. Thus, numbers in this row show an upper bound for Spearman scores of the individual measures (obtained from an optimal weighting of all individual measures).

Table 2 shows the Spearman correlation between ProNorm scores and the Skipgram measures (all significant at $p < 0.001$). As we hypothesized, WC stands out as the best predictor, suggesting that human responses to a cue word (when asked to name related words) are more likely to be found in the vicinity of the cue word within the context embedding space rather than within the word embedding space. The correlation between the ProNorm scores with WC is larger than with WW or AA scores. This indicates the importance of the knowledge encoded in the context embeddings, but specifically the prediction power of the asymmetric similarity measure compared to the symmetric ones. Interestingly, CW is not as good as WC in this task. This reveals the importance of the direction in associative relatedness between words such as baby and stork, which seems to correlate with their vector representations. Finally, the regression model, which applies an optimal weighing on different Skipgram measures finds the best fit, whereas AA which gives equal weights to symmetric and asymmetric measures fails to compete with WC alone. Comparisons between Tables 1 and 2 suggest that, similarity and relatedness are best approximated by symmetric and asymmetric measures, respectively.

We next examined the WordSim353 data to evaluate whether above implications apply also to ratings collected in implicit setup, i.e., where human subjects were not instructed to response based either on taxonomic similarity or associative relatedness. We examine each subset of WordSim353 separately and treat them like similarity and relatedness data. Table 3 shows results on these two collections of word pairs with best parameter setup; i.e., with *dim*=300 and *win*=3 and 6[4]. Similar to our previous experiments on the other datasets, relative ranking of similar word pairs is best predicted with commonly used measure WW alone, which is indicative of second-order co-occurrence similarity. For related word pairs, asymmetric measures WC and CW, which are indicative of first-level co-occurrence come out as better individual predictors compared to WW. However, the balanced combination of all, i.e., the AA measure seems to be the consistent winner across both datasets. This finding suggests that when similarity/relatedness is scored by people as an overall degree of closeness between words and without explicit instruction to focus on one aspect, the most reliable predictor would be a cosine measure that considers both symmetric and asymmetric types of relations between words.

[4]Results for *win*=10 were not as good as in other conditions for this experiment, therefore we only report the very best setups with *win*=3 and 6.

	Similar pairs		Related pairs	
	300-3	300-6	300-3	300-6
WW	0.79	**0.81**	0.62	0.63
CC	0.76	0.77	0.58	0.61
WC	0.76	0.79	**0.65**	**0.69**
CW	0.76	0.78	**0.65**	0.67
AA	**0.80**	**0.81**	**0.65**	0.67
AllReg	**0.80**	**0.82**	**0.69**	**0.70**

Table 3: Spearman correlation between scores from WordSim353 subsets and Skipgram measures.

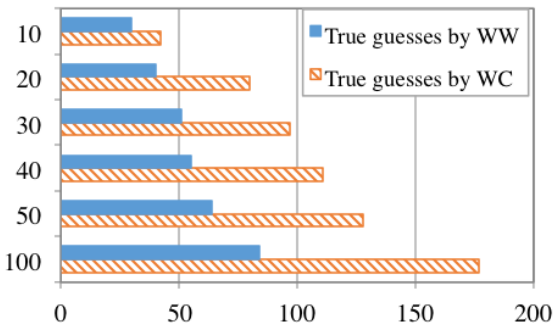

Figure 1: Number of human responses found in word and context embedding spaces near the word embedding of the cue (x-axis) as the search space is increased (y-axis).

6 Qualitative Evaluation

Our first experiment focused on discovering the best vector-based predictor for similarity and relatedness between two words. We found that considering context vectors in calculation of the similarity score produces a superior predictor, specially for relatedness, compared to the traditionally used measure (WW) based only on word vectors. The experiment in this section is a more tangible evaluation of the Word2Vec model in a relatedness task when a cue word is given. The aim is to simulate the production experiment with which the ProNorm data were collected and to evaluate whether using the WC measure will give us more true responses than WW.

For the purpose of this experiment, we use the Skipgram model with *dim*=300 and *win*=10 as these settings produced the best overall performance in the quantitative experiment on ProNorm data. The simulation procedure is as follows: For each cue word w_1 in the ProNorm dataset, each model generates the n most similar words in the vocabulary and we count how many of the human responses were contained in each set. The first model looks up nearest neighbors of w_1 within the word space (thus using WW as the proximity measure) and the second model searches for the nearest neighbors of w_1 within the Context space (thus using WC as the proximity measure). Variable n indicates the total number of guesses a model is allowed to make when responding to a given cue word. In other words, n is the size of the subspace explored around the cue word within each distributional semantic space. Since our previous experiment showed a higher correlation between WC and the relatedness norms, we expect that neighboring words within the context embedding space (in the vicinity of the cues word embedding) should be more populated with related words (i.e., human responses) compared to neighboring words within the word embedding space. Regarding the above procedure, we first extract the word embedding of the cue w_1 and then consider all human responses for that cue, i.e. w_2 of all existing pairs (w_1, w_2) in the dataset, within both the word and context embedding spaces. If, as results of the previous experiment suggest, WC is a better measure of forward relatedness, then a larger portion of human responses should be found in neighboring words within the context space than within the word space surrounding the cue word.

6.1 Results

Our distributional spaces are constructed based on Wikipedia text; therefore, the model vocabulary is very large and noisy. While the top-rank guesses of the model (both measures) are indeed similar/related to the cue words, a lot of them are more frequent in the training corpus genre, i.e. Wikipedia language, than in the simpler language humans (e.g., subjects of the ProNorm study) use when recalling direct relations. For example, in response to the cue word *restaurant* subjects of the ProNorm study generated words such as *plate*, *food*, *menu*, *drink*, and *chef*. In addition to correct guesses, both WW and WC models trained on web corpora generated words such as *bistro*, *eatery*, *hotel*, *grill* and *buffet* as closest words to *restaurant*. Another example would be the cue word *house*, which in the ProNorm experiment triggered *door*, *family*, *bricks*, *bed*, *window*, *roof*, *furniture*, *fireplace*, *chimney*, and *kitchen*. The WW model generated the following words as top candidates, which are in fact taxonomically sim-

ilar, to *house*: *mansion*, *farmhouse*, and *cottage*. WC model generates relatively more thematically related words, some of which are correct guesses (overlapping with human data) and some are not: *barn*, *residence*, *estate*, *dining*, *room*, *stables*, *fireplace*, *family*, and *kitchen*. On average, only one human response per cue can be found in the top 30 model responses. Blue and candy bars in Figure 6 show the total number of correct guesses by the model using WW and WC measures, respectively. This quantity is the total correct guesses for all 100 cues in the ProNorm dataset (x-axis), when the n most similar neighboring words are examined in each space (y-axis). We explored n values between 10 and 100.

Table 4 shows an example of our simulation for the word *car*. As the search space widens up to 100 most similar words in the vicinity of the cue word embedding, more overlap is observed between human responses and model responses. In addition to synonymous words such as *automobile*, the majority of incorrect guesses for the cue word *car* are names of automobile models such as *suv* and *bmw*. The W space around the cue word embedding is more populated with such taxonomically similar words compared to the C space around the cue word. On the other hand, as the results suggest, thematically related words such as *driver* and *steering wheel* can more easily be found within the surrounding C space. This pattern is very consistent across all the cue words in the ProNorm dataset, suggesting that WC is a more valid measure of forward thematic relatedness. This qualitative observation suggests that the differences between the Spearman correlations in Table 2 were meaningful, and vector-based measures of similarity and relatedness, i.e., WW and WC, return different sets of neighboring words to a given cue word.

n		Correct guesses by each measure
20	WW	tires
	WC	tires|driver
50	WW	tires
	WC	tires|driver|driving
100	WW	tires
	WC	tires|driver|driving|steering wheel

Table 4: Human responses for the cue word *car* found in top-n neighboring words within the word and context embedding spaces using WW and WC measures.

7 Related Work

Word embeddings learned from unlabeled text using different models such as Word2Vec and Glove are currently being used for representation of input to deep neural networks that carry out a variety of NLP tasks. Word similarity/relatedness datasets have been the basis for intrinsic evaluation of word embeddings. These datasets provide researchers with insights about how word relations are demonstrated in a distributional space. Previous work has employed WordSim353, SimLex999 and several other established similarity/relatedness datasets for evaluation of word embeddings (Baroni et al., 2014; Kiela et al., 2015; Levy et al., 2015; Melamud et al., 2016; Sahlgren and Lenci, 2016).

A closely related previous study to the current study is the comprehensive evaluation of Word2Vec and three other distributional semantic models by Levy et al. (2015), where they demonstrated that all the models could learn word relations to similar extent if hyper-parameters were carefully tuned. In particular, Levy et al. discussed the effect of averaging word and context vectors on capturing first and second-order similarity. However, the w+c option did not make it to their result tables because it was not selected as one of the generally optimal settings, while mentioned to be useful to test.

Asr and Jones (2017) looked more closely into this optional parameter setting in their study of count-based vs. predictive distributional semantic models (Word2Vec Skipgram vs. PPMI SVD). Using an artificial language framework, they showed that considering the w+c option would extend the range of word-to-word cosine similarity scores, and directly affect the topology of word clusters in the distributional space. However, none of the mentioned works studied the individual terms in the cosine similarity obtained from Word2Vec Skipgram when the w+c option is used, thus they left the question of using these terms for replicating psycholinguistic data on asymmetric association open. Another related line of research in NLP is work on retrofitting of word embeddings using additional lexical resources to reflect specific relations between words more strongly (Faruqui et al., 2015; Kiela et al., 2015). Kiela et al. (2015) looked into the particular case of similarity and relatedness. They pro-

posed using Thesaurus synonymy data and free-association data in training of the word embeddings to obtain vectors suitable for similarity and relatedness, respectively. In contrast to this category of work though, the objective of our research is elaborating the functionality of the word embedding algorithms and how their general-purpose output should be interpreted and queried rather than trying to maximize the performance of the model on a given task by modifying training data or the training mechanism.

Our study adds to the existing body of research by employing word relatedness data collected within a standard psychology experiment and showing how first- vs. second-order information accumulated on the two layers of the popular Skipgram model can be used for different tasks. We showed that the distributional measure for capturing asymmetric relatedness between two words is different from a measure that captures taxonomic similarity even though both types of information are obtained from a unified model trained on a single source of co-occurrence data.

8 Conclusions

Word and context embeddings produced by Word2Vec Skipgram are two different semantic representations of the vocabulary words within the same Euclidean space. We proposed several measures for complementary similarity and relatedness judgments computed based on these embeddings. Asymmetric measures obtained from the inner product of a vector from the word embedding space and a vector from the context embedding space are representative of first-order thematic relations between words.

We examined our proposal using a recently published dataset of production norms (Jouravlev and McRae, 2016) and confirmed when people were explicitly asked to recall thematically related words, their responses were more likely located within the context embedding space in the vicinity of the cues word embedding. In other words, WC, where W is the word embedding of the cue and C is the context embedding of the response, best measures forward thematic relatedness.

We also ran experiments on pure similarity judgment by employing a commonly used dataset of word pairs scored according to taxonomic similarity rather than other types of relations (Hill et al., 2015). Human judgments on word similarity taken from this data were best predicted by a symmetric measure, the classic WW cosine similarity between the word vectors. This suggests that the best measures of taxonomic similarity and thematic relatedness are different in distributional space, even though information involved in both measurements is collected from the same set of co-occurrence features.

Based on the observations made in the paper, we can also argue that the free recall task in the constraint manner where people are asked to name related words (such as in Jouravlev and McRae's study) is similar to the task of predicting context words for the given cue word. This is an important finding for the psycholinguitic research trying to study the mechanisms in lexical production tasks. For NLP research, these findings motivate taking different approaches in problems where thematic relations between words is important for the task, e.g., in assessment of text coherence, question answering, or language generation.

Finally, our experiments elaborated the functionality of the two transformation matrices in Word2Vec architecture. We repeated some of our experiments with GloVe, another popular word embedding model with two final sets of (word/context) vectors. We found similar patterns of relative goodness of measures: WW was consistently better in scoring similarity between two words and WC was better in measuring the thematic relatedness. However, the asymmetry between WC and CW did not come out clearly in these experiments and the overall performance of the GloVe model in the similarity task was much lower than Skipgram. A closer investigation of the GloVe model architecture will be necessary for argumentation about its different results (B Appendix includes results of our preliminary experiments with GloVe). Other vector space models obtained from non-neural architectures can also be examined in this framework. For example, Levy et al. (2015) showed that the w+c option (using the average of word and context embeddings as word vectors) could be simulated in a count-based model that applies SVD to the PMI matrix of word-context co-occurrences. Examining these models on similarity vs. relatedness using our proposed measures will be left for the future.[5]

[5]Code for running all experiments using Word2Vec and GloVe models is available at https://github.com/FTAsr/wordvet

Acknowledgments

We are thankful to our reviewers for their helpful feedback on the initial version of the paper and suggestions for extension of the work. This research was funded by grant R305A140382 from the Institute of Education Sciences, USA.

A Appendix

Supplementary results on SimLex and ProNorm datasets using Skipgram with *dim*=100 & 200 are presented here. Patterns of how WW and CW measures predict similarity and relatedness are consistently repeated in these parameter settings.

Measure	100-3	100-6	100-10
WW	**0.36**	**0.36**	**0.34**
WC	0.31	0.32	0.33
Measure	200-3	200-6	200-10
WW	**0.42**	**0.40**	**0.39**
WC	0.34	0.35	0.35

Table 5: Spearman correlation between similarity scores (SimLex) and Skipgram measures.

Measure	100-3	100-6	100-10
WW	0.18	0.19	0.19
WC	**0.19**	**0.21**	**0.21**
Measure	200-3	200-6	200-10
WW	0.19	0.20	0.21
WC	**0.22**	**0.24**	**0.24**

Table 6: Spearman correlation between similarity scores (ProNorm) and Skipgram measures.

B Supplementary Results using GloVe

Supplementary result on SimLex and ProNorm datasets using GloVe models with *dim*=300 and *win*=3/6/10 are presented in this section. GloVe had a general disadvantage in learning word similarity (SimLex) compared to Skipgram. Patterns of how WW and CW measures predict similarity and relatedness are nevertheless similar across models: WC is much better than WW for relatedness prediction.

Measure	300-3	300-6	300-10
WW	0.25	0.26	0.16
CC	**0.26**	0.25	0.18
WC	0.13	0.17	0.16
CW	0.15	0.17	0.14
AA	**0.26**	**0.27**	**0.20**
AllReg	**0.29**	**0.30**	**0.25**

Table 7: Spearman correlation between similarity scores (SimLex) and GloVe measures.

Measure	300-3	300-6	300-10
WW	0.15	0.16	0.14
CC	0.13	0.17	0.14
WC	**0.22**	**0.21**	0.21
CW	0.19	**0.21**	**0.22**
AA	0.20	**0.21**	0.19
AllReg	**0.22**	**0.21**	**0.24**

Table 8: Spearman correlation between relatedness scores (ProNorm) and GloVe measures.

References

Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, and Aitor Soroa. 2009. A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 19–27.

Tatsuya Aoki, Ryohei Sasano, Hiroya Takamura, and Manabu Okumura. 2017. Distinguishing japanese non-standard usages from standard ones. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2323–2328.

Fatemeh Torabi Asr and Michael Jones. 2017. An artificial language evaluation of distributional semantic models. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 134–142.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 238–247.

Christine Chiarello, Curt Burgess, Lorie Richards, and Alma Pollock. 1990. Semantic and associative priming in the cerebral hemispheres: Some words do, some words don't sometimes, some places. *Brain and language*, 38(1):75–104.

Simon De Deyne, Amy Perfors, and Daniel J Navarro. 2016. Predicting human similarity judgments with distributional models: The value of word associations. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1861–1870.

Simon De Deyne and Gert Storms. 2008. Word associations: Network and semantic properties. *Behavior Research Methods*, 40(1):213–231.

Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy, and Noah A Smith. 2015. Retrofitting word vectors to semantic lexicons. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1606–1615.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on World Wide Web*, pages 406–414. ACM.

Thomas L Griffiths, Mark Steyvers, and Joshua B Tenenbaum. 2007. Topics in semantic representation. *Psychological review*, 114(2):211.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

Keith A Hutchison. 2003. Is semantic priming due to association strength or feature overlap? a microanalytic review. *Psychonomic Bulletin & Review*, 10(4):785–813.

Michael N Jones, Melody Dye, and Brendan T Johns. 2017. Context as an organizing principle of the lexicon. In *Psychology of Learning and Motivation*, volume 67, pages 239–283. Elsevier.

Olessia Jouravlev and Ken McRae. 2016. Thematic relatedness production norms for 100 object concepts. *Behavior research methods*, 48(4):1349–1357.

Douwe Kiela, Felix Hill, and Stephen Clark. 2015. Specializing word embeddings for similarity or relatedness. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2044–2048.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Margery Lucas. 2000. Semantic priming without association: A meta-analytic review. *Psychonomic Bulletin & Review*, 7(4):618–630.

Paweł Mandera, Emmanuel Keuleers, and Marc Brysbaert. 2017. Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*, 92:57–78.

Oren Melamud, Ido Dagan, and Jacob Goldberger. 2015. Modeling word meaning in context with substitute vectors. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 472–482.

Oren Melamud, David McClosky, Siddharth Patwardhan, and Mohit Bansal. 2016. The role of context types and dimensionality in learning word embeddings. *arXiv preprint arXiv:1601.00893*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Douglas L. Nelson, Cathy L. McEvoy, and Thomas A. Schreiber. 1999. The university of south florida word association, rhyme, and fragment norms. retrieved february 5, 2004.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Magnus Sahlgren and Alessandro Lenci. 2016. The effects of data size and frequency range on distributional semantic models. *arXiv preprint arXiv:1609.08293*.

Amos Tversky and Itamar Gati. 1982. Similarity, separability, and the triangle inequality. *Psychological review*, 89(2):123.