Workshop on Linguistic and Neurocognitive Resources (LiNCr2020), pages 1–8
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Extrapolating Binder Style Word Embeddings to New Words

Jacob Turton[1], David Vinson[2], Robert Elliott Smith[1]
[1]Department of Computer Science, University College London
Gower St, London, WC1 EBT
{J.Turton, Rob.Smith}@cs.ucl.ac.uk
[2]Division of Psychology and Language Sciences, University College London
26 Bedford Way, London, WC1H 0AP
d.vinson@ucl.ac.uk

Abstract

Word embeddings such as Word2Vec not only uniquely identify words but also encode important semantic information about them. However, as single entities they are difficult to interpret and their individual dimensions do not have obvious meanings. A more intuitive and interpretable feature space based on neural representations of words was presented by Binder and colleagues (2016) but is only available for a very limited vocabulary. Previous research (Utsumi, 2018) indicates that Binder features can be predicted for words from their embedding vectors (such as Word2Vec), but only looked at the original Binder vocabulary. This paper aimed to demonstrate that Binder features can effectively be predicted for a large number of new words and that the predicted values are sensible. The results supported this, showing that correlations between predicted feature values were consistent with those in the original Binder dataset. Additionally, vectors of predicted values performed comparatively to established embedding models in tests of word-pair semantic similarity. Being able to predict Binder feature space vectors for any number of new words opens up many uses not possible with the original vocabulary size.

Keywords: Semantics, Word-Embeddings, Interpretation

1. Introduction

One of the biggest challenges in computational linguistics is finding representations of words that not only uniquely identify them, but also capture their semantic qualities. A popular approach is distributional semantics (Boleda, 2020), based on the assumption that "a word is characterised by the company it keeps" (Firth, 1957). In practice this means using the co-occurrences of words in large text corpora to derive word embeddings that represent their semantic meaning (Boleda, 2020). Utilising computers makes calculating word co-occurrences in large corpora trivial.

Matrix factorisation approaches such as Latent Semantic Analysis (Landauer et al, 2011) create a term-document matrix and from this produce embeddings for individual words. Alternative matrices can represent term-term co-occurrences or how often words co-occur in sliding window contexts, as is used in the Hyperspace Analogue to Language (HAL) model (Lund & Burgess, 1996).

More recently, models using neural network architectures have proven effective for creating word embeddings. Word2Vec (Mikolov et al, 2013) and GloVe (Pennington. Socher & Manning, 2014) both create word embeddings (typically 300 dimensional) which achieved state of the art results in semantic tasks at their time of introduction. These models are unsupervised; they learn the embeddings from raw text data.

To improve the embeddings, some researchers have proposed infusing them with additional explicit human semantic knowledge. This has resulted in models such as Numberbatch (Speer, Chin & Havasi, 2017), which retrofit the embeddings with information from human created semantic networks, achieving state of the art results in some tests of semantic meaning (e.g. Speer & Lowry-Duda, 2017).

A major difficulty with all word embedding models is interpreting the vectors and validating the semantic information that they capture. By mapping words into a vector space, the relative distance between the embeddings can be used to indicate semantic similarity (Schnabel et al, 2015). This allows word vectors to be understood in terms of their position in vector space in relation to other vectors, but as individual objects in isolation they are difficult to interpret. Furthermore, they offer little insight into *how* the words are related, just that certain words are semantically similar due to their proximity.

This paper proposes mapping word embeddings into a more interpretable feature space, based on the core semantic features of words (Binder et al, 2016). Unfortunately, this feature space currently only exists for a small 535 word vocabulary seriously limiting its uses. Whilst previous work (Utsumi, 2018) has shown that it is possible to derive these feature vectors from embeddings such as Word2Vec, it is still not known how well this scales to a large number of new words. Three experiments were carried out, the first demonstrating that Binder features can be predicted from word embeddings, the second showing that these predictions are sensible for large new word-sets and the third evaluating the performance of the new embeddings in semantic tasks.

By demonstrating that Binder features can be derived for any number of new words, this paper hopes to establish it as a legitimate embedding space.

2. Related Work

2.1 Word Embeddings

Word2Vec, Glove and Numberbatch all represent words as vectors. Word2Vec uses a simple neural network to predict which words should co-occur in a rolling window context. Glove embeddings are derived from a global word co-occurrence matrix. Glove embeddings have been shown to slightly outperform Word2Vec embeddings on certain semantic tasks (Pennington. Socher & Manning, 2014). Numberbatch combines both Word2Vec and GloVe embeddings with information from a semantic network to create a final ensemble embedding for each word. It uses ConceptNet (Speer, Chin & Havasi, 2017) a human created semantic network to inject human level

semantic information into the embeddings. To do this it uses a process called retrofitting whereby the vectors of words connected in the semantic network are pushed closer whilst still remaining as close as possible to their original values.

2.2 Interpreting Embeddings

There have been a number of attempts to improve the interpretability of word embeddings. Dimensionality reduction techniques such as Principle Component Analysis (PCA) or t-Distributed Neighbour Stochastic Embedding (t-SNE) allow the high dimensional embeddings to be visualised in lower two or three dimensional spaces (Liu et al, 2017). Word embeddings can then be interpreted in terms of which other words are visually close to them; a human friendly method of interpretation.

Alternatively, clustering methods can be used to group words according to their distances in vector space. The embeddings can then be interpreted in terms of the clusters created (Zhai, Tan & Choi, 2015).

The methods mentioned so far rely on the location of word embeddings in their vector space and their relative distance to other embeddings for them to be interpretable. Other methods try to make the embeddings more interpretable in themselves. Senel et al (2018) identified 110 semantic categories of words and developed word embeddings represented as weightings across these categories. Whilst this allowed embeddings to be interpreted in isolation, each embedding was now being interpreted in relation to other 'complex concepts'; the categories.

This actually relates to a larger issue in semantics, revolving around how words and concepts are defined. A common belief in cognitive linguistics is that people define concepts in terms of their constituent features (e.g. Cree and McRae, 2003). However, these features themselves are often complex concepts which must be defined in terms of yet more features (Binder et al, 2016). This makes defining a concept difficult and, even more troublingly, can result in circular definitions where concepts are defined in terms of each other. Whilst efforts have been made to identify a set of *primitives*: core irreducible features of meaning, results have been mixed (Drobnak, 2009).

2.3 Reflecting Human Semantic Understanding

Binder et al (2016) proposed an alternative method of defining concepts in terms of a core set of semantic features. In a meta-study, they identified 65 semantic features all thought to have specific neural correlates within the brain. The features were chosen to represent different types of meaning in relation to concepts, from visual, to auditory, to tactile and emotional. They then asked human participants to rate a collection of words across this feature set with scores from 0-5. For example when asked to rate a word for the feature 'Vision', participants were asked: 'To what degree is it something you can easily see?'. The authors collected scores for 535 words; 434 nouns, 62 verbs and 39 adjectives. They also made efforts to include words relating to abstract entities as well as concrete objects. Table 1 below gives an example of the mean scores for Vision, Motion and Time features for the first three words in the Binder dataset.

Word	Vision	Motion	Time	Pleasant	Angry
mosquito	2.9	3.6	0.3	0.2	2.9
ire	1.1	0.6	0.2	0.1	5.0
raspberry	4.6	0.0	0.5	4.1	0.2

Table 1: Example semantic feature scores (5 of 65) for three words from Binder et al (2016)

The features that Binder and colleagues proposed are an attractive embedding space as it allows words to be interpreted individually. Moreover, since each dimension is interpretable, *how* words relate or differ can be seen. Binder et al demonstrated that this could be used to differentiate words based on categories, either narrow e.g. mammals vs fish, or more broad e.g. concrete vs abstract. Moreover, they identified a number of important uses for their feature space, including identifying feature importances, representing how abstract concepts are experienced and understanding how concepts can be combined.

However, for the feature space to be useful it needs to cover a decent proportion of the English vocabulary and they only collected ratings for 535 words. Collecting human ratings for even a moderate coverage of the English vocabulary would be prohibitively expensive and time consuming. Instead, it may be possible to predict the feature scores using word embeddings. Abnar et al (2018) demonstrated that word embeddings could be used to predict neural activation associated with concrete nouns. Since the Binder features are intended to relate to specific neural correlates, the embeddings should be able to be used to predict them. In this direction, Utsumi (2018) demonstrated that Binder feature vectors could successfully be derived from word embeddings including Word2Vec and GloVe for words within the Binder dataset. Taking this further and demonstrating that the features can extrapolated to any number of new words with embeddings would massively expand the feature space vocabulary. Previous studies have shown that it is possible to extrapolate feature scores for new words using distributional embeddings (e.g. Mandera, Kueleers & Brysbaert, 2015) albeit for much smaller feature sets.

3. Experiment 1: Predicting Semantic Features

3.1 Introduction

The purpose of this first experiment was to determine whether the values of the 65 semantic features from Binder et al (2016) could be derived from word embeddings in line with Utsumi (2018). A wider range of regression models (five) were tested plus the previously untested Numberbatch embeddings were included. As Numberbatch embeddings combine both GloVe and Word2Vec and include extra human semantic knowledge, it is expected that they should perform best.

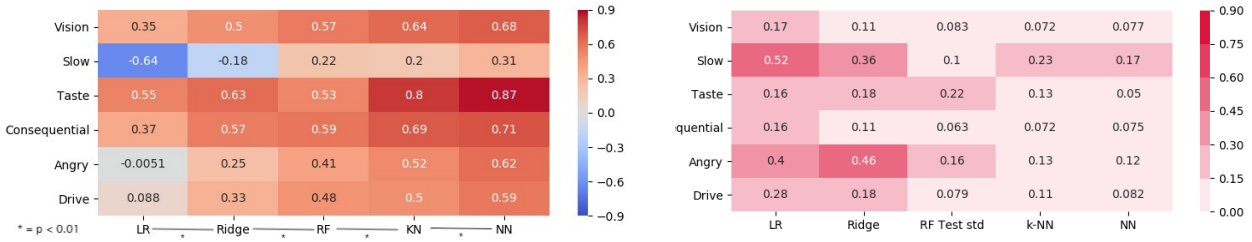

Figure 1: Mean (left) and standard deviation (right) of R-squared scores for six of the 65 Binder semantic features across the 100 test sets. Different models are compared horizontally

3.2 Data

Scores for 535 words across the 65 semantic features were retrieved from Binder et al (2016). Pre-trained Word2Vec (Google, 2013), GloVe (Pennington, Socher & Manning, 2014) and Numberbatch (Speer, Chin & Havasi, 2017) embeddings (all 300 dimensional) were retrieved from their respective online sources. Numberbatch retrofits Word2Vec and GloVe embeddings with information from the ConceptNet knowledge-base.

3.3 Method

Five different types of regression model were compared using GloVe embeddings: linear regressor (LR), ridge regressor (Ridge), random forest regressor (RF), k-nearest neighbours regressor (k-NN) and a 4-layer neural network regressor (NN). Each word in the dataset had a value between 0-5 for each of the 65 semantic features and a 300 dimensional word embedding. The word embeddings were fed into the models as the independent variables with the semantic features as the dependent variables. Separate regression models were trained for each of the features.

For evaluation, models were trained on a subset of the data and their predictions evaluated on a hold-out test set. Because the dataset was fairly small, especially in relation to the number of independent variables (300), there was a risk of overfitting and therefore it was important to maximise the training set size. However, having a test set too small may not appropriately test the models across a representative sample. Utsumi (2018) tackled this using leave-one-out cross validation. In this paper, a bootstrapping method was used instead. 95% of the data was committed for training, and the remaining 5% for testing, but it was repeated over 100 randomly selected train-test set splits. This allowed a large training set whilst testing over a large representative sample of the words overall. The mean and standard deviation of the results could be calculated across the 100 test sets and this also allowed significance testing to compare the model. To ensure fairness, all models were evaluated on the same random 100 train-test splits.

The three different types of word embeddings: Word2Vec, GloVe and Numberbatch were compared using the same method as above.

R-squared was used as the evaluation metric as it was the most intuitive to understand. A Wilcoxon Ranks-sums test was carried out (recommended by Demsar, 2006) to compare the performance of the different models and embeddings.

3.4 Results

Figure 1 above gives the mean R-squared and standard deviations across and six of the 65 features for test set. The six features chosen for Figure 1 represent a mix of concrete and more abstract Binder features.

Table 2 gives the overall mean and standard deviation of test set R-squared scores for each model.

Model	Mean R-sq.	Sd.
Linear Regression	0.03	0.35
Ridge	0.29	0.22
Random Forest	0.41	**0.10**
k-Nearest Neighbours	0.51	0.13
Neural Network	**0.61**	0.11

Table 2: Mean and standard deviation of R-squared scores across all semantic features for the different models.

Table 3 below gives the mean and standard deviation of test set R-squared scores for the different embedding types using the neural network model (best performing model).

Embedding	Mean R-sq.	Sd.
Word2Vec	0.60	0.12
GloVe	0.61	0.10
Numberbatch	**0.65**	**0.09**

Table 3: Overall R-squared mean and standard deviation of different word embeddings

3.5 Discussion

The aim of this experiment was to determine whether semantic feature values from Binder et al (2016) could be derived from word embeddings. In line with the results from Utsumi (2018), this was fairly successful with the best model (Neural Network) achieving an average R-squared of 0.61 across the semantic features, with some features up to ~0.8. Like Utsumi found, there was quite a lot of variation in how well the feature values were predicted, with some such as 'Slow' achieving a relatively low average R-squared (~0.3). Like Utsumi, certain groups of features tended to perform better than others. For example, sensorimotor features such as Toward and Away were more poorly predicted from the embeddings.

However, overall this suggests that for many features a substantial proportion of the variance in human ratings can be derived from word embeddings.

The Neural Network model was the best performing overall, significantly better ($p<0.01$) than the next best performing (k-NN). It was also more consistent than the k-NN model, achieving a lower standard deviation for the features on average. The linear regression model's poor performance may have been due to overfitting as the Ridge regression performed significantly better ($p<0.01$).

Of the word embeddings, Numberbatch (not previously tested in the Utsumi paper) performed the best (0.65), significantly better than both Word2Vec and GloVe ($p<0.01$ for both). This is perhaps not surprising as Numberbatch encourages words connected in a knowledge graph to have similar vectors, and these words will likely also share semantic features.

4. Experiment 2: Predicting Semantic Features for a Larger Vocabulary

4.1 Introduction

Experiment 1 demonstrated that Binder et al (2016) style semantic features could be predicted from word embeddings (albeit with varying success across the features). However, for this to be useful, it is important that the features can be predicted for a much larger vocabulary. Unfortunately, ground truth human ratings for the 65 features only exist for the small Binder et al (2016) dataset, which makes evaluating the predicted scores for new words difficult. Having human scorers evaluate the predicted feature values for new words would be slow and expensive.

One way to overcome this would be to look at the correlations between the semantic features in the human rated Binder dataset and check that they remain consistent for predicted values in a much larger dataset. Binder et al (2016) demonstrated that certain semantic features tended to correlate with each-other across words in their word-set. This pattern of correlations between features should remain consistent within a much larger word-set. Therefore, predicting the semantic values from word embeddings for a new larger word-set of previously unseen words, should give the same or very similar pattern of correlation between the semantic features if the predicted values as sensible.

However, what if the Binder word-set is not a good representation of the wider English vocabulary? As mentioned in the introduction the vast majority of words in the Binder set are nouns, with relatively few verbs or adjectives. The between feature correlations may remain consistent but the predicted semantic values may not be sensible when expanding to a larger new wordset with greater variety of words. Fortunately, much larger datasets of human rated words do exist, but for a much smaller (and slightly different) set of semantic features. The Lancaster Sensorimotor norms (LSN) (Lynott et al, 2019) is a dataset of nearly 40,000 words rated across 11 features by human participants. Some of the features such as *Vision* and *Taste* are very close to features from Binder et al (2016) and all of the words from Binder dataset are included in the larger LSN dataset.

Using the Binder word-set which has human ratings for all of the 65 Binder features and 11 LSN features, the correlations between the LSN and Binder features can be calculated. Then, if the Binder Semantic features are predicted for the larger LSN word-set, it can be checked whether these correlations remain consistent with the LSN features. Since human ratings exist for the 11 LSN features in this larger word-set, it 'grounds' the results. If the pattern of correlations remains consistent, it suggests that the predicted semantic feature values for the new words are sensible.

4.2 Data

The LSN dataset (Lynott et al, 2019) was obtained from their online repository. It consists of 39,707 words rated along 11 features between 0-3.

Numberbatch word embeddings and the Binder et al (2016) dataset from experiment 1 were used again.

4.3 Method

First, the Pearson's correlation was calculated between all 65 semantic features in the Binder et al (2016) dataset, creating a 65×65 correlation matrix. Using the neural network regression model trained in experiment 1 and using Numberbatch embeddings, values for the 65 Binder semantic features across the 39,707 words in the LSN dataset were predicted. The Pearson's correlation between the predicted 65 semantic features across these new words (excluding those also present in the Binder word-set) was calculated, creating another 65×65 correlation matrix. As a numerical measure of similarity, each of the 65 Binder semantic features was represented as a 65 dimensional vector of correlations to all other features, including itself (its row in the correlation matrix). For each feature, the cosine similarity was measured between its correlation vector from the Binder word-set and LSN word-set (ie. 'Vision' Binder vector and 'Vision' LSN vector). Under perfect circumstances, the similarity would be 1 indicating identical vectors. For comparison, cosine similarity of correlation vectors for mismatched features from the Binder and LSN word-set were calculated (e.g. 'Vision' and 'Shape'). It would be expected that these would give a cosine similarity much lower than 1.

The same procedure as above was used for comparing the 11 LSN features to the 65 Binder features. Each of the 11 LSN features was represented as a 65 dimensional vector of correlations with the 65 Binder features. For each feature the cosine similarity between their vectors from the Binder and LSN dataset were calculated. For comparison, the cosine similarity between each of the 11 LSN feature's correlation vectors from the Binder dataset and every other feature's LSN dataset vectors were calculated.

Additionally, a correlation heat-map was created between the features for the Binder and LSN word-sets each separately and then plotted for visual inspection.

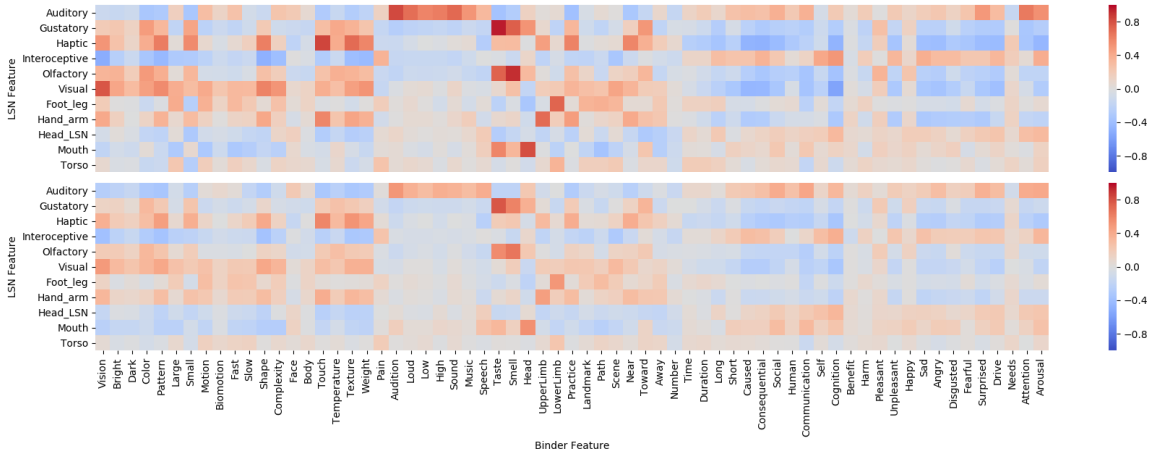

Figure 2: Correlations between the 11 LSN features and 65 Binder semantic features for the Binder word-set (top) and LSN word-set (bottom)

4.4 Results

Table 4 below gives the mean cosine similarity between the same feature correlation vectors from the Binder and LSN word-sets and between different feature vectors.

	Cosine Sim Mean	Cosine Sim S.d.
Same Feature	**0.985**	0.008
Different Features	0.063	0.490

Table 4: Binder semantic feature correlation vector cosine similarities between the Binder and LSN word-sets

Figure 2 above gives the heat-maps for correlations between the 11 LSN features and 65 Binder features for the Binder word-set (top) and LSN word-set (bottom).

4.5 Results

Table 4 below gives the mean cosine similarity between the same feature correlation vectors from the Binder and LSN word-sets and between different feature vectors.

	Cosine Sim Mean	Cosine Sim S.d.
Same Feature	**0.985**	0.008
Different Features	0.063	0.490

Table 4: Binder semantic feature correlation vector cosine similarities between the Binder and LSN word-sets

Figure 2 above gives the heat-maps for correlations between the 11 LSN features and 65 Binder features for the Binder word-set (top) and LSN word-set (bottom).
Table 5 gives the mean cosine similarity of LSN feature correlation vectors between the Binder and LSN word-sets.

	Cosine Sim Mean	Cosine Sim S.d.
Same Feature	**0.94**	0.04
Different Features	-0.02	0.56

Table 5: LSN feature correlation vector cosine similarities between the Binder and LSN word-sets

4.6 Discussion

Table 4 shows that the mean cosine similarity is very high (almost 1) for correlation vectors of the same semantic feature in the Binder and LSN word-sets. This is compared to the very low (almost 0) cosine similarity between the correlation vectors of different features from the Binder and LSN word-sets. This demonstrates that the patterns of correlations between the 65 Binder features remained fairly consistent in the larger LSN word-set where the values had been predicted using the neural network model.

For the 11 LSN features, the heat-maps show a similar pattern of correlations for the features between the Binder and LSN word-sets. The colours are slightly less intense in the LSN word-set suggesting the correlations are slightly weaker. However, this would be expected due to noise from errors in predicting the feature values. The mean cosine similarity is very high (nearly 1) for feature correlation vectors matched across the Binder and LSN word-sets and almost 0 for non-matching features.

Together these results suggest that the values predicted for the 65 semantic features from word embeddings are sensible even in a large and diverse new vocabulary such as the LSN word-set.

5. Experiment 3: Validation of the New Feature Space

5.1 Introduction

Experiments 1 and 2 demonstrated that the values of 65 semantic features could be successfully predicted from word embeddings, and that these appear to be consistent across a large vocabulary of previously unseen words.

Whilst this new feature space is not intended to replace existing embeddings (in fact since it is purely derived from them it almost certainly contains less information about the words) it is still important to demonstrate that it does capture sufficient semantic information.

One of the most common methods for validating word embeddings is using semantic similarity datasets. Typically, these datasets contain pairs of words which are

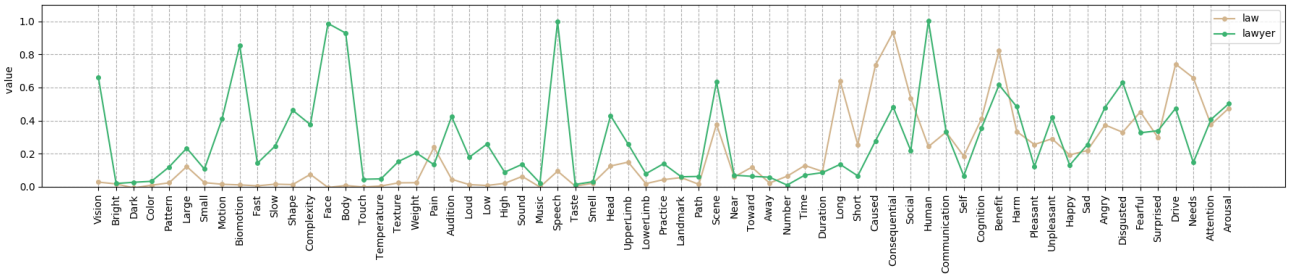

Figure 3: Predicted Semantic Features of words 'law' and 'lawyer'

rated for semantic similarity by human participants. Cosine similarity between word embeddings can be used as a measure of their semantic similarity according to the embedding model. Well performing models should give cosine similarities between words in the pairs that correlate closely to the human ratings.

In the Binder feature space, each word can be represented as a 65 dimensional vector with a value for each of the semantic features. For new words, these vectors can be created by predicting the values for each of the semantic features, similar to in experiments 1 and 2. The cosine similarity between these vectors can then be calculated as a measure of semantic similarity. The aim of this final experiment was to validate the similarity measurements between the predicted vectors against human ratings of similarity.

5.2 Data

Three different word similarity datasets were used: Wordsim-353 (Gabrilovich, 2002), Simlex-999 (Hill, Reichart & Korhonen, 2015) and Verbsim (Yang and Powers, 2006). The same pre-trained Word2Vec, Numberbatch and GloVe embeddings as experiment 1 were used. The same neural network trained on the Binder et al (2016) dataset as experiments 1 & 2 was used.

5.3 Method

A vocabulary was created consisting of all words from the three similarity datasets. Using the trained neural network model and the Numberbatch vectors for the words, values for the 65 semantic features were predicted for the words in the vocabulary. This resulted in a 65 dimensional vector for each of the words, where for each word each dimension was the value of each semantic feature for that word.

For each of the similarity datasets, the cosine similarity was calculated between the semantic feature vectors of the words in each pair. The cosine similarity was also calculated for the Word2Vec, GloVe and Numberbatch embeddings as a comparison.

Spearman's rank correlation was used as it compares the similarity rankings of the word pairs between the human ratings and the ratings derived from the embeddings.

5.4 Results

Table 6 below gives the Spearman's rank coefficient between the word embedding cosine similarity ratings and the ground truth human ratings across the three different word pair datasets.

Embeddings	WordSim353	SimLex999	SimVerb
Word2Vec	0.69	0.44	0.36
GloVe	0.72	0.41	0.28
Numberbatch	**0.83**	**0.63**	**0.57**
Predicted Binder	0.47	0.54	0.46

Table 6: Spearman's Rank correlation between model and human similarity ratings for different word-pair datasets

Whilst Numberbatch embeddings performed best on all datasets, the predicted feature embeddings performed fairly well, beating Word2Vec and GloVe on two of the three datasets. However, the predicted embeddings performed particularly poorly on the Wordsim-353 dataset, performing the worse by far.

5.5 Discussion

At first, the particularly poor performance of the predicted semantic vectors on the Wordsim-353 dataset seems discouraging. However, the Wordsim dataset has received criticism for containing a high proportion of word pairs rated high in similarity through association (e.g. law and lawyer) rather than pure semantic meaning (Hill et al, 2015). Since the predicted Binder semantic embedding space defines words in terms of semantic features, it is understandable that it would not capture similarity due to association as associated words do not necessarily share core features. The SimLex-999 dataset was specifically designed to avoid word pairs with high ratings due to association. The better performance of the predicted feature embeddings on this dataset indicates that the poor performance on the Wordsim dataset was likely due to these association word pairs.

The performance of the predicted embeddings on the SimVerb dataset is also encouraging seeing as there were relatively few verbs in the Binder et al (2016) dataset used for training the prediction model. And it indicates that the model should be suitable for predicting semantic features for new verbs.

Figure 3 above illustrates how two words with high human rated similarity due to association (law and lawyer) are represented by the predicted feature vectors. l this embedding space it can be seen that they are considered very different. Lawyer appears to be represented as a human concrete object: a person as a professional lawyer. The law on the other hand appears to be a more abstract concept (as indicated by the very low

scores across all visual, auditory and sensorimotor features). By these senses the concepts are very different, even though to a human they may seem very similar due to their high association.

Finally, whether word similarity datasets are the best way to evaluate word embedding semantics is debatable. Faruqui et al (2016) provide several shortcomings of word similarity tests for evaluating word embeddings. In light of this, the poorer performance of the Binder embeddings may not mean they do not hold important semantic information

6. General Discussion

The aim of this research was to demonstrate that the Binder semantic feature space for words can be extrapolated to a much larger vocabulary of words using word embeddings. This was important as the Binder word-set is limited to only 535 words.

In line with Utsumi (2018), Experiment 1 demonstrated that Binder features can be derived from word embeddings, with the previously untested Numberbatch embeddings giving the best performance. Like in the Ustumi paper, a neural network architecture model performed best. Experiment 2 demonstrated that the predicted values for a large set of new words appeared sensible, with the internal correlations between the features remaining consistent with human rated words. Finally, experiment 3 showed that this new embedding space retains important semantic information about words, performing comparatively to established embedding models. However, it does not capture associations between words well which may be an important aspect of semantic similarity that it fails on.

The purpose of mapping words to this new feature space is not to replace existing embedding models, but to provide an alternative way to view word embeddings. As Figure 3, on the previous page, illustrates the words represented in this feature space are quite easy to interpret. Furthermore, the semantic features that either differentiate or liken words can easily be identified. The fact that this feature space can be fully derived from existing word embeddings such as Numberbatch, suggests that this semantic information is all present within the word embeddings. However, the range in explained variance between the predicted features does suggest that some semantic information is better captured by word embeddings than other. This is something that Utsumi (2018) investigated in greater detail.

Finally, being able to predict the feature values from existing word embeddings allows the Binder feature space to be extrapolated to a much larger vocabulary. This makes many of the uses for the feature space, outlined in Binder et al (2016), more realistic as their original dataset was too limited in size.

7. Bibliographical References

Abnar, S., Ahmed, R., Mijnheer, M. & Zuidema, W. (2018). Experiential, Distributional and Dependency-based Word Embeddings have Complementary Roles in Decoding Brain Activity, *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics*, 57-66.

Binder, J. R., Conant, L. L., Humphries, C. J., Fernandino, L., Simons, S. B., Aguilar, M., & Desai, R. H. (2016). Toward a Brain-Based Componential Semantic Representation. *Cognitive neuropsychology*, 33(3-4):130-174.

Boleda, G. (2020). Distributional Semantics and Linguistic Theory. *Annual Review of Linguistics*. 6:213-234

Cree, G. S., & McRae, K. (2003). Analyzing the Factors Underlying the Structure and Computation of the Meaning of Chipmunk, Cherry, Chisel, Cheese, and Cello (and Many Other Such Concrete Nouns). *Journal of experimental psychology: general*, 132(2):163.

Demšar, J. (2006). Statistical Comparisons of Classifiers Over Multiple Data Sets. *Journal of Machine learning research*, 7:1-30.

Drobnak, F. T. (2009). On the Merits and Shortcomings of Semantic Primes and Natural Semantic Metalanguage in Cross-Cultural Translation. *ELOPE: English Language Overseas Perspectives and Enquiries*. 6(1-2):29-41.

Faruqui, M., Tsvetkov, Y., Rastogi, P. & Dyer, C. (2016) Problems With Evaluation of Word Embeddings Using Word Similarity Tasks, *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*. 30-35

Firth, J.R. (1957). Applications of General Linguistics, *Transactions of the Philological Society*. 56(1):1-14

Landauer, T. K., McNamara, D. S., Dennis, S., & Kintsch, W. (Eds.). (2013). *Handbook of latent semantic analysis*. Psychology Press. New York.

Liu, S., Bremer, P. T., Thiagarajan, J. J., Srikumar, V., Wang, B., Livnat, Y., & Pascucci, V. (2017). Visual Exploration of Semantic Relationships in Neural Word Embeddings. *IEEE transactions on visualization and computer graphics*, 24(1):553-562.

Lund, K., & Burgess, C. (1996). Producing High-Dimensional Semantic Spaces from Lexical Co-occurrence. *Behavior research methods, instruments, & computers*, 28(2):203-208.

Lynott, D., Connell, L., Brysbaert, M., Brand, J., & Carney, J. (2019). The Lancaster Sensorimotor Norms: Multidimensional Measures of Perceptual and Action Strength for 40,000 English Words. *Behavior Research Methods*, 1-21.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *arXiv preprint*. ArXiv:1301.3781

Mandera, P., Keuleers, E., & Brysbaert, M. (2015). How Useful are Corpus-Based Methods for Extrapolating Psycholinguistic Variables?. *The Quarterly Journal of Experimental Psychology*. 68(8):1623-1642.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global Vectors for Word Representation. *In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. Pages 1532-1543.

Schnabel, T., Labutov, I., Mimno, D., & Joachims, T. (2015, September). Evaluation Methods for Unsupervised Word Embeddings. *In Proceedings of the 2015 conference on empirical methods in natural language processing*. Pages 298-307.

Şenel, L. K., Utlu, I., Yücesoy, V., Koc, A., & Cukur, T. (2018). Semantic Structure and Interpretability of Word Embeddings. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(10):1769-1779.

Speer, R., Chin, J., & Havasi, C. (2017, February). Conceptnet 5.5: An Open Multilingual Graph of General Knowledge. *In Thirty-First AAAI Conference on Artificial Intelligence*.

Speer, R., & Lowry-Duda, J. (2017). Conceptnet at Semeval-2017 Task 2: Extending Word Embeddings With Multilingual Relational Knowledge. *arXiv preprint arXiv*:1704.03560.

Utsumi, A. (2018). A Neurobiologically Motivated Analysis of Distributional Semantic Models. *Proceedings of the 40th Annual Conference of the Cognitive Science Society,*1147-1152.

Zhai, M., Tan, J., & Choi, J. D. (2016, March). Intrinsic and Extrinsic Evaluations of Word Embeddings. *In Thirtieth AAAI Conference on Artificial Intelligence*.

8. Language Resource References

Hill, F., Reichart, R. & Korhonen, A. (2015). SimLex-999: Evaluating Semantic Models with (Genuine) Similarity Estimation. *Computational Linguistics*. https://fh295.github.io/simlex.html

Gabrivolvich, E. (2002). The WordSimilairty-353 Test Collection. http://www.cs.technion.ac.il/~gabr/resources/data/wordsim353/

Google. (2013). Google Code Archive, Word2Vec. https://code.google.com/archive/p/word2vec/

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global Vectors for Word Representation. *In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. https://nlp.stanford.edu/projects/glove/

Speer, R., Chin, J., & Havasi, C. (2017, February). Conceptnet 5.5: An Open Multilingual Graph of General Knowledge. *In Thirty-First AAAI Conference on Artificia Intelligence*. https://github.com/commonsense/conceptnet-numberbatch

Yang, D. & Powers, D. M. (2006). *Verb Similarity on the Taxonomy of WordNet*. Masaryk University.