SWOW-8500: Word Association Task for Intrinsic Evaluation of Word Embeddings

Avijit Thawani
IIT BHU / Varanasi
thawani@usc.edu

Biplav Srivastava
IBM / New York
biplav.srivastava@gmail.com

Anil Kumar Singh
IIT BHU / Varanasi
aksingh.cse@iitbhu.ac.in

Abstract

Downstream evaluation of pretrained word embeddings is expensive, more so for tasks where current state of the art models are very large architectures. Intrinsic evaluation using word similarity or analogy datasets, on the other hand, suffers from several disadvantages. We propose a novel intrinsic evaluation task employing large word association datasets (particularly the Small World of Words dataset). We observe correlations not just between performances on SWOW-8500 and previously proposed intrinsic tasks of word similarity prediction, but also with downstream tasks (eg. Text Classification and Natural Language Inference). Most importantly, we report better confidence intervals for scores on our word association task, with no fall in correlation with downstream performance.

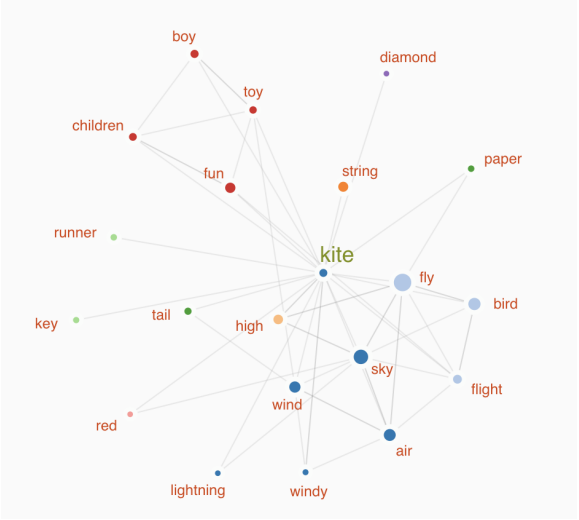

Figure 1: Visualization of the cue *Kite* and its associated words according to the SWOW dataset. Source: https://smallworldofwords.org/en/project/explore

1 Introduction

With the recent rise in popularity of distributional semantics, word embeddings have become the basic building block of several state-of-the-art models spanning multiple problems across Natural Language Processing and Information Retrieval. Word embeddings are essentially non-sparse representations of words in the form of one (relatively) small dimensional vector of real numbers for every word, and all of these vectors lie in the same continuous space.

Despite the clear benefits of these distributed representations, it is not obvious how to come up with apt word embeddings for a given NLP task. Approaches such as word2vec (Mikolov et al., 2013b), GloVe (Pennington et al., 2014), etc. have been shown to perform well on downstream tasks such as text classification, sequence labelling, question answering, text summarization, and machine translation.

Typically, word vectors are used in NLP models in two ways: fixed pretrained embeddings, and finetuning. In the first way, word vectors have already been trained on some large dataset (e.g. Wikipedia, Twitter, Blog corpus, etc.) using one of the aforementioned techniques. These vectors are taken as fixed weights and the model merely uses them as they are rather than learning them during the training phase. On the other hand, finetuning allows for these vectors to be modified too, using backpropagation. Here the word embeddings are taken only as initialized weights for the model's first layer.

It is of natural interest to the NLP community to identify evaluation metrics for word embeddings. Besides direct performance measurement on downstream tasks, there have also been proposed several intrinsic evaluation measures such as MEN, WordSim, SimLex, etc. These are small proxy tasks which word vectors are expected to perform well on, given the assumption that they capture semantics of words. While Extrinsic evaluations use word embeddings as input features to

Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP, pages 43–51
Minneapolis, USA, June 6, 2019.

a downstream task and measure changes in performance metrics specific to that task, Intrinsic evaluations directly test for syntactic or semantic relationships between words (Schnabel et al., 2015). For example, the word similarity task asks word embeddings to predict how similar are the meanings of two prompt words. The closer this estimate is to human judgements, higher is the score allotted to the (pretrained) word embedding.

Through this paper, we propose the Word Association task for evaluating non-contextualized pretrained word embeddings, with the help of word asssociation datasets originally collected for psychological research. The datasets were formed by asking participants to respond to certain cue words. For example, given the cue *tiger*, one could respond with the words *lion*, *panther*, *wild*, etc. Large datasets of this sort are now available online, and it can be argued that they capture a notion of which words are in close association with others (as perceived by human participants).

According to cognitive theories of the mind, people form associations between concepts based on similarity, contiguity, or contrast. Our task proposal stems from the following argument: Any model that claims to understand the semantics of words should be able to mimic human beings in recognizing the associations between pairs of words. For example, a distributed representation of words, i.e., word embeddings, should be able to tell that the word *tiger* is in some way associated with *lion* but not with, say, *kettle*, assuming such a statistic is observed in the word association dataset too.

Given the scale of these datasets, they seem like a lucrative way to evaluate pretrained word embeddings. We see them as a manually annotated corpus of word associations, though not originally meant for word embedding evaluation. Therefore, we must devise a convenient way to compare the semantics captured in a given set of pretrained word vectors with that captured in such word association datasets.

We make our scripts, along with several other resources, available at https://github.com/avi-jit/SWOW-eval

2 Related Work

2.1 Word Embedding Evaluation

There exist several intrinsic evaluation tasks for word embeddings. One way to tell apart intrinsic from extrinsic evaluations is the lack of any trainable parameters in the former. Schnabel et al. (2015) discuss word relatedness, analogy, selective preference, and categorization as types of intrinsic tasks.

Our proposed task is most similar to the word relatedness/similarity tasks, several of which have already been proposed in literature: WS-3533 (Finkelstein et al., 2002), WS-SIM and WS-REL (Agirre et al., 2009), RG-65 (Rubenstein and Goodenough, 1965), MC-30 (Miller and Charles, 1991), MTurk-2875 (Radinsky et al., 2011), MTurk-771 (Halawi et al., 2012), MEN7 (Bruni et al., 2012), YP-130 (Yang and Powers, 2006), Rare Words (Luong et al., 2013), etc. We list the ones above specifically since those are the ones we compare our proposed task to, using the online resource wordvectors.org (Faruqui and Dyer, 2014), whose code remains available on GitHub [1]. Association of Computational Linguistics[2] and Vecto AI[3] also maintain benchmark pages for word similarity.

Likewise, VecEval (Nayak et al., 2016) and Multilingual-embeddings-eval-portal (Ammar et al., 2016) are GitHub repositories for Extrinsic Evaluation of word embeddings. [4,5].

Another direction of work has been towards critiquing intrinsic evaluation, in a bid to understand its shortcomings and potential workarounds (Schnabel et al., 2015; Zhai et al., 2016). One of the key shortcomings is **the Absence of Statistical Significance** (Faruqui et al., 2016), which we aim to tackle through this proposal. We believe that a massive dataset of word associations can be used to circumvent issues related to confidence intervals of scores reported. We put this belief to test in later sections of this paper.

2.2 Word Association

Our prime motivation behind this work was Marvin Minsky's Society of Mind (1988) which theorizes that humans learn by linking concepts together, using what Minsky calls K-Lines. If we assign meanings to concepts by associating them

[1] http://github.com/mfaruqui/eval-word-vectors

[2] http://aclweb.org/aclwiki/Similarity_(State_of_the_art)

[3] http://github.com/vecto-ai/word-benchmarks

[4] http://github.com/NehaNayak/veceval

[5] http://github.com/wammar/multilingual-embeddings-eval-portal

with each other, artificial models of semantics should also be able to do the same.

Word Association games are those wherein a participant is asked to utter the first (or first few) words that occur to him/her when given a trigger/cue/stimulus word. For example, given *king*, one could respond with *rule*, *queen*, *kingdom*, or even *kong* (from the movie King Kong). Word associations have long intrigued psychologists including Carl Jung (1918) and hence large studies have been conducted in this direction. Some prominent datasets which collect user responses to word association games are:

1. University of Southern Florida: Free Association (USF-FA) (Nelson et al., 2004) has single-word association responses from an average of 149 participants per cue for a set of 5,019 cue words.

2. Edinburgh Association Thesaurus (EAT) (Kiss et al., 1973) collects 100 responses per cue for a total of 8,400 cues.

3. JeuxDeMots: over 5 million french words (Lafourcade, 2007).

4. **Small World of Words** (SWOW) (De Deyne et al., 2018): Word association and participant data for 100 primary, secondary and tertiary responses to 12,292 cues, collected from over 90,000 participants[6].

5. Birkbeck norms (Moss et al., 1996) contain 40 to 50 responses for over 2,600 cues in British English.

Among non-English word association norms, the largest resources available include 16000 cues in Dutch (De Deyne et al., 2013), 3900 cues in Korean (Jung et al., 2010), and 2100 cues in Japanese (Joyce, 2005).

The authors of SWOW and Jeux De Mots have even attempted to employ their word association datasets for learning word embeddings (De Deyne et al., 2016; Plu et al.) . They use both count-based and random walk based strategies to learn vector representations of words. Note that we differ in using the SWOW dataset not as a corpus to learn word vectors, but as a human annotated dataset for evaluating other pretrained word vectors.

cue	response	R123	N	R123.Str
would	should	63	288	0.220
would	could	63	288	0.220
would	will	24	288	0.083
would	can	11	288	0.038
...	...	...	...	...
stumble	fall	76	290	0.262
stumble	trip	68	290	0.234
stumble	upon	16	290	0.055

Table 1: A few example cue-response tuples from the SWOW dataset, along with their associated R123.Strength scores

Summary Statistic	Value
Sample Minimum (the smallest observation)	115
Lower Quartile (the first quartile)	270
Median (the middle value)	282
Upper Quartile (the third quartile)	289
Sample Maximum (the largest observation)	300

Table 2: The five number summary for N, i.e. the number of responses per cue

3 Dataset

Here onwards, we restrict ourselves to only the Small World of Words dataset (SWOW), a part of which can be seen in Table 1. For each cue-response pair C-R, the value $R123$ is the number of participants who responded with R when given the cue C. Note that out of at most three responses collected per cue per respondent, it does not matter to the $R123$ score whether R occurred in the first response or the third. N is the number of total responses given the cue C in the processed version of released SWOW dataset. The value $R123.Strength$ is simply equal to $\frac{R123}{N}$.

There are $978,908$ cue-response pairs in the latest release of SWOW dataset. The statistics for the number of responses per cue is shown in Table 2. For our own SWOW evaluations, we got rid of anything that was not a single word, e.g. *New York* or *get-together*. We further selected only the most

[6] http://smallworldofwords.org/en/project

frequently co-occurring word associations. In particular, we kept only those cue-response pairs that have $R123.strength$ (i.e., number of people who cited this response for this cue within any of the three responses they gave, divided by the total number of responses for this cue word) is greater than 0.2 which corresponds to saying that at least one fifth of all respondents believe this response is one of the three top associated words for the given cue. We were now left with 8500 cues and a few of their corresponding top responses each. While we restrict ourselves to experimenting only on the SWOW-8500 dataset, we make available the code and resources to create even larger datasets (with fewer restrictions on, say, minimum strength of association, needed). [7]

Note that word association datasets are asymmetric in that they treat the pairs C-R and R-C separately, i.e., for the cue *coffee*, the response *tea* might be the most frequent one but for the cue *tea*, the most frequent response could be *black*. We need to bear this in mind when using this dataset to evaluate word embeddings intrinsically, since usually intrinsic datasets give out a single value for a word pair. This also does not fit well with the traditional measure of similarity/relatedness between two words, i.e., cosine distance, which is a symmetric metric.

4 Methodology

Supplemental Table 4 in the original SWOW dataset paper (De Deyne et al., 2018) shows high correlations with a few of the word similarity datasets mentioned above. In this aspect, our work can be seen as their direct successor, since we build upon these correlations to propose a new (and larger) task for intrinsic evaluation of word vectors.

We wish to compare performances of any pretrained word embedding on (1) Our proposed task, (2) other Intrinsic evaluation tasks, and (3) Downstream Tasks. To that end, we first settle upon some candidate word embeddings. All embeddings had 300 dimensions, thereby avoiding different numbers of parameters to be learnt for downstream models. They were reduced to a very small common vocabulary of 7779 words. This helped in conveniently expressing results, without accounting for Out-of-Vocabulary words differently. We attempt to have a representative set of embeddings, including the best and most popular ones:

1. Word2Vec Skip Gram (Mikolov et al., 2013b,a) trained on Google News.[8]
2. GloVe (Pennington et al., 2014) trained on Wikipedia 2014 and Gigaword 5.[9]
3. FastText (Bojanowski et al., 2017) trained with subword information on Common Crawl (600B tokens).[10]
4. ConceptNet Numberbatch (Speer et al., 2017) trained on a big knowledge graph and some text corpora.[11]
5. Baroni and Lenci's (2014) count-based embeddings, which are the result of dimensionality reduction on a large count matrix.[12]
6. Random Baseline: a baseline developed by randomly allotting 300 floating numbers to each word in the common vocabulary of the above five embeddings.

We used intrinsic evaluations in the form of 13 word similarity tasks, provided by wordvectors.org (Faruqui and Dyer, 2014). For our proposed task SWOW-8500, and for a given pretrained embedding E, we ask E to predict top k responses for each of the 6481 cues (the ones in common between the 7779 sized vocabulary of our word vectors, and the 8500 cues in our proposed task). This corresponds to listing the top-k most similar words to the cue (which we have found from decreasing order of cosine similarity). We tried with several fixed values of k but finally report results keeping k variable, and always equal to the number of responses for that particular cue (in the SWOW-8500 dataset). Here k can be thought of as the number of guesses allotted to an image classifier. We then report how many of the correct responses (according to SWOW dataset) also occurred in the guesses made by E.

The True Positives are those words that occur both in SWOW-8500 as well as E's guesses. False

[7] https://github.com/avi-jit/SWOW-eval

[8] http://code.google.com/archive/p/word2vec/

[9] http://nlp.stanford.edu/projects/glove/

[10] http://fasttext.cc/docs/en/english-vectors.html

[11] http://github.com/commonsense/conceptnet-numberbatch

[12] http://clic.cimec.unitn.it/dm/

cue	Correct Guesses (TP)	Incorrect Guesses (FP)	Couldn't Guess (FN)
		ConceptNet Numberbatch	
assassination	murder	assassin, killing	president, kill
sect	religion, cult	religious	group
newt	salamander	democrat, republican	lizard, amphibian
		Baroni and Lenci (Count-based)	
assassination	-	killing, kidnapping, massacre	president, murder, kill
sect	cult	fraternity, republic	group, religion
newt	salamander	ladybird, alligator	lizard, amphibian

Table 3: Responses to Cues by two of the compared pretrained embeddings, along with ground truth responses

Positives are those words that were correct responses (according to SWOW-8500) but could not be guessed by E (not present in SWOW-8500). False Negatives correspond to wrong guesses by E. Note that since no ground truth responses are labelled as negative (i.e., we only have words that *should* be present in the response set for a given cue), the number of True Negatives is always 0. From a confusion matrix, we can report accuracy, error, precision, recall, F1 score, and also a confidence interval for the error score.

Lastly, we also conduct downstream evaluation of embeddings on five tasks (Sentiment Analysis, Chunking, Natural Language Inference, Named Entity Recognition, and POS Tagging) using the VecEval framework (Nayak et al., 2016). The original framework uses, on top of the embedding layers, LSTMs for some of the tasks. This brings up the question of which other architectures should then be tried out. Since bidirectional language modelling has been shown to outperform a simple left-to-right traversal (Devlin et al., 2018), should biLSTMs be used instead? What about Transformers, or self-attention layers (Vaswani et al., 2017)? To avoid a very large number of model parameters, and to conveniently report results only about the word embeddings like we intend to, we instead chose to go ahead with simple feed forward neural networks (one or two hidden layers) and no LSTM layers. Based on several experiments, we chose our hyperparameters as: 50 neurons per hidden layer, a dropout of 0.5, and 50 epochs with a batch size of 128. For details of the tasks and data involved, please refer to their paper or webpage. [13]

[13] http://veceval.com

5 Results

Table 3 is a sample from the cues and responses in the SWOW-8500 task. For each cue, the ground truth extracted from SWOW is the union of the words shown under columns *Correct Guesses (True Positives)* and *Couldn't Guess (False Negatives)*. It is noteworthy how (qualitatively) close-to-correct are the responses by ConceptNet as opposed to those by the Count-Based embedding, and as we shall see, the same holds in the quantitative scores assigned to the two, by SWOW-8500 task.

Table 4 shows performance of the selected pretrained embeddings on intrinsic evaluation: the upper half covering existing word similarity datasets and the lower half covers SWOW-8500. ConceptNet Numberbatch seems to outperform all the others, which could be attributed to it being based on a knowledge graph that links words based on what concepts people think are associated. Table 5 shows performances on Downstream tasks.

From Table 5 and the upper half of Table 4, one can see a good correlation between intrinsic and extrinsic evaluations, contrary to past reports (Faruqui et al., 2016), at least for Fixed versions of the tasks. However for model runs where Finetuning was allowed, and with a large enough training set, even Random Baseline embeddings quickly came at par with the others. This goes to show that, for the mostly classification-type tasks that we considered, requiring little linguistic knowledge and relying on topical semantics, our proposed task acts as a great proxy.

Within Table 4, we notice how the Precision, Recall, and F1 scores (from our proposed SWOW-8500 task) correlate well with all intrinsic evalua-

	CN	FT	GloVe	w2v	Count	Base	Pairs	OOV
EN-MEN-TR-3k	0.855	0.806	0.744	0.771	0.254	0.014	3000	423
CI width:	**0.027**	**0.036**	**0.046**	**0.041**	**0.095**	**0.102**		
EN-MC-30	0.932	0.940	0.902	0.916	0.658	0.264	30	10
CI width:	0.190	0.177	0.276	0.240	0.725	1.054		
EN-MTurk-771	0.839	0.740	0.659	0.685	0.228	0.034	771	192
CI width:	0.064	0.098	0.122	0.114	0.203	0.213		
EN-SIMLEX-999	0.638	0.426	0.359	0.435	0.179	0.030	999	113
CI width:	0.103	0.141	0.151	0.141	0.170	0.173		
EN-VERB-143	0.569	0.324	0.454	0.538	0.360	0.072	144	124
CI width:	0.833	1.026	0.940	0.864	1.005	1.105		
EN-YP-130	0.727	0.542	0.524	0.463	0.178	-0.077	130	48
CI width:	0.274	0.403	0.413	0.445	0.541	0.554		
EN-RW-STANFORD	0.815	0.666	0.552	0.648	0.401	-0.113	2034	1952
CI width:	0.200	0.325	0.402	0.338	0.480	0.558		
EN-RG-65	0.939	0.943	0.862	0.819	0.492	-0.084	65	20
CI width:	0.202	0.196	0.217	0.275	0.593	0.751		
EN-WS-353-ALL	0.814	0.738	0.615	0.707	0.287	-0.051	353	88
CI width:	0.108	0.145	0.198	0.160	0.295	0.315		
EN-WS-353-SIM	0.842	0.826	0.683	0.776	0.448	0.002	203	43
CI width:	0.121	0.132	0.221	0.166	0.327	0.406		
EN-WS-353-REL	0.771	0.709	0.608	0.659	0.090	-0.163	252	68
CI width:	0.157	0.192	0.242	0.217	0.375	0.369		
EN-MTurk-287	0.863	0.816	0.764	0.779	0.261	-0.253	287	187
CI width:	0.137	0.179	0.221	0.210	0.478	0.481		
EN-SimVerb-3500	0.580	0.337	0.208	0.341	0.088	-0.022	3500	694
CI width:	0.064	0.086	0.093	0.086	0.096	0.098		
Precision	0.254	0.223	0.171	0.169	0.059	0.000		
Recall	0.280	0.246	0.189	0.186	0.065	0.000		
F1 Score	0.266	0.233	0.180	0.177	0.061	0.000	8500	2019
Error	0.746	0.777	0.829	0.831	0.941	1.000		
(Error CI width)	**0.008**	**0.008**	**0.007**	**0.007**	**0.005**	**0.000**		

Table 4: Intrinsic tasks performance. CN: ConceptNet Numberbatch; FT: FastText; Count: Baroni and Lenci. Pairs: Number of word pairs in the dataset; OOV: Number of word pairs of which at least one word was missing (for upper half of table) or Number of cues missing (for lower half of table) in the common vocabulary shared by the six pre-trained embeddings. All Confidence Intervals (CI) reported at 99% confidence level.

tions. Thus SWOW task captures more or less the same properties already captured by existing word similarity datasets. So far the only added advantage is that it has already been built (along with others like USF and EAT), and therefore did not require additional expensive annotation efforts.

The large scale of SWOW also offers a solution to the underlying shortcomings in intrinsic evaluations: reporting statistical signficance. As evident from Table 4, SWOW-8500 offers up to three times narrower confidence intervals for error rate, as opposed to the best amongst word similarity datasets, i.e. EN-MEN-TR-3k. The table cites all values at Confidence Intervals 99%. Even at

	CN	FT	GloVe	w2v	Count	Random
Ques Fixed	0.6245	0.5055	0.6099	0.6264	0.5000	0.2234
Ques Finetuned	0.7015	0.7143	0.6978	0.7033	0.4506	0.7033
Senti Fixed	0.6984	0.6663	0.5436	0.6766	0.4874	0.5092
Senti Finetuned	0.6318	0.6445	0.6468	0.6480	0.5344	0.6640
Chunk Fixed	0.6598	0.6605	0.5980	0.6352	0.4168	0.3138
Chunk Finetuned	0.5824	0.5682	0.5925	0.6002	0.3863	0.3138
NLI Fixed	0.4142	0.4222	0.3234	0.3234	0.3345	0.3233
NLI Finetuned	0.4312	0.4303	0.4334	0.4280	0.3398	0.4245
NER Fixed	0.9264	0.9297	0.9226	0.9245	0.8332	0.8332
NER Finetuned	0.9145	0.9124	0.9190	0.9197	0.8332	0.8332
POS Fixed	0.6625	0.6695	0.6285	0.6547	0.3609	0.3244
POS Finetuned	0.5323	0.5305	0.5491	0.5456	0.3535	0.3244

Table 5: Downstream tasks performance. CN: ConceptNet Numberbatch; FT: FastText; Count: Baroni and Lenci

a more modest confidence level of 90%, for the largest intrinsic dataset, i.e. SimVerb with 3500 word pairs, the accuracy of Numberbatch embeddings at 90% confidence could be reported within a span of 0.039. The smallest dataset MC had 30 data points, leading to a 90% confidence span of 0.114. For SWOW with 6481 data points, the error rate can be reported with a 90% confidence span of 0.003. Thus, we have greater confidence in reporting SWOW evaluations than with previous intrinsic datasets, yet have little difference in actual (relative) scores reported.

The Confidence Intervals for correlation scores reported are based on the Fischer Transformation (Fisher, 1915). The transformation is defined as $z_r = \frac{\ln\left(\frac{1+r}{1-r}\right)}{2}$, where r is the correlation coefficient. Thereafter, the confidence interval (lower and upper limits) can be computed as: $\hat{z} = z_r \pm \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{N-3}}$, where N is the number of pairs of observations, (in our case the number of pairs shared with vocabulary).

Confidence Interval for the SWOW-8500, which is a classification task, is reported as the Wilson Score Interval (Wilson, 1927). The error interval (lower and upper limits) are defined as: $e = \hat{e} \pm z\sqrt{\frac{e(1-e)}{n}}$, where e is the error value, z is the constant (equal 2.58 for 99% CI), and n is the number of observations evaluated upon (equal to the total number of responses for all cues in SWOW-8500).

6 Conclusions and Future Work

In this paper, we've suggested a new breed of intrinsic evaluation tasks, that rely not on word similarity but on word association. More concretely, we use the Small World of Words dataset to create SWOW-8500, an intrinsic evaluation task We describe the task, and compare performance for six word embeddings, on (1) our proposed task, on (2) *thirteen* word similarity tasks, and on (3) *five* downstream tasks.

We find that the same sets of properties as captured by word similarity datasets, which have been shown to correlate with downstream tasks as well, are also captured by the Word Association task SWOW-8500. To add to that, we report higher confidence scores which shall help in reporting significance of results on intrinsic evaluation better. Thus we hope to dispel the suspicion over results reported using the (relatively) small word similarity datasets, since they are now corroborated with much larger human studies as well.

There remain several interesting directions to be explored, primarily the use of even more Word Association datasets (mentioned in Section 2.2). While in this paper, we've cited only the Response Prediction task, we tried out several others, including a Word Similarity task, and a Response Ordering task. With further experimentation, it would be interesting to see what properties of embeddings do these variations capture. Lastly, more downstream tasks could be tested for correlation, e.g. morphological analysis.

References

Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Pasca, 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, Boulder, Colorado. Association for Computational Linguistics.

Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A. Smith. 2016. Massively multilingual word embeddings. *CoRR*, abs/1602.01925.

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)*, pages 238–247, Baltimore, Maryland. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5(1):135–146.

Elia Bruni, Gemma Boleda, Marco Baroni, and Nam Khanh Tran. 2012. Distributional semantics in technicolor. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 136–145, Jeju Island, Korea. Association for Computational Linguistics.

Simon De Deyne, Daniel J Navarro, and Gert Storms. 2013. Associative strength and semantic activation in the mental lexicon: evidence from continued word associations. Cognitive Science Society.

Simon De Deyne, Danielle J Navarro, Amy Perfors, Marc Brysbaert, and Gert Storms. 2018. The small world of words english word association norms for over 12,000 cue words. *Behavior research methods*, pages 1–20.

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, Osaka, Japan. The COLING 2016 Organizing Committee.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Manaal Faruqui and Chris Dyer. 2014. Community evaluation and exchange of word vectors at wordvectors.org. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 19–24, Baltimore, Maryland. Association for Computational Linguistics.

Manaal Faruqui, Yulia Tsvetkov, Pushpendre Rastogi, and Chris Dyer. 2016. Problems with evaluation of word embeddings using word similarity tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 30–35, Berlin, Germany. Association for Computational Linguistics.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2002. Placing search in context: the concept revisited. *ACM Trans. Inf. Syst.*, 20(1):116–131.

R. A. Fisher. 1915. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, 10(4):507–521.

Guy Halawi, Gideon Dror, Evgeniy Gabrilovich, and Yehuda Koren. 2012. Large-scale learning of word relatedness with constraints. In *The 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '12, Beijing, China, August 12-16, 2012*, pages 1406–1414.

Terry Joyce. 2005. Constructing a large-scale database of japanese word associations. *Glottometrics*, 10:82–99.

Carl Gustav Jung. 1918. *Studies in word-association*. W. Heinemann, Limited.

Jaeyoung Jung, Li Na, and Hiroyuki Akama. 2010. Network analysis of korean word associations. In *Proceedings of the NAACL HLT 2010 First Workshop on Computational Neurolinguistics*, pages 27–35. Association for Computational Linguistics.

George R Kiss, Christine Armstrong, Robert Milroy, and James Piper. 1973. An associative thesaurus of english and its computer analysis. *The computer and literary studies*, pages 153–165.

Mathieu Lafourcade. 2007. Making people play for lexical acquisition with the jeuxdemots prototype. In *SNLP'07: 7th international symposium on natural language processing*, page 7.

Thang Luong, Richard Socher, and Christopher Manning. 2013. Better word representations with recursive neural networks for morphology. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 104–113, Sofia, Bulgaria. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013,*

Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, pages 3111–3119.

George A Miller and Walter G Charles. 1991. Contextual correlates of semantic similarity. *Language and cognitive processes*, 6(1):1–28.

Marvin Minsky. 1988. *Society of mind*. Simon and Schuster.

Helen Moss, Lianne Older, and Lianne JE Older. 1996. *Birkbeck word association norms*. Psychology Press.

Neha Nayak, Gabor Angeli, and Christopher D. Manning. 2016. Evaluating word embeddings using a representative suite of practical tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 19–23, Berlin, Germany. Association for Computational Linguistics.

Douglas L Nelson, Cathy L McEvoy, and Thomas A Schreiber. 2004. The university of south florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, 36(3):402–407.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Julien Plu, Kévin Cousot, Mathieu Lafourcade, Raphaël Troncy, and Giuseppe Rizzo. Jeuxdeliens: Word embeddings and path-based similarity for entity linking using the french jeuxdemots lexical semantic network. In *Actes de la conférence Traitement Automatique de la Langue Naturelle, TALN 2018*, page 529.

Kira Radinsky, Eugene Agichtein, Evgeniy Gabrilovich, and Shaul Markovitch. 2011. A word at a time: computing word relatedness using temporal semantic analysis. In *Proceedings of the 20th International Conference on World Wide Web, WWW 2011, Hyderabad, India, March 28 - April 1, 2011*, pages 337–346.

Herbert Rubenstein and John B. Goodenough. 1965. Contextual correlates of synonymy. *Commun. ACM*, 8(10):627–633.

Tobias Schnabel, Igor Labutov, David Mimno, and Thorsten Joachims. 2015. Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 298–307, Lisbon, Portugal. Association for Computational Linguistics.

Robert Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA.*, pages 4444–4451.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 6000–6010.

Edwin B. Wilson. 1927. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212.

Dongqiang Yang and David Martin Powers. 2006. *Verb similarity on the taxonomy of WordNet*. Masaryk University.

Michael Zhai, Johnny Tan, and Jinho D. Choi. 2016. Intrinsic and extrinsic evaluations of word embeddings. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA.*, pages 4282–4283.