Predicting Word Association Strengths

Andrew Cattle **Xiaojuan Ma**
Hong Kong University of Science and Technology
Department of Computer Science and Engineering
Clear Water Bay, Hong Kong
{acattle,mxj}@cse.ust.hk

Abstract

This paper looks at the task of predicting word association strengths across three datasets; WordNet Evocation (Boyd-Graber et al., 2006), University of Southern Florida Free Association norms (Nelson et al., 2004), and Edinburgh Associative Thesaurus (Kiss et al., 1973). We achieve results of $r = 0.357$ and $\rho = 0.379$, $r = 0.344$ and $\rho = 0.300$, an $\rho = 0.292$ and $\rho = 0.363$, respectively. We find Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) cosine similarities, as well as vector offsets, to be the highest performing features. Furthermore, we examine the usefulness of Gaussian embeddings (Vilnis and McCallum, 2014) for predicting word association strength, the first work to do so.

1 Introduction

Word embeddings such as Word2Vec (Mikolov et al., 2013) or GloVe (Pennington et al., 2014) have received increasing attention in the world of natural language processing and computational linguistics. Under such embeddings, the semantic relatedness of two words is generally taken to be the cosine similarity of their word vectors. Although this approach performs well for variety of applications, it is not without its limitations. First, it defines "relatedness" quite narrowly as the extent to which the two words appear in similar contexts. Second, it fails to capture how humans internally represent words (De Deyne et al., 2016b).

Word associations offer a more flexible view of semantic relatedness by leveraging "lexical knowledge acquired through world experience" (Nelson et al., 2004). While word embeddings capture distributional relationships, word associations are able to capture more nuanced relationships "which are based on human perception and experiences [and] are not reflected in common language usage." (Ma, 2013) For example, "yellow" is so closely associated with "banana" that many people would only specify a banana's colour if it is not yellow. This is backed up by De Deyne et al. (2016b) which found word associations performed better than word embeddings across a variety of semantic relatedness tasks.

Furthermore, word associations, unlike cosine similarities, are asymmetric; when presented with the word "beer", many people think of the word "glass" but when presented with the word "glass", few people think of the word "beer" (Ma, 2013). This directionality allows for more fine-grained exploration of semantic links, with applications in word similarity (Jabeen et al., 2013) and computational humour (Cattle and Ma, 2016).

Although several word association datasets exist, such as the Edinburgh Associative Thesaurus (EAT, Kiss et al., 1973), the University of South Florida Free Association Norms (USF, Nelson et al., 2004), or WordNet Evocation (Evocation, Boyd-Graber et al., 2006), their reliance on human annotations mean they all suffer from coverage issues relating to limited vocabularies or sparse connectivity (Cattle and Ma, 2016; De Deyne et al., 2016b). Although these issues would be somewhat alleviated by the creation of larger datasets, collecting human judgments for all possible word pairs is impractical. Therefore, the ability to predict association strengths between arbitrary word pairs represents the best solution to these coverage issues (Boyd-Graber et al., 2006).

Although the prediction of Evocation ratings has attracted some attention (Boyd-Graber et al., 2006; Hayashi, 2016), to the best of our knowledge this is the first work to focus on the prediction of USF or EAT strengths. As described in Sec-

Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 1294–1299
Copenhagen, Denmark, September 7–11, 2017.

tion 2, USF and EAT have several advantages over Evocation, such as the ability to work with ambiguous words instead of WordNet synsets. Following Hayashi (2016)'s work on Evocation prediction, we frame word association prediction as a supervised regression task and introduce several new and modified features, including the first use of Gaussian embeddings (Vilnis and McCallum, 2014) to better capture the asymmetric nature of word associations.

2 Previous Work

Word association has been used in psychological and psycholinguistic experiments for well over 100 years (Boyd-Graber et al., 2006; De Deyne and Storms, 2008). Word association datasets such as USF or EAT have typically framed word association as "a task that requires participants to produce the first word to come to mind that is related in a specified way to a presented cue" (Nelson et al., 2000). These datasets use *forward strengths*, the proportion of participants who produce a specific response, to "index the relative accessibility of related words in memory [for a given cue]" (Nelson et al., 2004).

This cue/response framework has several drawbacks. First, since forward strengths are relative, comparing strengths across different cue words is difficult. Second, both cues and responses are ambiguous, with each participant's responses being greatly influenced by how they chose to interpret a given cue. For example, someone responding to the cue "brother" with "monk" is considering a different sense of "brother" than someone who responds "sister" (Ma, 2013). As such, forward strengths are biased toward responses which presume more readily apparent cue word senses. Third, limiting participants to a single response can lead to weaker associations being underreported or omitted entirely.

Evocation solves the ambiguity issue by focusing on associations between WordNet synsets. Boyd-Graber et al. (2006) presented participants with randomly selected synset pairs and asked them to score how much the first synset *evoked* (i.e. brought to mind) the second. Unlike forward strengths, these Evocation ratings are absolute, meaning they can be directly compared across different cues. While randomly selecting synset pairs ensured that weaker associations would not be underreported, it did have the disadvantage that 67% of pairs were unanimously rated as having no connection (Boyd-Graber et al., 2006).

Despite attempts to address this spareness issue by expanding Evocation with data gathered from Amazon Mechanical Turk[1] (Nikolova et al., 2009) or word-sense disambiguated USF cue/response pairs (Ma, 2013), obtaining human judgments for all possible synset pairs is impractical. As such, the prediction of Evocation ratings presents the most promising solution to this coverage issue. Boyd-Graber et al. (2006) detailed a simple Evocation estimator which used a combination of WordNet structure-based features, WordNet definition-based features, and corpus-based word co-occurrence features. However, this approach is somewhat limited in that it frames Evocation prediction as a classification task, considering only five Evocation levels.

The main drawback of Evocation prediction as a classification task is that it is too coarse-grained to deal with very weak associations, such as those in remote triads (De Deyne et al., 2016a), or very slight variations in association strength, such as those useful for computational humour (Cattle and Ma, 2016). To this end, Hayashi (2016) framed Evocation prediction as a supervised regression task. They employed a combination of WordNet structure-based features, word embedding-based features, and lexical features and found that vector offsets, i.e. the mathematical difference between vectors, were a strong indicator of Evocation ratings.

While Evocation's use of unambiguous synsets is useful for many applications, it is not without its own drawbacks. First, it requires texts to be word sense disambiguated; a non-trivial task. Second, since humans do not conceptualize words as a discrete set of independent word senses, Evocation is unable to capture natural associations owing to homography, homophony, or polysemy (Ma, 2013). As such, despite their drawbacks, word associations may provide a more flexible, more holistic view of mental semantics.

By allowing participants to record more than one response, De Deyne and Storms (2008), and their derivative works De Deyne et al. (2013) and De Deyne et al. (2016b), were able to better represent weaker associations. However, this introduced its own set of problems as great care had to be taken to avoid chaining, i.e. responding to a

[1] https://mturk.com/

previous response instead of the cue, and retrieval inhibition. De Deyne and Storms (2008) frames word association collection as a continuous task, meaning not only that the vocabulary is ever growing but also that changes in associations over time can be observed and tracked. But despite the steps taken to improve the size and quality of their association dataset, practicality dictates that coverage issues cannot be completely eliminated.

3 System Definition

Our word association prediction system extends the method in Hayashi (2016) with several modifications to make it better suited to the USF and EAT datasets.

First, we modify Hayashi (2016)'s *lexVector*. Hayashi (2016) represent each word's part-of-speech (POS) using a one-hot encoded five dimensional vector (one of each POS in WordNet). Similarly, they represent each word's lexical category using a one-hot encoded 45 dimensional vector (one for each WordNet lexicographer file). This results in a 100 dimensional vector representing the POS and lexical categories of both the cue and the response. Since words in USF and EAT can be associated with multiple synsets and we want to be able to capture associations related to polysemy, instead using a one-hot encoding we employ count vectors specifying the number of synsets from each POS/lexical category each word belongs to.

Second, instead of computing Wu-Palmer similarity (WUP, Wu and Palmer, 1994) between a single synset pair, we compute it for all cue synset/response synset pairs and record the maximum and average values. Following Boyd-Graber et al. (2006) and Ma (2013), we also explored the use of path and Leacock-Chodorow (Leacock and Chodorow, 1998) similarities but found they did not add any advantage over WUP alone. We take a similar approach for adapting load and betweenness centralities (Barthelemy, 2004) as well as AutoExtend (AutoEx, Rothe and Schütze, 2015) similarity.

Third, we extend the notion of *dirRel*, introduced in Hayashi (2016) to leverage the semantic network structure of WordNet. Given a graph where nodes represent synsets and arcs represent WordNet relations such as hypernym/hyponym and holonym/meronym, *dirRel(s,t,k)* is the proportion of k-step neighbours of s that are also k-step neighbours of t. In the original formula, s and t are nodes representing a single synset. We instead consider a set of nodes S and a set of nodes T representing the set of synsets associated with the cue and response words, respectively, as shown in Equation 1. This may increase the probability that $|nb(S,k) \cap nb(T,k)| > 0$, a shortcoming of the original *dirRel* due to WordNet's "relatively sparse connective structure" (Hayashi, 2016).

$$dirRel(S,T,k) = \frac{|nb(S,k) \cap nb(T,k)|}{|nb(S,k)|} \quad (1)$$

Fourth, in addition to the Word2Vec (w2v) cosine similarity between cue/response pairs calculated using Google's pre-trained 300 dimension Word2Vec embeddings[2]. We also examine the effectiveness of Stanford's pre-trained 300 dimension GloVe embeddings[3].

Fifth, in order to better capture asymmetric word associations, we propose using Gaussian embeddings. Gaussian embeddings (Vilnis and McCallum, 2014) represent words not as a fixed point in vector space but as "potential functions", continuous densities in latent space; therefore, they are more suitable for capturing asymmetric relationships. More specifically, for each cue/response pair, we calculate both the KL-divergence and cosine similarities of their Gaussian embeddings. The embeddings have a dimensionality of 300 and are trained on English Wikipedia using the Word2Gauss[4] (w2g) and the hyperparameters reported by the developer[5]

Sixth, since cue and response words are not associated with a single synset, the AutoEx embeddings employed in Hayashi (2016) to compute vector offsets are not well suited for our task. Instead, we experiment with offsets calculated using w2v, GloVe, and w2g embeddings.

Finally, our 300 topic LDA model (Blei et al., 2003) was trained using Gensim[6] on full English Wikipedia instead of the subset of English Wikipedia used in Hayashi (2016).

Using the above features, we trained a multilayer perceptron for each of our three datasets; Evocation, USF, and EAT. In the case of Evocation, we discarded any synset information and

[2]https://code.google.com/archive/p/word2vec/
[3]https://nlp.stanford.edu/projects/glove/
[4]https://github.com/seomoz/word2gauss
[5]https://github.com/seomoz/word2gauss/issues/18#issuecomment-286203006
[6]https://radimrehurek.com/gensim/

Feature	Evocation		USF		EAT	
	r	ρ	r	ρ	r	ρ
Hayashi (2016)	0.374	0.401	—	—	—	—
All (w/ w2v offsets)	0.357	0.379	0.344	0.300	0.292	0.363
betweenness (max)	−0.000	0.004	0.008	0.019	0.035	0.112
betweenness (avg)	−0.002	−0.001	0.012	0.004	0.002	0.021
load (max)	−0.009	−0.010	0.017	0.025	0.039	0.118
load (avg)	−0.007	−0.006	0.002	0.004	0.002	0.017
WUP sim (max)	0.098	0.136	0.092	0.111	0.049	−0.026
WUP sim (avg)	0.051	0.062	0.045	0.051	0.033	0.014
lexVector	0.115	0.117	0.091	0.077	0.105	0.249
dirRel	0.177	0.149	0.152	0.130	0.124	0.049
LDA cos sim	0.129	0.033	0.054	0.040	0.046	0.007
AutoEx cos sim (max)	0.135	0.144	0.124	0.132	0.054	−0.034
AutoEx cos sim (avg)	0.148	0.174	0.082	0.089	0.045	0.019
w2v cos sim	**0.265**	**0.264**	**0.229**	0.226	0.150	0.094
GloVe cos sim	0.239	0.262	0.222	**0.232**	0.117	−0.010
w2g cos sim	0.227	0.246	0.173	0.185	0.109	0.046
w2g KL-divergence	0.110	0.185	−0.013	−0.011	0.086	0.205
w2v offsets	0.010	0.009	0.092	0.076	0.144	0.299
GloVe offsets	0.007	0.009	0.127	0.098	**0.162**	**0.344**
w2g offsets	−0.005	−0.003	0.073	0.065	0.111	0.186

Table 1: Individual feature performance after 50 epochs

Feature	Evocation		USF		EAT	
	r	ρ	r	ρ	r	ρ
All (w/ w2v offsets)	0.357	0.379	0.344	0.300	0.292	0.363
- betweenness (max)	0.360	0.383	0.341	0.301	0.270	0.360
- betweenness (avg)	0.357	0.376	0.331	0.298	0.284	0.353
- load (max)	0.358	0.382	0.339	0.299	0.290	0.375
- load (avg)	0.360	0.381	0.340	0.304	0.279	0.353
- WUP sim (max)	0.367	0.376	0.333	0.294	0.283	0.367
- WUP sim (avg)	0.355	0.374	0.335	0.296	0.291	0.364
- lexVector	0.351	0.365	0.336	0.300	0.275	0.339
- dirRel	0.356	0.375	0.334	0.293	0.283	0.362
- LDA cos sim	0.357	0.377	0.340	0.299	0.291	0.361
- AutoEx cos sim (max)	0.362	0.382	0.347	0.299	0.280	0.358
- AutoEx cos sim (avg)	0.358	0.377	0.346	0.305	0.278	0.357
- w2v cos sim	0.352	0.377	0.331	0.294	0.280	0.345
- GloVe cos sim	0.352	0.367	0.332	0.292	0.284	0.360
- w2v and GloVe sims	**0.329**	**0.342**	**0.284**	**0.255**	0.261	0.353
- w2g cos sim	0.358	0.378	0.348	0.304	0.284	0.357
- w2g KL-divergence	0.351	0.356	0.344	0.299	0.286	0.348
- w2v offsets	0.361	0.386	0.303	0.280	**0.239**	**0.271**

Table 2: Ablation performance after 50 epochs

simply use each synset's headword (e.g. given the sysnet *entity.n.01*, we only considered the word *entity*). Following the setup used in Hayashi (2016), all neural networks are trained using the Chainer[7] Python library with rectified linear units, dropout, and two hidden layers, each with 50% of the number of units in the input layer. All were trained on 80% of their respective dataset, with 20% held out for testing. Mean squared error was used as a loss function and optimization was performed using Adam algorithm (Kingma and Ba, 2014). To act as a baseline, we also reimplemented the system described in Hayashi (2016) and trained it on the same 80/20 split of Evocation as our system. In addition to the reported results, we also performed feature selection experiments using 20% of the training sets as validation.

4 Results and Discussion

The performance of individual features are reported in Table 1 while the results of our ablation experiments are reported in Table 2. For all experiments we report both the Pearson correlation coefficient (as r) and Spearman's rank correlation coefficient (as ρ).

The best performing single feature on Evocation and USF is w2v cosine similarity. However, its removal in the ablation test had little effect. This is likely due to redundancy between w2v and GloVe; not only does GloVe perform similarly to w2v but removing both features at the same time produced the largest drop in performance. It is unclear why word embedding cosine similarities in general performed relatively poorly on EAT. While the USF and EAT datasets are very similar, EAT does seem to contain a greater number of multi-word cues/responses which would not be in the word embedding vocabularies. In such cases, perhaps a multi-word embedding like Doc2Vec (Le and Mikolov, 2014) would be more appropriate. However, if this were indeed the issue, one would expect vector offsets to perform equally poorly. This is not the case, with GloVe offsets being the best performing single feature on EAT and the removal of w2v offsets causing the greatest drop in performance in the EAT ablation tests.

The results of our Hayashi (2016) implementation are roughly comparable to those reported in the original paper ($r = 0.374$, $\rho = 0.401$ compared to $r = 0.439$, $\rho = 0.400$). Our slightly lower Pearson's R may be due to differences in way we split our training and test data as well as due to randomness in the training process itself.

On Evocation, our system does not perform as well as Hayashi (2016). This is expected as we explicitly ignore any synset information and instead attempt to predict association strengths between word-sense ambiguous words. Despite this, our performance is not appreciably lower, indicating the fitness of our system.

The fact that we perform better on Evocation than either USF or EAT is quite interesting considering our system was designed with USF and EAT in mind. There are several possible explanations for this. First, as mentioned in Section 2, 67% of cue/response pairs in Evocation have a strength of zero. This uniformity in Evocation strengths may make them easier to predict. Second, due to the way USF and EAT were collected, there are no true zeros in the datasets. This lack of grounding

[7] http://chainer.org/

may skew the predictions. Third, this may be an indication that predicting associations in a word-sense ambiguous context is a harder task than predicting them in a word-sense disambiguated one. Boyd-Graber et al. (2006) explicitly told annotators to ignore associations based on polysemy or rhyme. It could be the case that these effects are more difficult to identify.

Another possible explanation for this relatively lower performance is a lack of bespoke features. For example, we heavily rely on WordNet-based features which make sense in a word-sense disambiguated context but are less suited for ambiguous contexts. In fact, removal of several of these features, such as betweenness or AutoEx similarity, seem to slightly improve performance. One explanation is that, despite noting in Section2 that word association strengths are influenced by word-sense frequencies, our system instead implicitly assumes all synsets are equally likely.

The most surprising finding was the poor performance of Gaussian embeddings overall, and KL-divergence in particular. Given the asymmetric nature of word associations, KL-divergence seemed to be a natural fit. However, it is vastly outperformed by even cosine similarity on the same set of embeddings. Despite this, the usefulness of Gaussian embeddings cannot be ruled out. While we used pre-trained embeddings for Word2Vec and GloVe, we had to train our own Gaussian embedding model. Not only were Word2Vec and GloVe trained on much larger corpora than Gaussian embedding's English Wikipedia, but the pre-trained embeddings likely underwent a greater degree of hyperparameter tuning.

5 Conclusions and Future Works

In this paper we explored the effectiveness of various features for predicting Evocation, USF, and EAT association strengths, finding GloVe and Word2Vec cosine similarities as well as vector offsets to be the most useful features. We also examined the effectiveness of Gaussian embeddings for capturing the asymmetric nature of word embeddings but found it to be less effective than traditional word embeddings.

Although we report a lower performance than that in Hayashi (2016), potentially indicating that predicting association strengths in word-sense ambiguous contexts is a harder task, we believe our results are a promising start. Training Gaussian embeddings on a larger corpus may lead to improved effectiveness. Future works should also consider incorporating word-sense frequencies or developing word-sense agnostic features, with a particular focus on asymmetric features.

References

Marc Barthelemy. 2004. Betweenness centrality in large complex networks. *The European Physical Journal B-Condensed Matter and Complex Systems* 38(2):163–168.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3(Jan):993–1022.

Jordan Boyd-Graber, Christiane Fellbaum, Daniel Osherson, and Robert Schapire. 2006. Adding dense, weighted connections to wordnet."in: Proceedings of the third global wordnet meeting, jeju island, korea, january 2006 .

Andrew Cattle and Xiaojuan Ma. 2016. Effects of semantic relatedness between setups and punchlines in twitter hashtag games. *PEOPLES 2016* page 70.

Simon De Deyne, Daniel J Navarro, Amy Perfors, and Gert Storms. 2016a. Structure at every scale: A semantic network account of the similarities between unrelated concepts. *Journal of Experimental Psychology: General* 145(9):1228.

Simon De Deyne, Daniel J Navarro, and Gert Storms. 2013. Better explanations of lexical and semantic cognition using networks derived from continued rather than single-word associations. *Behavior Research Methods* 45(2):480–498.

Simon De Deyne, Amy Perfors, and J. Daniel Navarro. 2016b. 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*. The COLING 2016 Organizing Committee, pages 1861–1870. http://aclweb.org/anthology/C16-1175.

Simon De Deyne and Gert Storms. 2008. Word associations: Norms for 1,424 dutch words in a continuous task. *Behavior Research Methods* 40(1):198–205.

Yoshihiko Hayashi. 2016. Predicting the evocation relation between lexicalized concepts. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. The COLING 2016 Organizing Committee, Osaka, Japan, pages 1657–1668. http://aclweb.org/anthology/C16-1156.

Shahida Jabeen, Xiaoying Gao, and Peter Andreae. 2013. Directional context helps: Guiding semantic

relatedness computation by asymmetric word associations. In *WISE (1)*. pages 92–101.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .

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.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*. pages 1188–1196.

Claudia Leacock and Martin Chodorow. 1998. Combining local context and wordnet similarity for word sense identification. *WordNet: An electronic lexical database* 49(2):265–283.

Xiaojuan Ma. 2013. Evocation: analyzing and propagating a semantic link based on free word association. *Language resources and evaluation* 47(3):819–837.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. 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 Simon Dennis. 2000. What is free association and what does it measure? *Memory & cognition* 28(6):887–899.

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.

Sonya Nikolova, Jordan Boyd-Graber, Christiane Fellbaum, and Perry Cook. 2009. Better vocabularies for assistive communication aids: connecting terms using semantic networks and untrained annotators. In *Proceedings of the 11th international ACM SIGACCESS conference on Computers and accessibility*. ACM, pages 171–178.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*. volume 14, pages 1532–1543.

Sascha Rothe and Hinrich Schütze. 2015. Autoextend: Extending word embeddings to embeddings for synsets and lexemes. In *Proceedings of the ACL*.

Luke Vilnis and Andrew McCallum. 2014. Word representations via gaussian embedding. *CoRR* abs/1412.6623. http://arxiv.org/abs/1412.6623.

Zhibiao Wu and Martha Palmer. 1994. Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, pages 133–138.