Robust Gram Embeddings

Taygun Kekeç and **D.M.J. Tax**
Pattern Recognition and Bioinformatics Laboratory
Delft University of Technology
Delft, 2628CD, The Netherlands
taygunkekec@gmail.com,D.M.J.Tax@tudelft.nl

Abstract

Word embedding models learn vectorial word representations that can be used in a variety of NLP applications. When training data is scarce, these models risk losing their generalization abilities due to the complexity of the models and the overfitting to finite data. We propose a regularized embedding formulation, called *Robust Gram* (RG), which penalizes overfitting by suppressing the disparity between target and context embeddings. Our experimental analysis shows that the RG model trained on small datasets generalizes better compared to alternatives, is more robust to variations in the training set, and correlates well to human similarities in a set of word similarity tasks.

1 Introduction

Word embeddings represent each word as a unique vector in a linear vector space, encoding particular semantic and syntactic structure of the natural language (Arora et al., 2016). In various lingual tasks, these sequence prediction models shown superior results over the traditional count-based models (Baroni et al., 2014). Tasks such as sentiment analysis (Maas et al., 2011) and sarcasm detection (Ghosh et al., 2015) enjoys the merits of these features.

These word embeddings optimize features and predictors simultaneously, which can be interpreted as a factorization of the word cooccurence matrix C. In most realistic scenarios these models have to be learned from a small training set. Furthermore, word distributions are often skewed, and optimizing the reconstruction of $\hat{C}$ puts too much emphasis on the high frequency pairs (Levy and Goldberg, 2014). On the other hand, by having an unlucky and scarce data sample, the estimated $\hat{C}$ rapidly deviates from the underlying true cooccurence, in particular for low-frequency pairs (Lemaire and Denhire, 2008). Finally, noise (caused by stemming, removal of high frequency pairs, typographical errors, etc.) can increase the estimation error heavily (Arora et al., 2015).

It is challenging to derive a computationally tractable algorithm that solves all these problems. Spectral factorization approaches usually employ Laplace smoothing or a type of SVD weighting to alleviate the effect of the noise (Turney and Pantel, 2010). Alternatively, iteratively optimized embeddings such as Skip Gram (SG) model (Mikolov et al., 2013b) developed various mechanisms such as undersampling of highly frequent hub words apriori, and throwing rare words out of the training.

Here we propose a fast, effective and generalizable embedding approach, called Robust Gram, that penalizes complexity arising from the factorized embedding spaces. This design alleviates the need from tuning the aforementioned pseudo-priors and the preprocessing procedures. Experimental results show that our regularized model 1) generalizes better given a small set of samples while other methods yield insufficient generalization 2) is more robust to arbitrary perturbations in the sample set and alternations in the preprocessing specifications 3) achieves much better performance on word similarity task, especially when similarity pairs contains unique and hardly observed words in the vocabulary.

Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 1060–1065,
Austin, Texas, November 1-5, 2016.

2 Robust Gram Embeddings

Let $|y| = V$ the vocabulary size and N be the total number of training samples. Denote x, y to be $V \times 1$ discrete word indicators for the context and target: corresponding to the context and word indicators c, w in word embedding literature. Define $\Psi_{d\times V}$ and $\Phi_{d\times V}$ as word and context embedding matrices. The projection on the matrix column space, Φx, gives us the embedding $\vec{x} \in R^d$. We use Φx and Φ_x interchangeably. Using a very general formulation for the regularized optimization of a (embedding) model, the following objective is minimized:

$$J = \sum_{i}^{N} \mathcal{L}(\Psi, \Phi, x_i, y_i) + g(\Psi, \Phi) \qquad (1)$$

where $\mathcal{L}(\Psi, \Phi, x_i, y_i)$ is the loss incurred by embedding example target y_i using context x_i and embedding parameters Ψ, Φ, and where $g(\Psi, \Phi)$ is a regularization of the embedding parameters. Different embedding methods differ in the form of specified loss function and regularization. For instance, the Skip Gram likelihood aims to maximize the following conditional:

$$\begin{aligned} \mathcal{L}(\Psi, \Phi, x, y) &= -\log p(y|x, \Phi, \Psi) \\ &= -\log \frac{\exp(\Psi_y^T \Phi_x)}{\sum_{y'} \exp(\Psi_{y'}^T \Phi_x)} \end{aligned} \qquad (2)$$

This can be interpreted as a generalization of Multinomial Logistic Regression (MLR). Rewriting $(\Psi y)^T(\Phi x) = (y^T \Psi^T \Phi x) = y^T W x = W_y x$ shows that the combination of Φ and Ψ become the weights in the MLR. In the regression the input x is transformed to directly predict y. The Skip Gram model, however, transforms both the context x and the target y, and can therefore be seen as a generalization of the MLR.

It is also possible to penalize the quadratic loss between embeddings (Globerson et al., 2007):

$$\mathcal{L}(.) = -\log \frac{\exp(-||\Psi_y - \Phi_x||^2)}{\sum_{y'} \exp(-||\Psi_{y'} - \Phi_x||^2)} \qquad (3)$$

Since these formulations predefine a particular embedding dimensionality d, they impose a low rank constraint on the factorization $W = \Psi^T \Phi$. This means that $g(\Psi, \Phi)$ contains $\lambda rank(\Phi^T \Psi)$ with a sufficiently large λ. The optimization with an explicit rank constraint is NP hard. Instead, approximate rank constraints are utilized with a Trace Norm (Fazel et al., 2001) or Max Norm (Srebro and Shraibman, 2005). However, adding such constraints usually requires semidefinite programs which quickly becomes computationally prohibitive even with a moderate vocabulary size.

Do these formulations penalize the complexity? Embeddings under quadratic loss are already regularized and avoids trivial solutions thanks to the second term. They also incorporate a bit weighted data-dependent ℓ_2 norm. Nevertheless, choosing a log-sigmoid loss for Equation 1 brings us to the Skip Gram model and in that case, ℓ_p regularization is not stated. Without such regularization, unbounded optimization of $2Vd$ parameters has potential to converge to solutions that does not generalize well.

To avoid this overfitting, in our formulation we choose g_1 as follows:

$$g_1 = \sum_{v}^{V} \lambda_1 \left(||\Psi_v||_2^2 + ||\Phi_v||_2^2\right) \qquad (4)$$

where Ψ_v is the row vector of words.

Moreover, an appropriate regularization can also penalize the deviance between low rank matrices Ψ and Φ. Although there are words in the language that may have different context and target representations, such as *the* [1], it is natural to expect that a large proportion of the words have a shared representation in their context and target mappings. To this end, we introduce the following regularization:

$$g_2 = \lambda_2 ||\Psi - \Phi||_F^2 \qquad (5)$$

where F is the Frobenius matrix norm. This assumption reduces learning complexity significantly while a good representation is still retained, optimization under this constraint for large vocabularies is going to be much easier because we limit the degrees of freedom.

The Robust Gram objective therefore becomes:

$$LL + \lambda_1 \sum_{v}^{V} \left(||\Psi_v||_2^2 + ||\Phi_v||_2^2\right) + \lambda_2 ||\Psi - \Phi||_F^2 \qquad (6)$$

[1] Consider prediction of *Suleiman* from *the*, and *the* from *oasis*. We expect *the* to have different vectorial representations.

where $LL = \sum_i^N \mathcal{L}(p(y_i|x_i, \Psi, \Phi))$ is the data log-likelihood, $p(y_i|x_i, \Psi, \Phi)$ is the loglinear prediction model, and $\mathcal{L}$ the cross entropy loss. Since we are in the pursuit of preserving/restoring low masses in $\hat{C}$, norms such as ℓ_2 allow each element to still possess a small probability mass and encourage smoothness in the factorized $\Psi^T\Phi$ matrix. As $\mathcal{L}$ is picked as the cross entropy, Robust Gram can be interpreted as a more principled and robust counterpart of Skip Gram objective.

One may ask what particular factorization Equation 6 induces. The objective searches for Ψ, Φ matrices that have similar eigenvectors in the vector space. A spectral PCA embedding obtains an asymmetric decomposition $W = U\Sigma V^T$ with $\Psi = U$ and $\Phi = \Sigma V$, albeit a convincing reason for embedding matrices to be orthonormal lacks. In the Skip Gram model, this decomposition is more symmetric since neither Ψ nor Φ are orthonormal and diagonal weights are distributed across the factorized embeddings. A symmetric factorization would be: $\Psi = U\Sigma^{0.5}$,$\Phi = \Sigma^{0.5}V^T$ as in (Levy and Goldberg, 2014). The objective in Eq. 6 converges to a more symmetric decomposition since $||\Psi - \Phi||$ is penalized. Still some eigenvectors across context and target maps are allowed to differ if they pay the cost. In this sense our work is related to power SVD approaches (Caron, 2000) in which one searches an a to minimize $||W - U\Sigma^a\Sigma^{1-a}V^T||$. In our formulation, if we enforce a solution by applying a strong constraint on $||\Psi - \Phi||_F^2$, then our objective will gradually converge to a symmetric powered decomposition such that $U \approx V$.

3 Experiments

The experiments are performed on a subset of the Wikipedia corpus containing approximately 15M words. For a systematic comparison, we use the same symmetric window size adopted in (Pennington et al., 2014), 10. Stochastic gradient learning rate is set to 0.05. Embedding dimensionality is set to 100 for model selection and sensitivity analysis. Unless otherwise is stated, we discard the most frequent 20 hub words to yield a final vocabulary of 26k words. To understand the relative merit of our approach [2] , Skip Gram model is picked as the baseline. To retain the learning speed, and avoid inctractability of maximum likelihood learning, we learn our embeddings with Noise Contrastive Estimation using a negative sample (Gutmann and Hyvärinen, 2012).

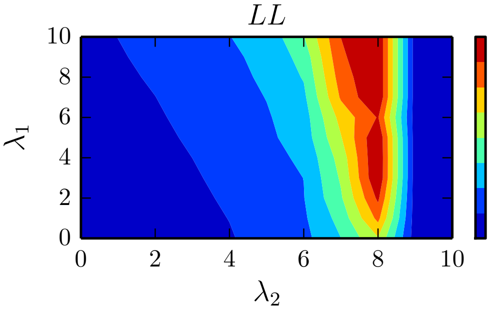

Figure 1: The LL objective for varying λ_1, λ_2.

3.1 Model Selection

For model selection, we are going to illustrate the log likelihood of different model instances. However, exact computation of the LL is computationally difficult since a full pass over the validation likelihood is time-consuming with millions of samples. Hence, we compute a stochastic likelihood with a few approximation steps. We first subsample a million samples rather than a full evaluation set, and then sample few words to predict in the window context similar to the approach followed in (Levy and Goldberg, 2014). Lastly, we approximate the normalization factor with one negative sample for each prediction score (Mnih and Kavukcuoglu, 2013)(Gutmann and Hyvärinen, 2012). Such an approximation works fine and gives smooth error curves. The reported likelihoods are computed by averaging over 5-fold cross validation sets.

Results. Figure 1 shows the likelihood LL obtained by varying $\{\lambda_1, \lambda_2\}$. The plot shows that there exits a unique minimum and both constraints contribute to achieve a better likelihood compared to their unregularized counterparts (for which $\lambda_1 = \lambda_2 = 0$). In particular, the regularization imposed by the differential of context and target embeddings g_2 contributes more than the regularization on the em-

[2] Our implementation can be downloaded from github.com/taygunk/robust_gram_embeddings

beddings Ψ and Φ separately. This is to be expected as g_2 also incorporates an amount of norm bound on the vectors. The region where both constraints are employed gives the best results. Observe that we can simply enhance the effect of g_2 by adding a small amount of bounded norm g_1 constraint in a stable manner. Doing this with pure g_2 is risky because it is much more sensitive to the selection of λ_2. These results suggest that the convex combination of stable nature of g_1 with potent regularizer of g_2, finally yields comparably better regularization.

3.2 Sensitivity Analysis

In order to test the sensitivity of our model and baseline Skip Gram to variations in the training set, we perform two sensitivity analyses. First, we simulate a missing data effect by randomly dropping out $\gamma \in [0, 20]$ percent of the training set. Under such a setting, robust models are expected to be effected less from the inherent variation. As an addition, we inspect the effect of varying the minimum cut-off parameter to measure the sensitivity. In this experiment, from a classification problem perspective, each instance is a sub-task with different number of classes (words) to predict. Instances with small cut-off introduces classification tasks with very few training samples. This cut-off choice varies in different studies (Pennington et al., 2014; Mikolov et al., 2013b), and it is usually chosen based on heuristic and storage considerations.

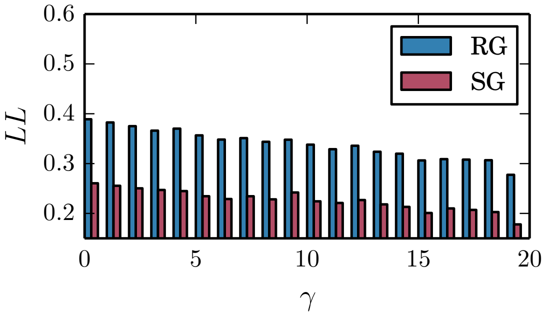

Figure 2: Training dropouts effect on LL.

Results. Figure 2 illustrates the likelihood of the Robust and Skip Gram model by varying the dropout ratio on the training set. As the training set shrinks, both models get lower LL. Nevertheless, likelihood decay of Skip Gram is relatively faster. When 20% drop is applied, the LL drops to 74% in the SG model. On the other hand the RG model not only starts with a much higher LL, the drop is also to 75.5%, suggesting that RG objective is more resistant to random variations in the training data.

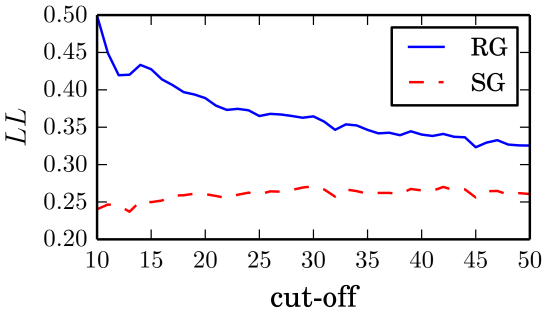

Figure 3: LL w.r.t the cut-off parameter.

Figure 3 shows the results of varying the rare-words cut-off threshold. We observe that the likelihood obtained by the Skip Gram is consistently lower than that of the Robust Gram. The graph shows that throwing out these rare words helps the objective of SG slightly. But for the Robust Gram removing the rare words actually means a significant decrease in useful information, and the performance starts to degrade towards the SG performance. RG avoids the overfitting occurring in SG, but still extracts useful information to improve the generalization.

3.3 Word Similarity Performance

The work of (Schnabel et al., 2015) demonstrates that intrinsic tasks are a better proxy for measuring the generic quality than extrinsic evaluations. Motivated by this observation, we follow the experimental setup of (Schnabel et al., 2015; Agirre et al., 2009), and compare word correlation estimates of each model to human estimated similarities with Spearman's correlation coefficient. The evaluation is performed on several publicly available word similarity datasets having different sizes. For datasets having multiple subjects annotating the word similarity, we compute the average similarity score from all subjects.

We compare our approach to set of techniques on the horizon of spectral to window based approaches. A fully spectral approach, HPCA, (Lebret and Lebret, 2013) extracts word embeddings by running a Hellinger PCA on the cooccurrence matrix. For this method, context vocabulary upper and lower bound parameters are set to $\{1, 10^{-5}\}$, as promoted by its author. GLoVe (Pennington et al., 2014) approach formulates a weighted least squares problem to combine global statistics of cooccurence and efficiency of window-based approaches. Its objective can be interpreted as an alternative to the cross-entropy loss of Robust Gram. The x_{max}, α values of the GLoVe objective is by default set to $100, 3/4$. Finally, we also compare to shallow representation learning networks such as Skip Gram and Continuous Bag of Words (CBoW) (Mikolov et al., 2013a), competitive state of the art window based baselines.

We set equal window size for all these models, and iterate three epochs over the training set. To yield more generality, all results obtained with 300 dimensional embeddings and subsampling parameters are set to 0. For Robust Gram approach, we have set $\lambda_1, \lambda_2 = \{0.3, 0.3\}$. To obtain the similarity results, we use the final Φ context embeddings.

Results. Table 1 depicts the results. The first observation is that in this setting, obtaining word similarity using HPCA and GLoVe methods are suboptimal. Frankly, we can conjecture that this scarce data regime is not in the favor of the spectral methods such as HPCA. Its poor performance can be attributed to its pure geometric reconstruction formulation, which runs into difficulties by the amount of inherent noise. Compared to these, CBoW's performance is moderate except in the RW dataset where it performs the worst. Secondly, the performance of the SG is relatively better compared to these approaches. Surprisingly, under this small data setting, RG outperforms all of its competitors in all datasets except for RG65, a tiny dataset of 63 words containing very common words. It is admissible that RG sacrifices a bit in order to generalize to a large variety of words. Note that it especially wins by a margin in MEN and Rare Words (RW) datasets, having the largest number of similarity query samples. As the number of query samples increases, RG embeddings' similarity modeling accuracy becomes clearly perceptible. The promising result Robust Gram achieves in RW dataset also sheds light on why CBoW performed worst on RW: CBOW overfits rapidly confirming the recent studies on the stability of CBoW (Luo et al., 2014). Finally, these word similarity results suggest that RG embeddings can yield much more generality under data scarcity.

	RG65	WS	WSS	WSR	MEN	RW
Size	63	353	203	252	3000	2034
CBoW	48.5	59.7	71.8	61.3	56.5	26.4
GloVe	48.9	56.2	61.5	59.1	53.0	30.0
SG	**59.2**	**71.7**	74.6	66.5	64.7	33.5
HPCA	32.1	48.6	52.9	51.5	49.9	30.7
RG	59.0	**71.7**	**74.8**	**66.7**	**65.8**	**34.0**

Table 1: Spearman's ρ coefficient. Higher is better.

4 Conclusion

This paper presents a regularized word embedding approach, called Robust Gram. In this approach, the model complexity is penalized by suppressing deviations between the embedding spaces of the target and context words. Various experimental results show that RG maintains a robust behaviour under small sample size situations, sample perturbations and it reaches a higher word similarity performance compared to its competitors. The gain from Robust Gram increases notably as diverse test sets are used to measure the word similarity performance.

In future work, by taking advantage of the promising results of Robust Gram, we intend to explore the model's behaviour in various settings. In particular, we plan to model various corpora, i.e. predictive modeling of sequentially arriving network packages. Another future direction might be encoding available domain knowledge by additional regularization terms, for instance, knowledge on synonyms can be used to reduce the degrees of freedom of the optimization. We also plan to enhance the underlying optimization by designing Elastic constraints (Zou and Hastie, 2005) specialized for word embeddings.

Acknowledgments

The authors acknowledge funding by the Dutch Organization for Scientific Research (NWO; grant 612.001.301). We also would like to thank Hamdi Dibeklioglu and Mustafa Unel for their kind support during this work.

References

Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, and Aitor Soroa. 2009. A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, pages 19–27.

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2015. Random walks on context spaces: Towards an explanation of the mysteries of semantic word embeddings. *CoRR*, abs/1502.03520.

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2016. Linear algebraic structure of word senses, with applications to polysemy. *CoRR*, abs/1601.03764.

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*, pages 238–247, June.

John Caron. 2000. Experiments with lsa scoring: Optimal rank and basis. In *Proc. of SIAM Computational Information Retrieval Workshop*.

Maryam Fazel, Haitham Hindi, and Stephen P. Boyd. 2001. A rank minimization heuristic with application to minimum order system approximation. In *In Proceedings of the 2001 American Control Conference*, pages 4734–4739.

Debanjan Ghosh, Weiwei Guo, and Smaranda Muresan. 2015. Sarcastic or not: Word embeddings to predict the literal or sarcastic meaning of words. In *EMNLP*, pages 1003–1012. The Association for Computational Linguistics.

Amir Globerson, Gal Chechik, Fernando Pereira, and Naftali Tishby. 2007. Euclidean embedding of co-occurrence data. *J. Mach. Learn. Res.*, 8:2265–2295.

Michael U. Gutmann and Aapo Hyvärinen. 2012. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *J. Mach. Learn. Res.*, 13(1):307–361, February.

Rémi Lebret and Ronan Lebret. 2013. Word emdeddings through hellinger PCA. *CoRR*, abs/1312.5542.

Benot Lemaire and Guy Denhire. 2008. Effects of high-order co-occurrences on word semantic similarities. *CoRR*, abs/0804.0143.

Omer Levy and Yoav Goldberg. 2014. Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems 27*, pages 2177–2185.

Qun Luo, Weiran Xu, and Jun Guo. 2014. A study on the cbow model's overfitting and stability. In *Proceedings of the 5th International Workshop on Web-scale Knowledge Representation Retrieval &#38; Reasoning*, Web-KR '14, pages 9–12. ACM.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 142–150.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546.

Andriy Mnih and Koray Kavukcuoglu. 2013. Learning word embeddings efficiently with noise-contrastive estimation. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2265–2273.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar.

Tobias Schnabel, Igor Labutov, David M. Mimno, and Thorsten Joachims. 2015. Evaluation methods for unsupervised word embeddings. In *EMNLP*, pages 298–307.

Nathan Srebro and Adi Shraibman. 2005. Rank, trace-norm and max-norm. In *COLT*, volume 3559 of *Lecture Notes in Computer Science*, pages 545–560. Springer.

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *J. Artif. Int. Res.*, 37(1):141–188, January.

Hui Zou and Trevor Hastie. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320.