Optimising Word Embeddings With Search-Based Approaches

Max Hort
University College London
London, United Kingdom
max.hort.19@ucl.ac.uk

Federica Sarro
University College London
London, United Kingdom
f.sarro@ucl.ac.uk

ABSTRACT

Word embeddings have rapidly become an all-purpose tool for a diverse range of real world applications. This development is nurtured by the availability and applicability of pre-trained models. However, their usage faces the risk of being inaccurate when used in domains different from the ones they were trained on.

In this paper, we formulate the adaptation of word embeddings as a vector multiplication problem, which enables us to apply search methods to explore potential word embedding adaptations with respect to their semantic correctness. To assess the effectiveness of our proposal, we empirically investigate the use of both local and global search-based approaches (i.e. Hill Climbing, Tabu Search and Genetic Algorithm) in order to maximise the semantic correctness of a popular Word2Vec pre-trained model (namely GoogleNews) when applied to another domain (i.e. the MEN dataset).

The results of our study reveal that Hill Climbing, Tabu Search and Genetic Algorithm perform equally well and all outperform the original GoogleNews model as well as a baseline model based on Random Search. This shows that optimising word embeddings with search-based approaches is possible and effective.

CCS CONCEPTS

• **Theory of computation → Optimization with randomized search heuristics**; • **Software and its engineering → Search-based software engineering**;

KEYWORDS

Word Embedding, Semantic Similarity, Optimization

ACM Reference Format:
Max Hort and Federica Sarro. 2020. Optimising Word Embeddings With Search-Based Approaches. In *Proceedings of Genetic and Evolutionary Computation Conference Companion (GECCO '20 Companion).* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Word embeddings have rapidly become an all-purpose tool for real world tasks. These models are trained on a vast amount of data and perform well for multiple natural language tasks [7]. However, this training process requires large text corpora and extensive training time [8]. Therefore, not every user is able to perform such a training process, either due to lack of time [4], computational capabilities or an insufficient amount of training data. In this case, users can resort to pre-trained models that are made publicly available, such as Word2Vec [5]. While pre-trained models can be effective, they are likely to be trained on different data than the problem domain under investigation therefore affecting their accuracy.

Different techniques for adapting word embeddings [2, 4, 6] have been proposed, however local and global search techniques have not been extensively analyzed.

We therefore investigate if search-based approaches can be effective in improving the semantic correctness of pre-trained word embedding models during adaptation. To this end, we design and experiment with three different local and global search techniques to optimize for the semantic correctness of a popular Word2Vec pre-trained word embedding model trained on news articles (namely GoogleNews [5]). We benchmark the search-based approaches against the original model and a Random Search approach. Our findings show that all the search-based approaches we investigate are able to significantly improve the semantic correctness of the original GoogleNews model.

2 PROBLEM FORMULATION

In the following, we explain the formulation of word embedding adaptation as a searchable problem.

Representation. Each solution $\overrightarrow{s}$ represents a vector to recompute and adapt a collection of word vectors. Vector multiplication is used to modify word vectors. A word vector $\overrightarrow{w}$ is multiplied by a solution vector $\overrightarrow{s}$, of the same size, element by element. The result is a modified word vector $\overrightarrow{w'} = \overrightarrow{w} \circ \overrightarrow{s}$. This procedure is applied to every word vector of an embedding model.

Initialization and Neighbor Creation. A candidate solution $\overrightarrow{s}$ is initialized by adding a small uniform noise vector $\overrightarrow{noise}$ to a vector of all ones $\overrightarrow{1}$:

$$\overrightarrow{1} + \overrightarrow{noise} = \overrightarrow{s} \tag{1}$$

As vectors have continuous values, it is not feasible to generate all possible neighboring solutions. We therefore consider stochastic methods for neighbor creation. The following neighbor creation methods are evaluated in course of the experiments:

Noise value Add a small noise value to a single element of $\overrightarrow{s}$;
Noise vector Add a small uniform noise vector, in the range $[-0.05, 0.05]$, to $\overrightarrow{s}$.

Fitness Function. Two methods exist to evaluate the semantic correctness of word embedding models: extrinsic and intrinsic [10]. While extrinsic methods evaluate the ability of word embeddings on downstream tasks, intrinsic methods compare word embeddings with judgements made by humans. In order to **evaluate the semantic correctness** of word embeddings, we use a fitness function based on the intrinsic word similarity method with Spearman's rank correlation coefficient [11] as done in previous work [3, 12].

GECCO '20 Companion, July 8–12, 2020, Cancún, Mexico
© 2020 Copyright held by the owner/author(s).
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

3 EXPERIMENTAL DESIGN

The adaptation of word embeddings can potentially have disruptive effects on the semantic correctness of the model, therefore we investigate the following research question:

RQ: Are search-based approaches able to optimize semantic correctness of pre-trained word embedding models?

To answer this question, we investigate the effectiveness of three search-based approaches (Hill Climbing, Tabu Search and Genetic Algorithm) to maximize the semantic correctness of the pre-trained GoogleNews model applied to the MEN dataset [1].[1]

These search algorithms are benchmarked against two baselines: the semantic correctness obtained by using the original pre-trained GoogleNews model and a Random Search, which is recommended as a sanity check for more sophisticated search algorithms[9]. A brief description of these algorithms and their setting is given below.

Random Search (RS) generates solutions by randomly adding noise vectors to the unit vector. We perform RS with different levels of uniform noise, ranging from $0.05 - 0.5$ with a step size of 0.05 and 10.000 repetitions. A noise level of 0.05 achieves the best performance and is used to initialize candidate solutions in subsequent experiments. Hill Climbing (HC) is a stochastic, local optimization algorithm. Based on an original solution, HC evaluates neighboring solutions and selects them if it improves the original fitness. Tabu Search (TS) is a heuristic search algorithm that can be used to augment other heuristics. In this context, TS enhances HC. Both HC and TS use *noise values* to create neighboring solutions (see Section 2). We evaluate different levels of noise, from $0.02 - 0.2$ and a step size of 0.02 for neighbor creation in HC. A noise level of 0.4 achieves the best performance and is subsequently used for neighbor creation with *noise values.* Tabu list sizes of $\{5, 10, 25, 50, 75, 100, 150\}$ are compared for TS, where a size of 150 performs best. Genetic Algorithm (GAs) is an evolutionary, global search, technique. We experimented with different GA's configurations and the following final setting was used: population size of 50, two-point crossover, 0.6 crossover probability, 0.2 mutation probability, tournament selection with a $s = 5$, *noise vectors* for mutation and no fitness re-computation of unchanged individuals.

Each experiment is limited to 10.000 fitness evaluations, repeated 100 times to account for randomness in stochastic searches, and average results are compared.

4 RESULTS

Figure 1 shows the semantic correctness of search methods on the MEN test set. The semantic correctness (i.e. Spearman's correlation) of the original GoogleNews model on the MEN test set is 0.771. This is the baseline benchmark for the search-based approaches investigated herein. We can observe that HC, TS and GA significantly outperform both the GoogleNews model and RS (all p-values are < 0.001). Furthermore, HC, TS and GA exhibit similar performances, i.e. based on the Wilcoxon signed-rank test (with $\alpha < 0.05$) one cannot reject the null hypothesis.

[1] We use version 0.2, released on the 30/04/2012 https://staff.fnwi.uva.nl/e.bruni/MEN. Since the words $\{colour, grey, harbour, theatre\}$ are not present in the GoogleNews embedding model, we change them to $\{color, gray, harbor, theater\}$, respectively, to maintain the size of the dataset.

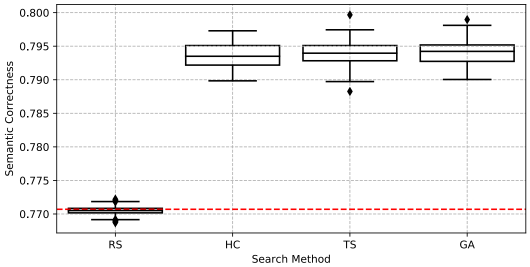

Figure 1: Boxplots of search algorithm performance on the MEN test set over 100 runs for HC and TS and 10.000 for RS. The dashed red line shows the semantic correctness of the original GoogleNews model, which is used as a baseline.

5 CONCLUSIONS AND FUTURE WORK

In this paper, we have formulated the word embedding adaptation problem as a search-based problem and evaluated the effectiveness of both local and global search techniques to optimize the semantic correctness of a popular pre-trained word embedding model (i.e. the GoogleNews Word2Vec). To the best of our knowledge, our work is the first to investigate search-based approaches aiming to improve semantic correctness of word embeddings. We found that single objective search techniques, local and global, are able to improve the semantic correctness of word embedding models. Future work will investigate other datasets and pre-trained models as well as the use of other search-based approaches and their application to downstream tasks.

ACKNOWLEDGMENTS

M. Hort and F. Sarro are supported by the ERC grant 741278 (EPIC).

REFERENCES

[1] E. Bruni, N.-K. Tran, and M. Baroni. 2014. Multimodal distributional semantics. *Journal of Artificial Intelligence Research* 49 (2014), 1–47.

[2] M. Faruqui, J. Dodge, S.K. Jauhar, C. Dyer, E. Hovy, and N.A. Smith. 2015. Retrofitting Word Vectors to Semantic Lexicons. In *Procs. of HLT-NAACL.* 1606–1615.

[3] M. Faruqui and C. 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.* 19–24.

[4] I. Labutov and H. Lipson. 2013. Re-embedding words. In *Procs. of ACL.* 489–493.

[5] T. Mikolov, I. Sutskever, K. Chen, GS. Corrado, and J. Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems.* 3111–3119.

[6] N. Mrkšić, D. Ó Séaghdha, B. Thomson, M. Gašić, L. Rojas-Barahona, P-H. Su, D. Vandyke, T-H. Wen, and S. Young. 2016. Counter-fitting Word Vectors to Linguistic Constraints. In *Procs. of HLT-NAACL.*

[7] Y. Qi, D. Sachan, M. Felix, S. Padmanabhan, and G. Neubig. 2018. When and Why Are Pre-Trained Word Embeddings Useful for Neural Machine Translation?. In *Procs. of HLT-NAACL.* 529–535.

[8] SN. Rezaeinia, A. Ghodsi, and R. Rahmani. 2017. Improving the accuracy of pre-trained word embeddings for sentiment analysis. *arXiv:1711.08609* (2017).

[9] F. Sarro, F. Ferrucci, M. Harman, A. Manna, and J. Ren. 2017. Adaptive Multi-Objective Evolutionary Algorithms for Overtime Planning in Software Projects. *IEEE Transactions on Software Engineering* 43, 10 (2017), 898–917.

[10] T. Schnabel, I. Labutov, D. Mimno, and T. Joachims. 2015. Evaluation methods for unsupervised word embeddings. In *Procs. of EMNLP.* 298–307.

[11] C. Spearman. 1904. The Proof and Measurement of Association Between Two Things. *The American Journal of Psychology* 15 (jan 1904), 72–101.

[12] Y. Tsvetkov, M. Faruqui, W. Ling, G. Lample, and C. Dyer. 2015. Evaluation of word vector representations by subspace alignment. In *Procs. of EMNLP.* 2049–2054.