CONSISTENT ALIGNMENT OF WORLD EMBEDDING MODELS*

Cem Safak Sahin, Rajmonda S. Caceres, Brandon Oselio & William M. Campbell
MIT Lincoln Laboratory
244 Wood Street
Lexington, MA 02421, USA

ABSTRACT

Word embedding models offer continuous vector representations that can capture rich contextual semantics based on their word co-occurrence patterns. While these word vectors can provide very effective features used in many NLP tasks such as clustering similar words and inferring learning relationships, many challenges and open research questions remain. In this paper, we propose a solution that aligns variations of the same model (or different models) in a joint low-dimensional latent space leveraging carefully generated synthetic data points. This generative process is inspired by the observation that a variety of linguistic relationships is captured by simple linear operations in embedded space. We demonstrate that our approach can lead to substantial improvements in recovering quality embeddings of local neighborhoods aligned and fused across different input word models.

1 INTRODUCTION

Recently there has been a growing interest in continuous vector representations of linguistic entities, most often referred to as embeddings. This includes techniques that rely on matrix factorization (Levy & Goldberg (2014); Pennington et al. (2014)), as well as currently popular neural network methods (Le & Mikolov (2014); Mikolov et al. (2013a;b)). These embeddings are able to capture complex semantic patterns such as linguistic analogies and have shown remarkable performance improvements across various NLP tasks.

Nonetheless, continuous word representations generated by neural network models are not well understood and evaluations of these representations are still nascent. It is not clear what the dimensions of the word vectors represent, and as such, we are often unable to easily evaluate the quality of representation except with reference to performance of downstream tasks (e.g., clustering, domain adaptation). It is also difficult to compare word embeddings of different dimensions, and when we do this naively, we often see wildly different local properties including from models trained on the same dataset. Herein, we highlight some these fundamental limitations of word representations, in particular with respect to their ability to be embedded and aligned in a lower-dimensional space.

Manifold alignment techniques have a rich history of addressing high dimensionality challenges within the domain adaptation and transfer learning research areas (Wang & Mahadevan (2009; 2011)), but they have primarily been applied to data sources such as images, and genomic data. In Wang et al. (2016) manifold alignment techniques are used to discover *logical* relationships in supervised settings. We believe that there is a great opportunity to further leverage the same techniques in unsupervised settings. However, it is not clear if these techniques will easily translate to alignment of continuous vector spaces when labels are not available.

Our main contribution consists of an approach that overcomes some of the effects of artificial high dimensionality by leveraging synthetically generated neighboring points, or as we refer to them, *latent words*. Inspired by the surprising insight that in high dimensional space, semantically similar words relate to one another via simple linear operations (Hashimoto et al. (2016); Mikolov et al. (2013a;b)), we conjecture that unseen words and word co-occurrences in the training datasets can

*This material is based upon work supported by the Assistant Secretary of Defense for Research and Engineering under Air Force Contract No. FA8721-05-C-0002 and/or FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Assistant Secretary of Defense for Research and Engineering.

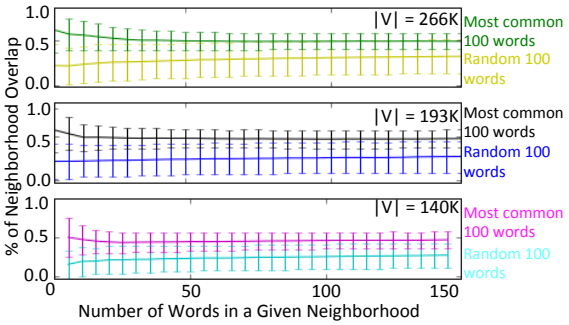

Figure 1: Quality of local neighborhood preservation as a function of undersampling in the Wikipedia dataset. We consider different subsets of the data, where we vary the size of the vocabulary (represented as $|V|$) from 140K to 266K dictionary words.

be imputed in the high dimensional space via simple local linear functions. Data imputation has been successfully used in traditional statistical analysis of incomplete datasets to improve learning and inference. The application of this concept however to improve the quality of word embedding representations is novel. As we demonstrate in the later sections, local densification of word point clouds allows us to take much better advantage of manifold embedding and alignment techniques. Additionally, we can fuse and enrich different word embedding models without the need of model retraining and access to more training data.

2 The Problem

We begin by illustrating how inherent randomness in the word embedding model as well as curse of dimensionality cause the output representation to be unstable and inconsistent. We report experimental results using the Wikipedia dataset (wiki (2016)) as presented in Fig. 1. The word embedding model is trained using the gensim library (Řehůřek & Sojka (2010))on three subsets of the Wikipedia dataset, containing $140K, 193K$ and $260K$ vocabulary words. The smaller datasets are proper subsets of the larger ones. This almost doubling of vocabulary size causes only about 24% increase of average frequencies of 100 most common words. In all cases, the words are mapped as points in $\mathcal{R}^{200}$. For a fixed local neighborhood size, we re-train the same model, using the same parameters, on the same training dataset. We then measure model stability as a function of neighborhood overlap across consecutive re-trained model instances. We measure the quality of the embedding by looking at a subsample of words in the training dataset, as well as words that fall in their local neighborhoods. We sample first naively, by picking a random subset of words, and then more strategically, by picking words that occur with high frequency, the idea being that, these words and their neighborhoods would be more accurately represented in high dimensional space.

We observe that embedding continuous word representations is highly unstable. Even though average neighborhood overlap is initially (at lower neighborhood sizes) higher, the variation is much higher as well. As we increase the size of the neighborhood, or improve the quality of our sample by only picking the most frequent words, we do observe a reduction in embedding instability. However, the amount of improvement that we get is very small relative to the increase in sample size and sample quality. Finally, even after the embedding quality stabilizes as we consider higher neighborhood sizes, it becomes unlikely to change at relatively low values of neighborhood overlap, an indication that continuous word representations are difficult to embed consistently. Next, we present an approach to addressing the adverse effects of sparsely distributed word representations.

3 Stabilizing Embeddings via Latent Words

Assume that we are given word embedding models, $W^i|_{i\in[1,k]}$ where k is the total number of models. V^i denotes a vocabulary of W^i. A word l in V^i is shown as w^i_l such that $W^i(w^i_l) \rightarrow \mathbb{R}^m$, and m is the size of the latent space. Let $d(.,.)$ represent the similarity function between two vector representations and we use the cosine similarity measure defined by $d(w^i_l, w^i_m) = \frac{\vec{w}^i_l \cdot \vec{w}^i_m}{||w^i_l|||w^i_m|}$. Since, input co-occurrence frequencies are normalized, we expect similar results if we used Euclidean distance instead. Let $n^i_\epsilon|_{w^i_l}$ be a set of words that fall in the ϵ-neighborhood of a word w^i_l, such that $n^i_\epsilon|_{w^i_l} = \{w^i_m | d(w^i_m, w^i_l) < \epsilon, w^i_l, w^i_m \in V^i\}$. Our goal is to find a lower-dimensional joint sub-

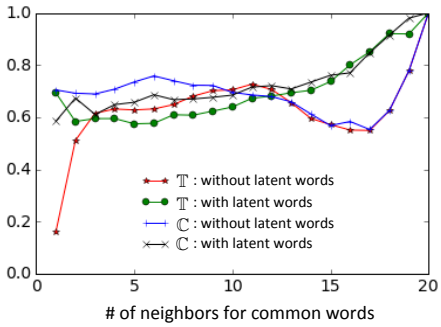

Figure 2: Trustworthiness ($\mathbb{T}$) and Continuity ($\mathbb{C}$) for common words in $n^i_\epsilon|_{w^i_l}$ and $n^j_\epsilon|_{w^j_l}$ after finding a joint union manifold and then mapping to R^{50} (larger numbers are better).

space of these models by using common words and their neighborhoods for given models since these neighborhoods of common words are composed of polysemic and semantically drifted words.

We first assume that the word embedding representation lies on an underlying manifold and that this manifold is locally continuous, linear and smooth. We then leverage the property of continuous word models to express linguistic relationships via simple linear operations. As illustrated in Mikolov et al. (2013b), if x, y and z are vector representations of *king*, *woman*, and *man*, respectively, then *queen* can be extrapolated by a simple linear combination of these vectors $x + y - z$. In addition to analogies, other linguistic inductive tasks such as *series completion* and *classification* can also be solved with vector operations on word embeddings Hashimoto et al. (2016); Lee (2015). Inspired by these insights, we surmise that words can be added and/or subtracted to recover unobserved analogies, series completions and classifications, and as a result, to recover unobserved words. Even though, we don't have the exact generative model for semantically similar unobserved words, we assume they occur nearby the observed words and this gives us a mechanism for densifying the original local neighborhoods. Note that for the purposes of our learning objective, we don't need to make the case that the latent words are real words; we only need to ensure that they are geometrically close to where we would expect the real words to be placed. We ensure this property by construction: a latent word is generated by linear vector operations of original words within an ϵ-neighborhood and is only included if it falls within this neighborhood. A latent word w^i_* can be generated by $w^i_* = \sum(\alpha_n \times w^i_{r_n})$, where α_n is a randomly chosen integer from $[-1, +1]$ and $w^i_{r_n} \in n^i_\epsilon|_{w^i_l}$. Note that w^i_* is a valid latent word if and only if $d(w^i_l, w^i_*) < \epsilon$. These latent points are only leveraged as anchor points to improve dimensionality reduction and alignment of the original embedding models.

Following, we present preliminary results for aligning a neighborhood of two models, W^i and W^j, generated by the same input data and training parameters explained in Section 2 for $|V| = 266K$. Since, we are performing manifold alignment, we illustrate via a successful alignment technique, the low rank alignment LRA given in Boucher et al. (2015), which is an extension of LLE. We also consider two relevant metrics to help us evaluate the quality of lower-dimensional embeddings: (i) trustworthiness ($\mathbb{T}$) and (ii) continuity ($\mathbb{C}$) van der Maaten et al. (2008). We could think of $\mathbb{T}$ and $\mathbb{C}$ as measuring components of the symmetric difference between a neighborhood in high and low dimensional space. In that sense, they emphasize inconsistencies with preserving the neighborhood structure in embedded space (unlike the neighborhood overlap which emphasizes the portion of the structure which is preserved). $\mathbb{T}$ and $\mathbb{C}$ values are analyzed for $n^i_\epsilon|_{w^i_l} \cap n^j_\epsilon|_{w^j_l}$ as presented in Fig. 2. As seen in this figure, these metrics are slightly better for our approach for up to 12 neighbors. Beyond that point, the addition of latent words causes considerable improvement in the alignment performance. In the figure, we have illustrated alignment of W^i and W^j using a most frequent common word, but we have observed similar or worst behavior when we pick less frequent vocabulary words or fewer common words across different models. The same is true when it comes to the improvements we can achieve via the injection of latent words.

In conclusion, our tailored manifold alignment approach offers a platform for fusing different word embedding models and generating richer semantic representations. Furthermore, a common representation of different models allows for explicit comparison and evaluation of the quality of the representation. In this paper, we only provide empirical results for alignment of a selected subset of local neighborhoods. As a future work, we plan (i) to extend our approach to generate a holistic embedding model that optimizes alignment across all local neighborhoods, (ii) to add more constraints into LRA process such that weights of latent words in the similarity matrix will be a part of loss function to minimize and (iii) to characterize the effects of anchor points in different models.

REFERENCES

Thomas Boucher, CJ Carey, Sridhar Mahadevan, and M. Darby Dyar. Aligning mixed manifolds. In *Proc. of 29th AAAI Conference on AI*, pp. 2511–2517, 2015. ISBN 0-262-51129-0.

Tatsunori Hashimoto, David Alvarez-Melis, and Tommi Jaakkola. Word embeddings as metric recovery in semantic spaces. *Trans. of the Assoc. for Comp. Linguistics*, 4:273–286, 2016. ISSN 2307-387X.

Quoc V. Le and Tomas Mikolov. Distributed representations of sentences and documents. *CoRR*, abs/1405.4053, 2014.

Lisa Seung-Yeon Lee. On the linear algebraic structure of distributed word representations. *CoRR*, abs/1511.06961, 2015.

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

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

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, pp. 3111–3119, 2013b.

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

Radim Řehůřek and Petr Sojka. Software Framework for Topic Modeling with Large Corpora. In *Proc. of LREC 2010 Workshop on New Challenges for NLP Frameworks*, pp. 45–50, 2010.

L.J.P. van der Maaten, E. O. Postma, and H. J. van den Herik. Dimensionality reduction: A comparative review, 2008.

Chang Wang and Sridhar Mahadevan. Manifold alignment without correspondence. In *Proc of 21st Inter. Joint Conf. on Artificial Intelligence*, pp. 1273–1278, CA, USA, 2009.

Chang Wang and Sridhar Mahadevan. Heterogeneous domain adaptation using manifold alignment. In *Proc. of 22nd Inter. Joint Conf. on Artificial Intelligence - Vol. 2*, pp. 1541–1546, 2011.

Chang Wang, Liangliang Cao, and James Fan. Building joint spaces for relation extraction. *Relation*, 1:16, 2016.

wiki. Latest wikipedia dump. `http://dumps.wikimedia.org/enwiki/latest/enwiki-latest-pages-articles.xml.bz2`, 2016.