Identifiability and Invariance Issues for Word Embeddings

Rachel Carrington, Karthik Bharath, Simon Preston

University of Nottingham

December 2019

Summary

- Word embeddings obtained as optimizers of objective functions in which the word and context matrices (U and V respectively) appear only through their product (UV) are not unique. (E.g. LSA, word2vec, Glove)
- The multiple solutions can perform differently on test data.
- Disparity in test-data performance between word embeddings can sometimes be due to different solutions being selected.
- We propose two ways of addressing this non-identifiability:
  - Imposing constraints on optimisation to ensure uniqueness of word embedding solution.
  - Optimizing test-data performance over the solution set.

Introduction

Notation: Let X be a representation of the data, V the matrix of word embeddings, U an auxiliary matrix, and D a test set.

Most word embedding models (e.g. GloVe, word2vec, LSA) can be written as an optimization

$$\min_{V} f(X, UV)$$

e.g., in LSA $f(X, UV) = ||X - UV||_F$.

Word embeddings are assessed by a function $g(D, V)$. Usually g is based on cosine similarity between columns of V. For example, g can be taken as the correlation coefficient between cosine similarities between word pairs and human-assigned similarity scores in D.

Non-identifiability

We can replace

$$(U, V) \to (UC^{-1}, CV)$$

where C is an invertible $d \times d$ matrix, without changing the value of f.
However, this will change the value of g, unless C is an orthogonal matrix.

We want to find the set of transformations to which f is invariant, but not g.

Group Theory

What is the set of transformations which leave f invariant but not g? It is helpful to use some group theory here to formalise this.

f is invariant to transformation of the word embeddings by $\mathrm{GL}(d)$, the set of invertible $d \times d$ matrices.

g is invariant to transformations by the set $c\mathrm{O}(d) = \{cQ \in \mathrm{GL}(d) : c \in \mathbb{R}, Q \in \mathrm{O}(d)\}$.

Let $\mathcal{F}_d$ be the set of transformations to which f is invariant, but g is not. Then $\mathcal{F}_d = \tilde{\mathcal{F}}_d - c\mathcal{I}$, where $\tilde{\mathcal{F}}_d = \mathrm{GL}(d) \setminus \mathrm{O}(d)$. The set $\tilde{\mathcal{F}}_d$ can be identified with the set $\mathrm{UT}(d)$ of upper triangular $d \times d$ matrices.

Solution 1: Imposing constraints

We can redefine f as a constrained optimization

$$\underset{U,V:V\in\mathfrak{C}_v}{\arg\min}\; f(X, UV)$$

where

$$\mathfrak{C}_v = \{V \in \mathbb{R}^{d\times p} : VV^T = I_d\}$$

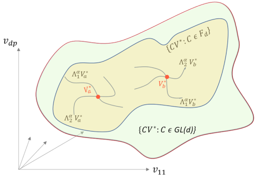
Claim: Any solution which satisfies this constrained optimization problem will be related by an orthogonal transformation, so all solutions will give the same value for g.

By also imposing constraints on U we can get a unique solution.

Solution 2: Optimization over $\mathcal{F}_d$

Alternatively, we can optimize over the solution set of f to get embeddings which perform best with respect to g.

One-dimensional optimization[1]: For SVD embeddings, we approximate the matrix X by $X \approx A_d \Sigma_d B_d^T$. We can then choose to take $V = \Sigma_d^\alpha B_d^T$, where $\alpha \in \mathbb{R}$.



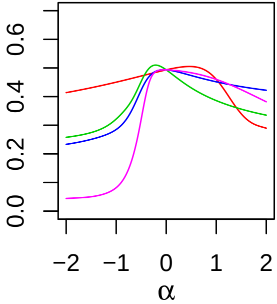


Figure: The graph shows the test scores for $\Lambda^\alpha V^*$, where $V^* = B_d^T$, for different choices of diagonal Λ. B_d^T was found using the SVD of the document-term matrix of the Corpus of Historical American English (COHA), with $d = 300$. The red line is $\Lambda = \Sigma_d$. This does not seem to perform significantly better than the other choices of Λ, so there doesn't seem to be any reason to restrict the optimization to this particular subset.

[1] [Bullinaria and Levy, 2012], [Turney, 2013]

Solution 2: Optimization over $\mathcal{F}_d$

Optimization over $\text{UT}(d)$**:**

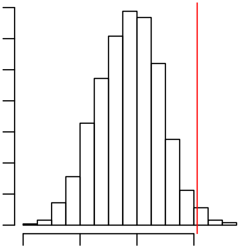


Figure: Histogram of performance of CV^*, where V^* is a word2vec embedding and C is a random element of $\text{UT}(d)$. The red line shows the performance of the base embedding.

Embedding	Spearman	Pearson
V^*_{word2vec}	0.700	0.652
Optimized V^*_{word2vec}	0.797	0.838
V^*_{GloVe}	0.601	0.603
Optimized V^*_{GloVe}	0.679	0.760

Table: Test scores for word2vec and GloVe embeddings on the WordSim-353 test set. The optimization is over ΛV^* where Λ is diagonal. In all cases performance can be significantly improved by optimization.

Conclusions

Our main findings are summarized follows:

- For many word embedding methods, the objective function does not have a unique optimum. However, different solutions can perform differently on test data.
- This means that the disparity in performance of different embedding sets, for example those tuned using hyperparameters, may be due to different elements of the solution set being selected.
- One way to deal with non-identifiability is to impose constraints on the solution via constrained optimization.
- Alternatively, we can try to optimize performance of the embeddings over the solution set of the objective function. In some cases performance can be significantly improved by selecting a different solution than that selected by the embedding algorithm.

References

Bullinaria, J. A. and Levy, J. P. (2012).
Extracting semantic representations from word co-occurrence statistics: stop-lists, stemming, and svd.
Behavior research methods, 44(3):890–907.

Turney, P. D. (2013).
Distributional semantics beyond words: Supervised learning of analogy and paraphrase.
Transactions of the Association for Computational Linguistics, 1:353–366.