Frustratingly Easy Meta-Embedding – Computing Meta-Embeddings by Averaging Source Word Embeddings

Joshua N Coates
Department of Computer Science
University of Liverpool
`j.n.coates@liverpool.ac.uk`

Danushka Bollegala
Department of Computer Science
University of Liverpool
`danushka@liverpool.ac.uk`

Abstract

Creating accurate meta-embeddings from pre-trained source embeddings has received attention lately. Methods based on global and locally-linear transformation and concatenation have shown to produce accurate meta-embeddings. In this paper, we show that the arithmetic mean of two distinct word embedding sets yields a performant meta-embedding that is comparable or better than more complex meta-embedding learning methods. The result seems counter-intuitive given that vector spaces in different source embeddings are not comparable and cannot be simply averaged. We give insight into why averaging can still produce accurate meta-embedding despite the incomparability of the source vector spaces.

1 Introduction

Distributed vector representations of words, henceforth referred to as word embeddings, have been shown to exhibit strong performance on a variety of NLP tasks (Turian et al., 2010; Zou et al., 2013). Methods for producing word embedding sets exploit the distributional hypothesis to infer semantic similarity between words within large bodies of text, in the process they have been found to additionally capture more complex linguistic regularities, such as analogical relationships (Mikolov et al., 2013c). A variety of methods now exist for the production of word embeddings (Collobert and Weston, 2008; Mnih and Hinton, 2009; Huang et al., 2012; Pennington et al., 2014; Mikolov et al., 2013a). Comparative work has illustrated a variation in performance between methods across evaluative tasks (Chen et al., 2013; Yin and Schütze, 2016).

Methods of "meta-embedding", as first proposed by Yin and Schütze (2016), aim to conduct a complementary combination of information from an ensemble of distinct word embedding sets, each trained using different methods, and resources, to yield an embedding set with improved overall quality.

Several such methods have been proposed. 1TON (Yin and Schütze, 2016), takes an ensemble of K pre-trained word embedding sets, and employs a linear neural network to learn a set of meta-embeddings along with K global projection matrices, such that through projection, for every word in the meta-embedding set, we can recover its corresponding vector within each source word embedding set. 1TON+ (Yin and Schütze, 2016), extends this method by predicting embeddings for words not present within the intersection of the source word embedding sets. An unsupervised locally linear meta-embedding approach has since been taken (Bollegala et al., 2017), for each source embedding set, for each word; a representation as a linear combination of its nearest neighbours is learnt. The local reconstructions within each source embedding set are then projected to a common meta-embedding space.

The simplest approach considered to date, has been to concatenate the word embeddings across the source sets (Yin and Schütze, 2016). Despite its simplicity, concatenation has been used to provide a good baseline of performance for meta-embedding.

A method which has not yet been proposed is to conduct a direct averaging of embeddings. The validity of this approach may perhaps not seem obvious, owing to the fact that no correspondence exists between the dimensions of separately trained word embedding sets. In this paper we first provide some analysis and justification that, despite this dimensional disparity, averaging can provide an approximation of the performance of concatenation without increasing the dimension of the embeddings. We give empirical results demonstrating the quality of average meta-embeddings. We

make a point of comparison to concatenation since it is the most comparable in terms of simplicity, whilst also providing a good baseline of performance on evaluative tasks. Our aim is to highlight the validity of averaging across distinct word embedding sets, such that it may be considered as a tool in future meta-embedding endeavours.

2 Analysis

To evaluate semantic similarity between word embeddings we consider the Euclidean distance measure. For ℓ_2 normalised word embeddings, Euclidean distance is a monotonically decreasing function of the cosine similarity, which is a popular choice in NLP tasks that use word embeddings such as semantic similarity prediction and analogy detection (Levy et al., 2015; Levy and Goldberg, 2014). We defer the analysis of other types of distance measures to future work. By evaluating the relationship between the Euclidean distances of pairs of words in the source embedding sets and their corresponding Euclidean distances in the meta-embedding space we can obtain a view as to how the meta-embedding procedure is combining semantic information. We begin by examining concatenation through this lens, before moving on to averaging.

2.1 Concatenation

We can express concatenation by first zero-padding our source embeddings, before combining them through addition.

Without loss of generality, we consider both concatenation and averaging over only two source word embedding sets for ease of exposition. Let $\mathcal{S}_1$ and $\mathcal{S}_2$ be unique embedding sets of real-valued continuous embeddings. We make no assumption that $\mathcal{S}_1$ and $\mathcal{S}_2$ were trained using the same method or resources. Consider two semantically similar words $\boldsymbol{u}$ and $\boldsymbol{v}$ such that $\boldsymbol{u}, \boldsymbol{v} \in \mathcal{S}_1 \cap \mathcal{S}_2$. Let $\boldsymbol{u}_{\mathcal{S}_1}$ and $\boldsymbol{v}_{\mathcal{S}_1}$, and $\boldsymbol{u}_{\mathcal{S}_2}$ and $\boldsymbol{v}_{\mathcal{S}_2}$ denote the specific word embeddings of $\boldsymbol{u}$ and $\boldsymbol{v}$ within the embeddings $\mathcal{S}_1$, and $\mathcal{S}_2$ respectively.

Let the dimensions of embeddings $\mathcal{S}_1$, and $\mathcal{S}_2$ be denoted $d_{\mathcal{S}_1}$, and $d_{\mathcal{S}_2}$ respectively. We zero-pad embeddings from $\mathcal{S}_1$ by front-loading $d_{\mathcal{S}_2}$ zero entries to each word embedding vector. In contrast, we zero-pad embeddings from $\mathcal{S}_2$ by adding $d_{\mathcal{S}_1}$ zero entries to the end of each embedding vector. The resulting embeddings from $\mathcal{S}_1$ and $\mathcal{S}_2$ now share a common dimension of $d_{\mathcal{S}_1} + d_{\mathcal{S}_2}$. Denote

the resulting embeddings of any word $\boldsymbol{u} \in \mathcal{S}_1 \cap \mathcal{S}_2$, as $\boldsymbol{u}_{\mathcal{S}_1}^{zero}$ and $\boldsymbol{u}_{\mathcal{S}_2}^{zero}$ respectively. Now, combining our source embeddings through addition we obtain equivalency to concatenation.

$$\boldsymbol{u}_{\mathcal{S}_1}^{zero} + \boldsymbol{u}_{\mathcal{S}_2}^{zero} = \begin{bmatrix} \boldsymbol{u}_{\mathcal{S}_{2(1)}} \\ \boldsymbol{u}_{\mathcal{S}_{2(2)}} \\ \vdots \\ \boldsymbol{u}_{\mathcal{S}_{2(d_{\mathcal{S}_2})}} \\ \boldsymbol{u}_{\mathcal{S}_{1(1)}} \\ \boldsymbol{u}_{\mathcal{S}_{1(2)}} \\ \vdots \\ \boldsymbol{u}_{\mathcal{S}_{1(d_{\mathcal{S}_1})}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{u}_{\mathcal{S}_2} \\ \boldsymbol{u}_{\mathcal{S}_1} \end{bmatrix} \quad (1)$$

Note that the zero-padded vectors are orthogonal.

Let the Euclidean distance between these words in each embedding be denoted by $E_{\mathcal{S}_1}$ and $E_{\mathcal{S}_2}$. Note that for any vector $\boldsymbol{u} \in \mathbb{R}^n$ the addition of zero-valued dimensions does not affect the value of its ℓ_2-norm. So we have

$$E_{\mathcal{S}_1} = ||\boldsymbol{u}_{\mathcal{S}_1} - \boldsymbol{v}_{\mathcal{S}_1}||_2 = ||\boldsymbol{u}_{\mathcal{S}_1}^{zero} - \boldsymbol{v}_{\mathcal{S}_1}^{zero}||_2 \quad (2)$$
$$E_{\mathcal{S}_2} = ||\boldsymbol{u}_{\mathcal{S}_2} - \boldsymbol{v}_{\mathcal{S}_2}||_2 = ||\boldsymbol{u}_{\mathcal{S}_2}^{zero} - \boldsymbol{v}_{\mathcal{S}_2}^{zero}||_2 \quad (3)$$

Consider the Euclidean distance between $\boldsymbol{u}$ and $\boldsymbol{v}$ after concatenation.

$$\begin{aligned} E_{CONC} \\ &= \left|\left|\begin{bmatrix} \boldsymbol{u}_{\mathcal{S}_2} \\ \boldsymbol{u}_{\mathcal{S}_1} \end{bmatrix} - \begin{bmatrix} \boldsymbol{v}_{\mathcal{S}_2} \\ \boldsymbol{v}_{\mathcal{S}_1} \end{bmatrix}\right|\right|_2 \\ &= ||(\boldsymbol{u}_{\mathcal{S}_1}^{zero} + \boldsymbol{u}_{\mathcal{S}_2}^{zero}) - (\boldsymbol{v}_{\mathcal{S}_1}^{zero} + \boldsymbol{v}_{\mathcal{S}_2}^{zero})||_2 \\ &= ||(\boldsymbol{u}_{\mathcal{S}_1}^{zero} - \boldsymbol{v}_{\mathcal{S}_1}^{zero}) - (\boldsymbol{v}_{\mathcal{S}_2}^{zero} - \boldsymbol{u}_{\mathcal{S}_2}^{zero})||_2 \\ &= \sqrt{(E_{\mathcal{S}_1})^2 + (E_{\mathcal{S}_2})^2 - 2E_{\mathcal{S}_1}E_{\mathcal{S}_2}\cos(\theta)} \\ &= \sqrt{(E_{\mathcal{S}_1})^2 + (E_{\mathcal{S}_2})^2 - 2E_{\mathcal{S}_1}E_{\mathcal{S}_2}(0)} \\ &= \sqrt{(E_{\mathcal{S}_1})^2 + (E_{\mathcal{S}_2})^2} \end{aligned}$$

For any two words belonging to the resultant embedding obtained by concatenation, the distance between these words in the resultant space is the root of the sum of squares of Euclidean distances between these words in $\mathcal{S}_1$ and $\mathcal{S}_2$.

2.2 Average word embeddings

Here we now make the assumption that $\mathcal{S}_1$ and $\mathcal{S}_2$ have common dimension d.[1]

---

[1]Without loss of generality, source embeddings with different dimensionality can be appropriately padded to have the same dimensionality.

Despite there being no obvious correspondence between dimensions of $\mathcal{S}_1$ and $\mathcal{S}_2$ we can show that the average embedding set retains semantic information through preservation of the relative distances between words.

Consider the positioning of words $\boldsymbol{u}$, and $\boldsymbol{v}$ after performing a word-wise average between the source embedding sets. The Euclidean distance between $\boldsymbol{u}$ and $\boldsymbol{v}$ in the resultant meta-embedding is given by

$$
\begin{aligned}
E_{AVG} &= \left\| \frac{(\boldsymbol{u}_{\mathcal{S}_1} + \boldsymbol{u}_{\mathcal{S}_2})}{2} - \frac{(\boldsymbol{v}_{\mathcal{S}_1} + \boldsymbol{v}_{\mathcal{S}_2})}{2} \right\|_2 \\
&= \frac{1}{2} \left\| (\boldsymbol{u}_{\mathcal{S}_1} - \boldsymbol{v}_{\mathcal{S}_1}) - (\boldsymbol{v}_{\mathcal{S}_2} - \boldsymbol{u}_{\mathcal{S}_2}) \right\|_2 \\
&\propto \sqrt{(E_{\mathcal{S}_1})^2 + (E_{\mathcal{S}_2})^2 - 2 E_{\mathcal{S}_1} E_{\mathcal{S}_2} \cos(\theta)}
\end{aligned}
$$

Now in this case, unlike concatenation, we have not designed our source embedding sets such that they are orthogonal to each other, and so it seems we are left with a term dependant on the angle between $(\boldsymbol{u}_{\mathcal{S}_1} - \boldsymbol{v}_{\mathcal{S}_1})$ and $(\boldsymbol{v}_{\mathcal{S}_2} - \boldsymbol{u}_{\mathcal{S}_2})$. However, Cai et al. (2013) showed that, if $\mathcal{X}$ is a set of random points $\in \mathbb{R}^n$ with cardinality $|\mathcal{X}|$, then the limiting distribution of angles, as $|\mathcal{X}| \to \infty$, between pairs of elements from $\mathcal{X}$, is Gaussian with mean $\pi/2$. In addition, Cai et al. (2013) showed that the variance of this distribution shrinks as the dimensionality increases.

Word embedding sets typically contain in the order of ten thousand or more points, and are typically of relatively high dimension. Moreover, assuming the difference vector between any two words in an embedding set is sufficiently random, we may approximate the limiting Gaussian distribution described by Cai et al. (2013). In such a case the expectation would then be that the vectors $(\boldsymbol{u}_{\mathcal{S}_1} - \boldsymbol{v}_{\mathcal{S}_1})$ and $(\boldsymbol{v}_{\mathcal{S}_2} - \boldsymbol{u}_{\mathcal{S}_2})$ are orthogonal, leading to the following result.

$$
\mathbb{E}[E_{AVG}] = \frac{1}{2} \sqrt{(E_{\mathcal{S}_1})^2 + (E_{\mathcal{S}_2})^2} \propto E_{CONC}
\tag{4}
$$

To summarise, if word embeddings can be shown to be approximately orthogonal, then averaging will approximate the same information as concatenation, without increasing the dimensionality of the embeddings.

3 Experiments

We first empirically test our theory that word embeddings are sufficiently random and high dimensional, such that they are approximately all orthogonal to each other. We then present an empirical evaluation of the performance of the meta-embeddings produced through averaging, and compare against concatenation.

3.1 Datasets

We use the following pre-trained embedding sets that have been used in prior work on meta-embedding learning (Yin and Schütze, 2016; Bollegala et al., 2017) for experimentation.

- **GloVe** (Pennington et al., 2014). 1,917,494 word embeddings of dimension 300.

- **CBOW** (Mikolov et al., 2013b). Phrase embeddings discarded, leaving 929,922 word embeddings of dimension 300.

- **HLBL** (Turian et al., 2010). 246,122 hierarchical log-bilinear (Mnih and Hinton, 2009) word embeddings of dimension 100.

Note that the purpose of this experiment is not to compare against previously proposed meta-embedding learning methods, but to empirically verify averaging as a meta-embedding method and validate the assumptions behind the theoretical analysis. By using three pre-trained word embeddings with different dimensionalities and empirical accuracies, we can evaluate the averaging-based meta-embeddings in a robust manner.

We pad HLBL embeddings to the rear with 200 zero-entries to bring their dimension up to 300. For GloVe, we ℓ_2 normalise each dimension of the embedding across the vocabulary, as recommended by the authors. Every individual word embedding from each embedding set is then ℓ_2-normalised. The proposed averaging operation, as well as concatenation, operate only on the intersection of these embeddings. The intersectional vocabularies GloVe $\cap$ CBOW, GloVe $\cap$ HLBL, and CBOW $\cap$ HLBL contain 154,076; 90,254; and 140,479 word embeddings respectively.

3.2 Empirical distribution analysis

We conduct an empirical analysis of the distribution of the angle $\sphericalangle[(\boldsymbol{u}_{\mathcal{S}_1} - \boldsymbol{v}_{\mathcal{S}_1}), (\boldsymbol{v}_{\mathcal{S}_2} - \boldsymbol{u}_{\mathcal{S}_2})]$ for each pair of datasets. Table 1 shows the mean and variance of these distributions, obtained
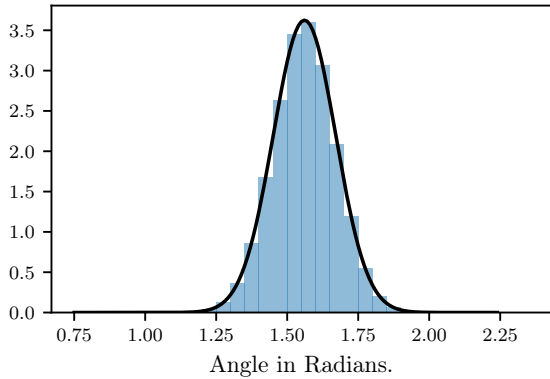
Figure 1: Distribution of angles between embeddings within GloVe ∩ CBOW.

from samples of 200,000 random pairs of words from each intersectional vocabulary. We find that the angles are approximately normally distributed around $\pi/2$.

Embeddings	μ	σ^2
GloVe & CBOW	1.5609	0.0121
GloVe & HLBL	1.5709	0.0129
CBOW & HLBL	1.5740	0.0126

Table 1: Observed distribution parameters.

Figure 1 shows a normalised histogram of the results for GloVe ∩ CBOW, along with a normal distribution characterised by the sample mean and variance. GloVe ∩ HLBL, and CBOW ∩ HLBL plots are not shown due to space limitations, but are similarly normally distributed. This result shows that the pre-trained word embeddings approximately satisfy the predictions made by Cai et al. (2013), thereby empirically justifying the assumption made in the derivation of (4).

3.3 Evaluation Tasks

3.3.1 Semantic Similarity

We measure the similarity between words by calculating the cosine similarity between their embeddings; we then calculate Spearman correlation against human similarity scores. The following datasets are used: **RG** (Rubenstein and Goodenough, 1965), **MC** (Miller and Charles, 1991), **WS** (Finkelstein et al., 2001), **RW** (Luong et al., 2013), and **SL** (Hill et al., 2015).

3.3.2 Word Analogy

Using the Google dataset **GL** (Mikolov et al., 2013b) (19544 analogy questions), we solve questions of the form *a is to b as c is to what?*, using

Embeddings	RG	MC	WS	RW	SL	GL
sources						
HLBL 100	35.3	49.3	35.7	19.1	22.1	15.0
CBOW 300	76.0	82.2	69.8	53.4	44.2	67.1
GloVe 300	82.9	87.0	75.4	48.7	45.3	68.7
AVG						
CBOW+HLBL 300	69.2	81.0	60.1	48.7	37.3	49.4
GloVe+CBOW 300	82.2	87.0	74.5	52.9	**46.5**	73.8
GloVe+HLBL 300	73.7	74.1	64.2	44.6	38.8	49.5
CONC						
CBOW+HLBL 400	68.7	80.2	62.9	49.1	39.6	53.2
GloVe+CBOW 600	**83.0**	**88.8**	**76.4**	**54.8**	46.3	**75.5**
GloVe+HLBL 400	73.7	80.1	65.5	46.4	40.0	53.8

Table 2: Results on word similarity, and analogical tasks. Best performances bolded per task. Dimensionality of the meta embedding is shown next to the source embedding names.

the CosAdd method (Mikolov et al., 2013c) shown in (5). Specifically, we determine a fourth word d such that the similarity between $(b - a + c)$ and d is maximised.

$$\mathrm{CosAdd}(a : b, c : d) = \cos(b - a + c, d) \quad (5)$$

3.4 Discussion of results

Table 2 shows task performance for each source embedding set, and for both methods on every pair of datasets. In our experiments concatenation obtains better overall performance. However, averaging offers improvements over the source embedding sets for semantic similarity task **SL** and word analogy task **GL**, on the combination of CBOW and GloVe. HLBL has a negative effect on CBOW and GloVe, but the performance of averaging is close to that of concatenation. An advantage of averaging when compared against concatenation, is that the dimensionality of the produced meta-embedding is not increased beyond the maximum dimension present within the source embeddings, resulting in a meta-embedding which is easier to process and store.

4 Conclusion

We have presented an argument for averaging as a valid meta-embedding technique, and found experimental performance to be close to, or in some cases better than that of concatenation, with the additional benefit of reduced dimensionality. We propose that when conducting meta-embedding, both concatenation and averaging should be considered as methods of combining embedding spaces, and their individual advantages considered.

References

Danushka Bollegala, Kohei Hayashi, and Ken-ichi Kawarabayashi. 2017. Think globally, embed locally—locally linear meta-embedding of words. *arXiv preprint arXiv:1709.06671* .

Tony Cai, Jianqing Fan, and Tiefeng Jiang. 2013. Distributions of angles in random packing on spheres. *The Journal of Machine Learning Research* 14(1):1837–1864.

Yanqing Chen, Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2013. The expressive power of word embeddings. *arXiv preprint arXiv:1301.3226* .

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*. ACM, pages 160–167.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on World Wide Web*. ACM, pages 406–414.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics* 41(4):665–695.

Eric H Huang, Richard Socher, Christopher D Manning, and Andrew Y Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*. Association for Computational Linguistics, pages 873–882.

Omer Levy and Yoav Goldberg. 2014. Linguistic regularities in sparse and explicit word representations. In *CoNLL*.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of Association for Computational Linguistics* 3:211–225.

Thang Luong, Richard Socher, and Christopher D Manning. 2013. Better word representations with recursive neural networks for morphology. In *CoNLL*. pages 104–113.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* .

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. pages 3111–3119.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013c. Linguistic regularities in continuous space word representations. In *Hlt-naacl*. volume 13, pages 746–751.

George A Miller and Walter G Charles. 1991. Contextual correlates of semantic similarity. *Language and cognitive processes* 6(1):1–28.

Andriy Mnih and Geoffrey E Hinton. 2009. A scalable hierarchical distributed language model. In *Advances in neural information processing systems*. pages 1081–1088.

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

Herbert Rubenstein and John B Goodenough. 1965. Contextual correlates of synonymy. *Communications of the ACM* 8(10):627–633.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*. Association for Computational Linguistics, pages 384–394.

Wenpeng Yin and Hinrich Schütze. 2016. Learning word meta-embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, pages 1351–1360.

Will Y Zou, Richard Socher, Daniel M Cer, and Christopher D Manning. 2013. Bilingual word embeddings for phrase-based machine translation. In *EMNLP*. pages 1393–1398.