Context Mover's Distance & Barycenters: Optimal Transport of Contexts for Building Representations

Sidak Pal Singh EPFL
Andreas Hug EPFL
Aymeric Dieuleveut EPFL and École Polytechnique
Martin Jaggi EPFL

Abstract

We present a framework for building unsupervised representations of entities and their compositions, where each entity is viewed as a probability distribution rather than a vector embedding. In particular, this distribution is supported over the contexts which co-occur with the entity and are embedded in a suitable low-dimensional space. This enables us to consider representation learning from the perspective of Optimal Transport and take advantage of its tools such as Wasserstein distance and barycenters. We elaborate how the method can be applied for obtaining unsupervised representations of text and illustrate the performance (quantitatively as well as qualitatively) on tasks such as measuring sentence similarity, word entailment and similarity, where we empirically observe significant gains (e.g., 4.1% relative improvement over Sent2vec, GenSen).

The key benefits of the proposed approach include: (a) capturing uncertainty and polysemy via modeling the entities as distributions, (b) utilizing the underlying geometry of the particular task (with the ground cost), (c) simultaneously providing interpretability with the notion of optimal transport between contexts and (d) easy applicability on top of existing point embedding methods. The code, as well as pre-built histograms, are available under https://github.com/context-mover/.

1 Introduction

One of the driving factors behind recent successes in machine learning has been the development of better methods for data representation. Examples include continuous vector representations for language (Mikolov et al., 2013; Pennington et al., 2014), CNN based feature representations for images and text (LeCun et al., 1998; Collobert and Weston, 2008; Kalchbrenner et al., 2014), or via the hidden state representations of LSTMs (Hochreiter and Schmidhuber, 1997; Sutskever et al., 2014). Pre-trained unsupervised representations, in particular, have been immensely useful as general purpose features for model initialization (Kim, 2014), downstream tasks (Severyn and Moschitti, 2015; Deriu et al., 2017), and in domains with limited supervised information (Qi et al., 2018). A shared theme across these methods is to map input entities to dense vector embeddings lying in a low-dimensional latent space where the semantics of inputs are preserved. Thus, each entity of interest (e.g., a word) is represented directly as a single point (i.e., its embedding vector) in space, which is typically Euclidean.

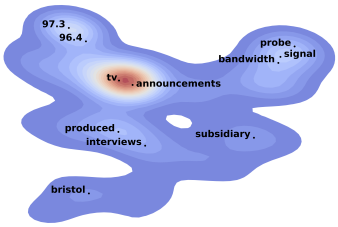


Figure 1: Distributional estimate for the entity 'radio'.

In contrast, we approach the problem of building unsupervised representations in a fundamentally different manner. We focus on the co-occurrence information between the entities and their contexts, and represent each entity as a *probability distribution (histogram) over its contexts*. Here the contexts themselves are embedded as points in a suitable low-dimensional space. This allows us to cast finding distance between entities as an instance of the *Optimal Transport problem* (Monge, 1781; Kantorovich, 1942; Villani, 2008). So, our resulting framework intuitively compares the cost of moving the contexts of a given entity to the contexts of another, which motivates the naming: *Context Mover's Distance* (CMD).

We call this distribution over contexts embeddings as

Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

the *distributional estimate* of our entity of interest (see Figure 1), while we refer to the individual embeddings of contexts as *point estimates*. More precisely, the contexts refer to any generic entities or objects (such as words, phrases, sentences, images, etc.) co-occurring with the entities to be represented.

The main motivation for our proposed approach originates from the domain of natural language, where the entities (words, phrases, or sentences) generally have different semantics depending on the context in which they occur. Hence, it is important to consider representations that effectively capture such inherent uncertainty and polysemy, and we will argue that distributional estimates capture more of this information compared to point-wise embedding vectors alone.

The co-occurrence information that is the crucial building block of our approach in building the distributions is actually inherent to a wide variety of problems, for instance, recommending products such as movies or web-advertisements (Grbovic et al., 2015), nodes in a graph (Grover and Leskovec, 2016), sequence data, or other entities (Wu et al., 2017). Particularly when training point-wise embeddings for textual data, the co-occurrence information is already computed as the first step, like in GloVe (Pennington et al., 2014), but does not get utilized beyond this.

Lastly, the connection to optimal transport at the level of entities and contexts paves the way to make better use of its vast toolkit (e.g., Wasserstein distances, barycenters, barycentric coordinates, etc.) for applications, which in the case of NLP has primarily been restricted to document distances (Kusner et al., 2015; Huang et al., 2016).

Contributions: 1) Employing the notion of optimal transport of contexts as a distance measure, we illustrate how our framework can be beneficial for important tasks involving word and sentence representations, such as sentence similarity, hypernymy (entailment) detection and word similarity. The method can be readily used on top of existing embedding methods and does not require any additional learning.

2) The resulting representations, as portrayed in Figures 1, 3, 4, capture the various senses under which the entity occurs. Further, the transport map obtained through CMD (see Figure 2) gives a clear interpretation of the resulting distance obtained between two entities.

3) CMD can be used to measure any task-specific distance (even asymmetric costs) between words, by defining a suitable underlying cost on the movement of contexts, which we show can lead to a state-of-the-art metric for unsupervised word entailment.

4) Defining the transport over contexts has the additional benefit that the representations are compositional - they directly extend from entities to groups of entities (of any size), such as from word to sentence representations. To this end, we utilize the notion of Wasserstein barycenters, which to the best of our knowledge has never been considered in the past. This results in a significant performance boost on multiple datasets, and even outperforming popular supervised methods like InferSent (Conneau et al., 2017) and GenSen (Subramanian et al., 2018) by a decent margin.

2 Related Work

Vector representations. The idea of using vector space models for natural language dates back to Bengio et al. (2003), but in particular, has been popularized by Word2vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014). One of the problems that still persists is the inability to capture, within just a point embedding, the multiple senses or semantics associated with the occurrence of a word. This has resulted in works that either maintain multiple embeddings Huang et al. (2012), or utilize bilingual parallel corpora Guo et al. (2014), or learn embeddings that capture some specific information Levy and Goldberg (2014a), but a general solution still remains to be found.

Representing with distributions. This line of work is fairly recent, mainly originating from Vilnis and McCallum (2014), who proposed to represent words with Gaussian distributions, and later extended to mixtures of Gaussians in (Athiwaratkun and Wilson, 2017). Concurrent to this work, Muzellec and Cuturi (2018) and Sun et al. (2018) have suggested using elliptical and Gaussian distributions endowed with a Wasserstein metric respectively. While these methods[1] already provide richer information than typical vector embeddings, their form restricts what could be gained by allowing for arbitrary distributions (in terms of being free from assumptions on their shape or modality) as possible here. Our proposal of distributional estimate (i.e., distribution over context embeddings), inherently relies upon the empirically obtained co-occurrence information of a word and its contexts. Hence, this naturally allows for the use of optimal transport (or Wasserstein metric) in the space containing the contexts, and leads to an interpretation[2] (Figure 2) which is not available in the above approaches.

After the release of our initial technical report[3], Frogner

[1] Elliptical embeddings (Muzellec and Cuturi, 2018) also depend on WordNet supervision in the case of hypernymy.

[2] We explicitly connect the ground space to the space of contexts, which enables a better interpretation of the transport map.

[3] The first version of this article appeared on 5th June, 2018 at https://openreview.net/forum?id=Bkx2jd4Nx7 titled as 'Wasserstein is all you need'.

et al. (2019) also independently propose to embed entities as discrete distributions in the Wasserstein space. A key distinction is that the training procedure required to learn such representations in their and the above-mentioned methods is not necessary for our approach, since we can just utilize the existing pre-trained point-embeddings together with the co-occurrence information. Further, these methods (except for Frogner et al. (2019)) don't provide a way to represent composition of entities (e.g. sentences) which is available via our framework (see Section 5).

Optimal Transport in NLP. The primary focus of the explorations of optimal transport in NLP has been on transporting words or sets of words directly, and for downstream applications rather than representation learning in general. Existing examples include document distances (Kusner et al., 2015; Huang et al., 2016), topic modelling (Rolet et al., 2016; Xu et al., 2018), document clustering (Ye et al., 2017), and others (Zhang et al., 2017; Grave et al., 2018). For example, the Word Mover's Distance (WMD; Kusner et al., 2015) considers computing the distance between documents as an optimal transport between their bag-of-words, and in itself doesn't lead to a representation. When the transport is defined at the level of words, like in these approaches, it can not be used to represent words themselves. In our approach, the transport is considered over contexts instead, which enables us to develop representations for words and extend them to represent composition of words (i.e., sentences, documents) in a principled manner, as illustrated in Sections 5 and 6.

3 Background on Optimal Transport

Optimal Transport (OT) provides a way to compare two probability distributions defined over a space $\mathcal{G}$ (commonly known as the ground space), given an underlying distance or more generally the cost of moving one point to another in the ground space. In other terms, it lifts a distance between points to a distance between distributions. In contrast, Kullback-Leibler (KL), or f-divergences in general, only focus on the probability mass values, thus ignoring the geometry of the ground space: something which we exploit via OT. Also, $\mathrm{KL}(\mu||\nu)$ is defined only when the distribution μ is absolutely continuous with respect to ν. Below is a brief background on OT in the discrete case.

Linear Program (LP) formulation. Consider an empirical probability measure of the form $\mu = \sum_{i=1}^{n} a_i \delta(\boldsymbol{x}^{(i)})$ where $X = (\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(n)}) \in \mathcal{G}^n$, $\delta(\boldsymbol{x})$ denotes the Dirac (unit mass) distribution at point $\boldsymbol{x} \in \mathcal{G}$, and $(a_1, \dots, a_n)$ lives in the probability simplex $\Sigma_n := \{\boldsymbol{p} \in \mathbb{R}_+^n \mid \sum_{i=1}^{n} p_i = 1\}$. Now given a second empirical measure, $\nu = \sum_{j=1}^{m} b_j \delta(\boldsymbol{y}^{(j)})$, with $Y = (\boldsymbol{y}^{(1)}, \dots, \boldsymbol{y}^{(m)}) \in \mathcal{G}^m$, and $(b_1, \dots, b_m) \in \Sigma_m$, and if the ground cost of moving from point $\boldsymbol{x}^{(i)}$ to $\boldsymbol{y}^{(j)}$ is denoted by M_{ij}, then the OT distance between μ and ν is the solution to the following LP.

$$\mathrm{OT}(\mu, \nu; M) := \min_{\substack{T \in \mathbb{R}_+^{n \times m} \\ \text{s.t. } \forall i, \sum_j T_{ij} = a_i,\ \forall j, \sum_i T_{ij} = b_j}} \sum_{ij} T_{ij} M_{ij} \quad (1)$$

The optimal $T \in \mathbb{R}_+^{n \times m}$ is referred to as the *transportation matrix*, where T_{ij} denotes the optimal amount of mass to move from point $\boldsymbol{x}^{(i)}$ to point $\boldsymbol{y}^{(j)}$. Intuitively, OT is concerned with the problem of moving a given supply of goods from certain factories to meet the demands at some shops, such that the overall transportation cost is minimal.

Distance. When $\mathcal{G} = \mathbb{R}^d$ and the cost is defined with respect to a metric $D_{\mathcal{G}}$ over $\mathcal{G}$ (i.e., $M_{ij} = D_{\mathcal{G}}(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(j)})^p$ for any i, j), OT defines a distance between empirical probability distributions. This is the p-Wasserstein distance, defined as $\mathcal{W}_p(\mu, \nu) := \mathrm{OT}(\mu, \nu; D_{\mathcal{G}}^p)^{1/p}$. In most cases, we are only concerned with the case where $p = 1$ or 2.

Barycenters. In Section 5, we will make use of the notion of averaging in the Wasserstein space. More precisely, the Wasserstein barycenter (Agueh and Carlier, 2011) is a probability measure that minimizes the sum of (p-th power) Wasserstein distances to the given measures. Formally, given N measures $\{\nu_1, \dots, \nu_N\}$ with corresponding weights $\boldsymbol{\eta} = \{\eta_1, \dots, \eta_N\} \in \Sigma_N$, the Wasserstein barycenter can be written as

$$\mathcal{B}_p(\nu_1, \dots, \nu_N) = \arg\min_{\mu} \sum_{i=1}^{N} \eta_i \mathcal{W}_p(\mu, \nu_i)^p. \quad (2)$$

Regularization and Sinkhorn iterations. The cost of exactly solving OT scales at least in $\mathcal{O}(n^3 \log(n))$ (n being the cardinality of the support of the empirical measure) when using network simplex or interior point methods. Following Cuturi (2013), we consider the entropy regularized Wasserstein distance, $\mathcal{W}_{p,\lambda}(\mu, \nu) := \mathrm{OT}_\lambda(\mu, \nu; D_{\mathcal{G}}^p)^{1/p}$, where the search space for the optimal T is instead restricted to a smooth solution close to the extreme points of the linear program identical to (2), but subtracting $\lambda H(T)$ from the linear objective, where $H(T) = -\sum_{ij} T_{ij} \log T_{ij}$. The regularized problem ($\lambda > 0$) can then be solved efficiently using Sinkhorn iterations (Sinkhorn, 1964). While the cost of each Sinkhorn iteration is quadratic in n, it has been shown in Altschuler et al. (2017) that convergence can be attained in a number of iterations that is independent of n, thus resulting in an overall complexity of $\widetilde{O}(n^2)$. Similarly, we consider the regularized barycenter ($\mathcal{B}_{p,\lambda}$) (Cuturi and Doucet, 2014) which uses regularized Wasserstein distances $\mathcal{W}_{p,\lambda}$ in Eq. (2) and can be cheaply computed by iterative Bregman projections (Benamou et al., 2015) to yield an

approximate solution. Overall, thanks to this entropic regularization, OT computations can be carried out efficiently in a parallel and batched manner on GPUs (in our use case, $n \approx 300$).

4 Methodology

In this section, we define the distributional estimate that we use to represent each entity and the corresponding OT based distance measure. Since we take the guiding example of building text representations, we consider each entity to be a word for clarity.

Distributional Estimate ($\mathbb{P}_V^w$). For a word w, its distributional estimate is built from a histogram H^w over the set of contexts $\mathcal{C}$, and an embedding of these contexts into a space $\mathcal{G}$. The histogram measures how likely it is that a word w occurred in a particular context c, i.e., probability $p(c|w)$. In absence of an exact closed-form expression, we can use its empirical estimate given by the frequency of the word w in context c, relative to the total frequency of word w in the corpus.

Thus one natural way to build this histogram is to maintain a co-occurrence matrix between words in our vocabulary and all possible contexts, where each entry indicates how often a word and context occur in a (symmetric) window of fixed size L. Then, the bin values $(\mathrm{H}^w)_{c\in\mathcal{C}}$ of the histogram can be viewed as the row corresponding to w in this co-occurrence matrix.

Next, the simplest embedding of contexts is into the space of one-hot vectors of all the possible contexts. However, this induces a lot of sparsity in the representation and the distance between such embeddings of contexts does not reflect their semantics. A classical solution would be to instead find a dense low-dimensional embedding of contexts that captures the semantics, possibly using techniques such as SVD or deep neural networks. We denote by $V = (\mathbf{v}_c)_{c\in\mathcal{C}}$ an embedding of the contexts into this low-dimensional space $\mathcal{G} \subset \mathbb{R}^d$, which we refer to as the *ground space*.

Combining the histogram H^w and the context embeddings V, we represent the word w by the following empirical distribution, referred to as the *distributional estimate* of the word:

$$\mathbb{P}_V^w := \sum_{c\in\mathcal{C}} (\mathrm{H}^w)_c \; \delta(\mathbf{v_c}). \tag{3}$$

Distance. If we equip the ground space $\mathcal{G}$ with a meaningful metric $D_\mathcal{G}$ and use distributional estimates $(\mathbb{P}_V^w)$ to represent the words, then we can define a distance between two words w_i and w_j as the solution to the following optimal transport problem:

$$\mathrm{CMD}(w_i, w_j; D_\mathcal{G}^p) \stackrel{\text{def}}{=} \mathrm{OT}(\mathbb{P}_V^{w_i}, \mathbb{P}_V^{w_j}; D_\mathcal{G}^p) \simeq \mathcal{W}_{p,\lambda}(\mathbb{P}_V^{w_i}, \mathbb{P}_V^{w_j})^p. \tag{4}$$

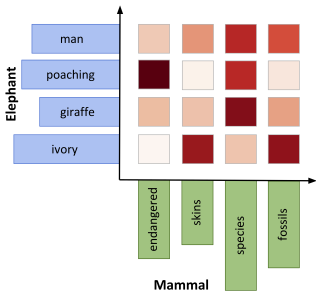


Figure 2: *Illustration of Context Mover's Distance (CMD) (Eq. (4)) between elephant & mammal* (when represented with their distributional estimates and using entailment ground metric discussed in Section 6). Here, we pick four contexts at random from their top 20 contexts in terms of PPMI. The square cells denote the entries of the transportation matrix T obtained while computing CMD. The darker a cell, larger the amount of mass moved.

Intuition. Two words are similar in meaning if the contexts of one word can be easily transported to the contexts of the other, with this cost of transportation being measured by $D_\mathcal{G}$. This idea still remains in line with the distributional hypothesis (Harris, 1954; Rubenstein and Goodenough, 1965) that words in similar contexts have similar meanings, but provides a precise way to quantify it. We thus call the distance in Eq.(4) the *Context Mover's Distance* (CMD).

Interpretation. The particular definition of CMD in Eq.(4), lends a pleasing interpretation (c.f. Figure 2) in terms of the transportation map T. This can be useful in understanding why and how are the two words being considered as similar in meaning, by looking at this movement of contexts.

Additionally, CMD between two words can be thought of as computing the WMD between some hypothetical documents associated to each word, which contain all possible contexts of the respective words.

Mixed Distributional Estimate. Based on a given task, it might be useful to reduce the cost incurred from extraneous contexts, or in other words, to adjust the amounts of "distribution" and "point" nature needed for the representation. This can be done by adding the point estimate of the target entity as an additional context in the distributional estimate, with a particular mixing weight m. The other contexts in the distributional estimate are reweighted to sum to $1 - m$.

Concrete Framework. For simplicity, we limit the contexts to consist of single words and discuss how the framework can be concretely applied as follows:

(i) Making associations better. It is commonly understood that co-occurrence counts alone may not necessarily suggest a strong association between a word and a context. The well-known Positive Pointwise Mutual

Information (PPMI) matrix (Church and Hanks, 1990; Levy et al., 2015) addresses this shortcoming, and in particular, we use its smoothed and shifted variant called SPPMI (Levy and Goldberg, 2014b). Overall, this enables us to extract better semantic associations from the co-occurrence matrix. Hence, the bin values (at context c) for the histogram of word w in Eq. (3) can be written as: $(\mathrm{H}^w)_c := \frac{\mathrm{SPPMI}(w,c)}{\sum_{c \in \mathcal{C}} \mathrm{SPPMI}(w,c)}$. Building this histogram information comes almost for free while learning point embeddings, as in GloVe (Pennington et al., 2014).

(ii) Computational considerations: A natural question could arise that CMD might be computationally intractable in its current formulation, as the possible number of contexts can be enormous. Since the contexts are mapped to dense embeddings, it is possible to only consider K *representative contexts*[4] , each covering some part $\mathcal{C}_k$ of the set of contexts $\mathcal{C}$. The histogram for word w with respect to these contexts can then be written as $\tilde{\mathbb{P}}^w_{\tilde{V}} = \sum_{k=1}^{K} (\tilde{\mathrm{H}}^w)_k \ \delta(\tilde{\mathbf{v}}_k)$, where $\tilde{\mathbf{v}}_k \in \tilde{V}$ is the point estimate of the k^{th} representative context, and $(\tilde{\mathrm{H}}^w)_k$ denotes the new histogram bin values (formed by combining the SPPMI contributions). More details on this, including precise definitions of SPPMI and the effect of number of clusters, are given in the supplementary sections S1.2 and S2.

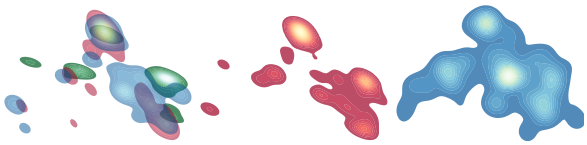
Another way would have been to consider the top-K contexts by SPPMI but we don't go this route, since the computations can't be batched when the supports are different. Also this would require reducing the support of the obtained barycenter, back to K, everytime.

Overall efficiency. Thus, with the batched implementations on a Nvidia TitanX GPU, it is possible to compute $\approx$ **13,700** Wasserstein-distances/second (for histogram size 100), and **4,600** Wasserstein-barycenters/second (for sentence length 25 and histogram size 100).

5 Sentence Representations

The goal of this task is to develop a representation for sentences, that captures the semantics conveyed by it. Most unsupervised representations proposed in the past rely on the composition of word embeddings, through additive, multiplicative, or other ways (Mitchell and Lapata, 2008; Arora et al., 2017; Pagliardini et al., 2017). As before, our aim is to represent sentences by distributional estimates to better capture the inherent uncertainty and polysemy. We hypothesize that a sentence, $S = (w_1, w_2, \ldots, w_N)$, can be effectively represented via the Wasserstein barycenter of the distributional estimates of its words, $\tilde{\mathbb{P}}^S_V := \mathcal{B}_{p,\lambda}\left(\tilde{\mathbb{P}}^{w_1}_V, \tilde{\mathbb{P}}^{w_2}_V, \ldots, \tilde{\mathbb{P}}^{w_N}_V\right)$.

[4]In practice, these are the cluster centroids obtained by applying K-means to context embeddings under $D_{\mathcal{G}}$.



(a) Dist. estimate (b) Euclidean avg (c) W-barycenter.

Figure 3: Distributional estimate of *'my'* (green), *'love'* (red) and *'life'* (blue). Illustrates how Wasserstein barycenter (and thus CoMB) utilizes the geometry of ground space, while the Euclidean averaging just focuses on the probability mass. [6]

The motivation is that since the barycenter minimizes the sum of optimal transports, cf. Eq. (2), it should result in a representation which best captures the simultaneous occurrence of the words in a sentence. Henceforth, we refer to this representation as *Context Mover's Barycenters* (CoMB).

To give a better idea of the nature of Wasserstein barycenter underlying CoMB, consider two Diracs measures, $\delta(\boldsymbol{x})$ and $\delta(\boldsymbol{y})$, with equal weights and under Euclidean ground metric. Then, the Wasserstein barycenter is $\delta(\frac{\boldsymbol{x}+\boldsymbol{y}}{2})$ while simple averaging gives $\frac{1}{2}(\delta(\boldsymbol{x}) + \delta(\boldsymbol{y}))$. Figure 3 highlights this interpolating nature of Wasserstein barycenter in the ground space, and illustrates how it is better suited for using the innate geometry of tasks (here context embeddings) as compared to the simple Euclidean averaging.

Averaging of point-estimates, in many variants (Iyyer et al., 2015; Arora et al., 2017; Pagliardini et al., 2017), has been shown to be surprisingly effective for multiple NLP tasks including sentence similarity. Interestingly, this can be seen as a special case of CoMB, when the distribution associated to a word is just a Dirac at its point estimate. It becomes apparent that having a rich distributional estimate for a word could be beneficial.

Since with CoMB, each sentence is also a distribution over the ground space $\mathcal{G}$ containing the contexts, we utilize the Context Mover's Distance (CMD) defined in Eq.(4) to define the distance between two sentences S_1 and S_2 as follows, $\mathrm{CMD}(S_1, S_2; D^p_{\mathcal{G}}) := \mathrm{OT}(\tilde{\mathbb{P}}^{S_1}_V, \tilde{\mathbb{P}}^{S_2}_V; D^p_{\mathcal{G}}) \simeq \mathcal{W}_{p,\lambda}(\tilde{\mathbb{P}}^{S_1}_V, \tilde{\mathbb{P}}^{S_2}_V)^p$. Here, the ground metric $D_{\mathcal{G}}$ is typically Euclidean or angular distance between the point embeddings.

Experimental Setup. To evaluate the effectiveness of an unsupervised sentence representation, we consider the semantic textual similarity (STS) tasks across 24 datasets from SemEval (Agirre et al., 2012, 2013, 2014, 2015, 2016), containing sentences from domains such as news headlines, forums, Twitter, etc. The objective

[6]For visualization purposes in Figures 1, 3, 4, we compute a 2D representation of the actual context embeddings using t-SNE (Maaten and Hinton, 2008) and use a kernel density estimate to smooth the distributions.

Model	Corpus (# tokens)	Val. Set	Test Set				
		STS16	STS12	STS13	STS14	STS15	**Avg.**
*(a) Unsupervised methods based on **GloVe** embeddings*							
NBoW	TBC (0.9 B)	19.2	21.1	13.5	25.0	30.7	22.6
SIF		26.6	32.4	23.0	34.1	35.3	31.2
SIF$_{\text{(PC removed)}}$		57.6	41.0	50.1	51.9	52.8	49.0
Euclidean avg.		50.7	45.7	39.0	49.9	53.5	47.0
CoMB		52.4	48.2	42.2	54.9	53.8	49.8
CoMB $_{\text{Mix}}$		60.2	50.5	51.0	58.3	60.5	55.1
CoMB $_{\text{Mix + PC removed}}$		63.0	49.3	56.5	60.8	64.0	57.7
*(b) Unsupervised methods based on **Sent2vec** embeddings*							
Sent2vec	TBC (0.9 B)	69.1	55.6	57.1	68.4	74.1	63.8
Sent2vec $_{\text{(PC removed)}}$		69.0	57.0	62.8	70.1	72.8	65.7
CoMB $_{\text{Mix}}$		70.1	56.1	59.7	68.8	73.7	64.6
CoMB $_{\text{Mix + PC removed}}$		70.6	57.9	64.2	**70.3**	73.1	**66.4**
(c) Unsupervised methods across different corpora							
Skip-thought (Arora et al., 2017)	TBC (0.9 B)	NA	30.8	24.8	31.4	31.0	29.5
WME (Word2vec)	Google News (100 B)	NA	**60.6**	54.5	65.5	61.8	60.6
SIF $_{\text{PC removed}}$ (GloVe)	Common Crawl (840 B)	NA	56.2	56.6	68.5	71.7	63.3
CoMB $_{\text{Mix + PC removed}}$ (GloVe)	TBC + News Crawl (5 B)	**72.0**	54.9	**67.2**	67.5	72.0	65.4
(d) Supervised methods							
GenSen (Kiros and Chan, 2018)	AllNLI, TBC, WMT, etc.	66.4	**60.6**	54.7	65.8	**74.2**	63.8
InferSent	AllNLI (26 M)	71.5	59.2	58.9	69.6	71.3	64.8

Table 1: Average *Pearson correlation* ($\times 100$) on STS tasks for CoMB and other baselines. *'Mix'* denotes the mixed distributional estimate. *'PC removed'* refers to removing contribution along the principal component of point estimates (as in SIF). See S3 for detailed results and hyperparameters.

here is to give a similarity score to each sentence pair and rank them, which is evaluated against the ground truth ranking via Pearson correlation.

We benchmark the performance of CoMB using SentEval (Conneau and Kiela, 2018) against a variety of unsupervised methods such as (a) Neural Bag-of-Words (NBoW) averaging of point estimates, (b) SIF from Arora et al. (2017) who regard it as a "simple but tough-to-beat baseline" and utilize weighted NBoW averaging with principal component removal, (c) Sent2vec (Pagliardini et al., 2017) which learns word embeddings so that their average works well as a sentence representation, (d) Skip-thought (Kiros et al., 2015) which trains an LSTM-based encoder to predict surrounding sentences, and (e) Word Mover's Embedding (WME; Wu et al., 2018) which is a recent variant of WMD. For comparison, we also show the performance of recent supervised methods such as InferSent (Conneau et al., 2017) and GenSen (Subramanian et al., 2018), although these methods are clearly at an advantage due to training on labeled corpora.

Empirical Results. *(i) Ground Metric: GloVe.* Table 1 (a) compares the performance of CoMB against other methods using the same GloVe embeddings trained on the common Toronto Book Corpus (TBC) (Zhu et al., 2015). We observe that the vanilla CoMB significantly outperforms SIF and NBoW, showing the benefit of having the distributional estimate instead of just a Dirac. Also, it is better than SIF $_{\text{PC removed}}$ on average across the test set, and using the mixed distributional estimate (CoMB $_{\text{Mix}}$) further improves the average test performance by 10%. Next, when the PC removal is carried out for point estimates during mixing (i.e., CoMB $_{\text{Mix + PC removed}}$), the average performance increases to 57.7. Both of these are for mixing weight $m = 0.4$ towards the point estimate. Also, we see empirical evidence that the Euclidean average of the distributional estimates (Figure 3b) performs worse than Wasserstein barycenter (CoMB), when measuring the sentence similarity using CMD for both.

(ii) Ground Metric: Sent2Vec. Our method is not specific to GloVe embeddings, and in Table 1 (b), we see the effect of using an improved ground metric, by employing word vectors from Sent2vec. Here, we notice that our best variant, CoMB$_{\text{Mix + PC removed}}$, results in a relative improvement of 4% over Sent2vec, which is a decent gain considering that for unstructured text corpora it is a state-of-the-art unsupervised method.

(iii) Overall comparisons: To facilitate an accurate comparison with baselines which typically use huge corporas, in Table 1 (c) we report our results (with ground metric GloVe) by using the News Crawl corpus (Bojar et al., 2018) combined with TBC. First of all, this increase in data boosts the performance of CoMB from 57.7 to 65.4, which outperforms WME despite using a smaller corpus. Thus, pointing towards the advantage of defining transport over contexts than words. Further, CoMB also outperforms SIF $_{\text{PC removed}}$ trained on Common Crawl and popular supervised

sentence embedding methods[7] such as GenSen and InferSent which utilize labeled corpora.

Qualitative analysis and ablation. We discuss this extensively in our supplementary section, but the main observations include: (a) Section S4.3: we qualitatively analyze the averaging of distributional estimates versus point estimates and find that the nature of errors made by CoMB and SIF are complementary in nature. CoMB outperforms when the difference in sentences stems from predicate while SIF is better when the distinguishing factor is the subject of the sentences. (b) Section S2.3: we observe that by around $K = 300$ to 500, the performance gained by increasing the number of clusters starts to plateau, implying that it is sufficient to only consider the representative contexts. (c) Section S5: CoMB shows promise for application in a downstream task like sentence completion, although a quantitative evaluation remains beyond the scope.

Summary and further prospects. Overall, this highlights the advantage of distributional estimates for words, that can be extended to give meaningful representation of sentences via CoMB in a principled manner. In terms of efficiency, it takes about 3 minutes on one GPU to get results on all the STS tasks comprising 25,000 sentences (see S1.4 for details). A future avenue would be to utilize the non-associativity of Wasserstein barycenters (i.e., $B_p(\mu, B_p(\nu, \xi)) \neq B_p(B_p(\mu, \nu), \xi)$), to take into account the word order with various aggregation strategies (like parse trees).

6 Hypernymy Detection

In linguistics, hypernymy is a relation between words where the semantics of one word (the *hyponym*) are contained within that of another word (the *hypernym*). A simple form of this relation is the *is-A* relation, e.g., *cat* is an *animal*. Hypernymy is a special case of the more general concept of lexical entailment, detecting which is relevant for tasks such as Question Answering.

Early unsupervised approaches for this task exploited various linguistic properties of hypernymy (Weeds and Weir, 2003; Kotlerman et al., 2010; Santus et al., 2014; Rimell, 2014). While most of these are count-based, point embedding methods (Chang et al., 2017; Henderson and Popa, 2016) have become popular in recent years. Other approaches represent words by Gaussian distributions with KL-divergence as an entailment measure (Vilnis and McCallum, 2014; Athiwaratkun and Wilson, 2017). These methods have proven powerful, as they capture not only the semantics, but also the uncertainty about the contexts in which a word appears.

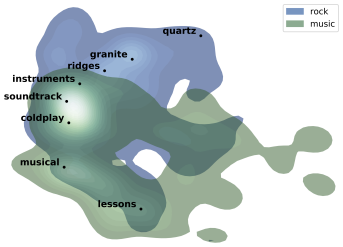


Figure 4: Distributional estimates of *rock* and *music*. The two words have an overlapping mode (for rock in the sense of rock music) and separate modes for other senses (such as rock in the sense of a stone).

Therefore, hypernymy detection is a great testbed to verify the effectiveness of our approach to represent each entity by the distribution of its contexts. The intuitive idea for the applicability of our method to this task originates from the *Distributional Inclusion Hypothesis* (Geffet and Dagan, 2005), which states that a word v entails another word w if "the most characteristic contexts of v are expected to be included in all w's contexts (but not necessarily amongst the most characteristic ones for w)". The inclusion of the contexts for the words *rock* and *music* is illustrated in Figure 4. We view our method as a relaxation of this strict inclusion condition by modeling it more softly with the optimal transport between the set of contexts corresponding to the hyponym and hypernym. Hence, it is natural to make use of the Context Mover's Distance (CMD), Eq. (4), but with a ground cost that measures entailment relations.

For this, we utilize a recently proposed method by Henderson et al. (Henderson and Popa, 2016; Henderson, 2017), which explicitly models what information is known about a word, by interpreting each entry of the embedding as the degree to which a certain feature is present. Based on the logical definition of entailment they derive an operator measuring the entailment similarity between two so-called entailment vectors defined as follows: $\boldsymbol{v}_i \circledcirc \boldsymbol{v}_j = \sigma(-\boldsymbol{v}_i) \cdot \log \sigma(-\boldsymbol{v}_j)$, where the sigmoid σ and log are applied componentwise on the embeddings. Thus, we use as ground cost $D_{ij}^{\text{Hend.}} := -\boldsymbol{v}_i \circledcirc \boldsymbol{v}_j$. This asymmetric ground cost shows that our framework can be flexibly used with an arbitrary cost function defined on the ground space.

Evaluation. In total, we evaluate our method on 10 standard datasets using the HypEval[10] evaluation toolkit. The foremost thing that we would like to check is the benefit of having a distributional estimate in comparison to just the point embeddings. Here,

[7] USE (Cer et al., 2018), which relies on a labeled corpus, doesn't report results on STS12-15 but according to (BERT official repo, 2019), its performance is 67.5 which is close to CoMB's unsupervised performance of 66.4. See also BERT's performance in BERT official repo (2019).

[9] Scores for GE+C, GE+KL, and DIVE are taken from (Chang et al., 2017) as we use the same evaluation setup.

[10] https://github.com/context-mover/HypEval

	Validation Set	Test Set						
Method	HypeNet-Train (Shwartz et al., 2016)	HypeNet-Test (Shwartz et al., 2016)	EVALution (Santus et al., 2015)	LenciBenotto Benotto (2015)	Weeds (Weeds et al., 2014)	Turney (Turney and Mohammad, 2015)	Baroni (Baroni and Lenci, 2011)	BIBLESS (Kiela et al., 2015)
$D^{\text{Hend.}}$	29.0	28.8	31.6	44.8	60.8	<u>56.6</u>	<u>78.3</u>	67.7
$\text{CMD}_{K=200}+D^{\text{Hend.}}$	<u>53.4</u>	<u>53.4</u>	**38.1**	<u>50.1</u>	<u>63.9</u>	56.0	67.5	**75.4**
$\text{CMD}_{K=250}+D^{\text{Hend.}}$	**53.6**	**53.7**	<u>37.1</u>	49.9	63.8	56.3	67.3	<u>75.2</u>
GE + Cosine	NA	21.6	26.7	43.3	52.0	53.9	69.7	NA
GE + KL	NA	23.7	29.6	45.1	51.3	52.0	64.6	NA
DIVE	NA	32.0	33.0	**50.4**	**65.5**	**57.2**	**83.5**	NA
Poincaré GloVe	NA	NA	NA	NA	NA	NA	NA	65.2

Table 2: Comparison of the entailment vectors from Henderson (2017) used alone ($D^{\text{Hend.}}$), and when used together with our Context Mover's Distance (CMD_K) as the underlying ground metric, with state-of-the-art unsupervised methods. The two listed CMD variants are the ones with best validation performance for $K = 200$ and 250 clusters. The scores are **AP@all (%)**[9], except for BIBLESS where it is accuracy. More details about the training setup and results on other datasets can be found in Section S1.1, and Table S12 in Section S6.2. Best results for each column are in **bold** and the 2nd best are <u>underlined</u>.

	Validation Set	Test Set											
Method	MEN + SimVerb-D [2499]	MC [30]	MTurk-287 [285]	MTurk-771 [771]	RG [65]	RW [1493]	SimLex [998]	Verb [144]	WS-ALL [352]	WS-REL [251]	WS-SIM [203]	YP [130]	SimVerb-T [2999]	**Wt. Average** [7721]
D^{GloVe}	**61.1**	67.8	66.8	61.7	**73.7**	33.7	**34.8**	**26.4**	53.5	40.9	68.2	**54.5**	**19.4**	35.0
$\text{CMD}_{\mathbf{K=400}}+D^{\text{GloVe}}$	60.8	**69.6**	**67.8**	**62.1**	73.1	**38.9**	33.9	23.6	**55.3**	**44.1**	**69.2**	53.0	18.9	**35.9**

Table 3: Performance on standard word similarity tasks measured by Spearman's rho x 100. Both methods are based on *Toronto Book Corpus (0.9 B)* to ensure fair comparison. The last column is the weighted average of test set performance, with weights as [# of word pairs] present in the vocabulary for each dataset.

we observe that employing CMD along with the entailment embeddings, leads to a significant boost on most of the datasets, except on Baroni and Turney, where the performance is still competitive with the other state of the art methods like Gaussian embeddings (GE). The more interesting observation is that on some datasets (EVALution, HypeNet, LenciBenotto) we even outperform or match state-of-the-art performance (cf. Table 2), by simply using CMD together with this ground cost $D_{ij}^{\text{Hend.}}$ based on the entailment embeddings. Further, on BIBLESS (equivalent to WBLESS), CMD performs better than the state-of-the-art unsupervised method, Poincaré GloVe, as reported in Tifrea et al. (2018). Also, qualitative analysis can be found in Table S18/S19 of the supplementary. Lastly, these results can be efficiently computed in less than 3 minutes on a single GPU for all datasets (>100,000 pairs) and check Table S15 for details.

7 Word Similarity

The hypernymy detection results indicate the advantage gained by representing with a distribution over contexts than a point embedding for a word-level task. Nevertheless, we present results for another standard word-level task: namely word similarity and relatedness (Faruqui and Dyer, 2014) in Table 3. We utilize GloVe embeddings with cosine similarity as a baseline for point embedding methods and compare the performance by using our Context Mover's Distance (CMD) on top. But some other point embedding method can also be plugged into CMD similarly.

In the above, we use the combined development sets of MEN and SimVerb for validation, as these are the only datasets with pre-defined development and test splits. The mixing weight m (see Section 4) is 0.8, the PPMI smoothing $\alpha = 0.15$ and the number of clusters $K = 400$ for the CMD experiment. We observe that on a majority of the test datasets, CMD results in a performance gain and also performs better on average. An extensive analysis with different embeddings and corpora is however outside the current scope.

8 Conclusion

We advocate for representing entities by a distributional estimate on top of any given co-occurrence structure. For each entity, we jointly consider the histogram information (with its contexts) as well as the point embeddings of the contexts. We show how this enables the use of optimal transport over distributions of contexts. Our framework results in an efficient, interpretable and compositional metric to represent and compare entities (e.g. words) and groups thereof (e.g. sentences), while leveraging existing point embeddings. We demonstrate its performance on several NLP tasks such as word and sentence similarity, as well as hypernymy detection. A practical take-home message is: *do not throw away the co-occurrence information* (e.g. when using GloVe), but *instead pass it on to our method.* Motivated by the promising results, learning the distributional estimates and applying the proposed framework on co-occurrence structures beyond NLP are exciting future directions.

Acknowledgments

We would like to acknowledge Alexis Conneau, Tom Bosc, and anonymous reviewers, for their helpful comments. Also, we thank all the members of MLO for fruitful discussions. SPS is indebted to Marco Cuturi for teaching him about Optimal Transport and Honda Foundation for sponsoring that visit.

References

Eneko Agirre, Mona Diab, Daniel Cer, and Aitor Gonzalez-Agirre. Semeval-2012 task 6: A pilot on semantic textual similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 385–393. Association for Computational Linguistics, 2012. 5

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. * sem 2013 shared task: Semantic textual similarity. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, volume 1, pages 32–43, 2013. 5

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. Semeval-2014 task 10: Multilingual semantic textual similarity. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, pages 81–91, 2014. 5

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Inigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, et al. Semeval-2015 task 2: Semantic textual similarity, english, spanish and pilot on interpretability. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 252–263, 2015. 5

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 497–511, 2016. 5

Martial Agueh and Guillaume Carlier. Barycenters in the wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011. 3

Jason Altschuler, Jonathan Weed, and Philippe Rigollet. Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. In *Advances in Neural Information Processing Systems*, pages 1964–1974, 2017. 3

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. *ICLR*, 2017. 5, 6

Ben Athiwaratkun and Andrew Gordon Wilson. Multimodal word distributions. *arXiv preprint arXiv:1704.08424*, 2017. 2, 7

Marco Baroni and Alessandro Lenci. How we blessed distributional semantic evaluation. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 1–10. Association for Computational Linguistics, 2011. 8

Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré. Iterative bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2): A1111–A1138, 2015. 3

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. A neural probabilistic language model. *J. Mach. Learn. Res.*, 3:1137–1155, March 2003. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=944919.944966. 2

Giulia Benotto. Distributional models for semantic relations: A study on hyponymy and antonymy. *PhD thesis, University of Pisa*, 2015. 8

BERT official repo. Universal Sentence Encoder results on STS 12-16. https://github.com/google-research/bert/issues/128#issuecomment-451896503, 2019. Accessed: 23 May, 2019. 7

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. Findings of the 2018 conference on machine translation (WMT18). In *Proc. of WMT*, 2018. URL https://www.aclweb.org/anthology/W18-6401. 6

Daniel Cer, Yinfei Yang, Sheng yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. Universal sentence encoder, 2018. 7

Haw-Shiuan Chang, ZiYun Wang, Luke Vilnis, and Andrew McCallum. Distributional inclusion vector embedding for unsupervised hypernymy detection. *arXiv preprint arXiv:1710.00880*, 2017. 7

Kenneth Ward Church and Patrick Hanks. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29, 1990. 5

Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural

networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM, 2008. 1

Alexis Conneau and Douwe Kiela. SentEval: An Evaluation Toolkit for Universal Sentence Representations. *arXiv preprint arXiv:1803.05449*, 2018. 6

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364*, 2017. 2, 6

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pages 2292–2300, 2013. 3

Marco Cuturi and Arnaud Doucet. Fast computation of wasserstein barycenters. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 685–693, Bejing, China, 22–24 Jun 2014. PMLR. 3

Jan Deriu, Aurelien Lucchi, Valeria De Luca, Aliaksei Severyn, Simon Müller, Mark Cieliebak, Thomas Hofmann, and Martin Jaggi. Leveraging Large Amounts of Weakly Supervised Data for Multi-Language Sentiment Classification. In *WWW 2017 - International World Wide Web Conference*, pages 1045–1052, Perth, Australia, 2017. 1

Manaal Faruqui and Chris Dyer. Community evaluation and exchange of word vectors at wordvectors.org. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 19–24, Baltimore, Maryland, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/P14-5004. URL https://www.aclweb.org/anthology/P14-5004. 8

Charlie Frogner, Farzaneh Mirzazadeh, and Justin Solomon. Learning embeddings into entropic wasserstein spaces, 2019. 2, 3

Maayan Geffet and Ido Dagan. The distributional inclusion hypotheses and lexical entailment. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 107–114. Association for Computational Linguistics, 2005. 7

Edouard Grave, Armand Joulin, and Quentin Berthet. Unsupervised Alignment of Embeddings with Wasserstein Procrustes. *arXiv preprint arXiv:1805.11222*, 2018. 3

Mihajlo Grbovic, Vladan Radosavljevic, Nemanja Djuric, Narayan Bhamidipati, Jaikit Savla, Varun Bhagwan, and Doug Sharp. E-commerce in your inbox: Product recommendations at scale. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1809–1818. ACM, 2015. 2

Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *KDD 2016 - Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 855–864. ACM, 2016. 2

Jiang Guo, Wanxiang Che, Haifeng Wang, and Ting Liu. Learning sense-specific word embeddings by exploiting bilingual resources. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 497–507, 2014. 2

Zellig S Harris. Distributional structure. *Word*, 10(2-3): 146–162, 1954. 4

James Henderson. Learning word embeddings for hyponymy with entailment-based distributional semantics. *arXiv preprint arXiv:1710.02437*, 2017. 7, 8

James Henderson and Diana Nicoleta Popa. A vector space for distributional semantics for entailment. *arXiv preprint arXiv:1607.03780*, 2016. 7

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 1

Eric H Huang, Richard Socher, Christopher D Manning, and Andrew Y Ng. 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*, pages 873–882. Association for Computational Linguistics, 2012. 2

Gao Huang, Chuan Guo, Matt J Kusner, Yu Sun, Fei Sha, and Kilian Q Weinberger. Supervised word mover's distance. In *Advances in Neural Information Processing Systems*, pages 4862–4870, 2016. 2, 3

Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1681–1691, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-1162. URL https://www.aclweb.org/anthology/P15-1162. 5

Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. A Convolutional Neural Network for Modelling Sentences. In *ACL - Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 655–665, 2014. 1

Leonid V Kantorovich. On the translocation of masses. In *Dokl. Akad. Nauk. USSR (NS)*, volume 37, pages 199–201, 1942. 1

Douwe Kiela, Laura Rimell, Ivan Vulić, and Stephen Clark. Exploiting image generality for lexical entailment detection. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 119–124. Association for Computational Linguistics, 2015. 8

Yoon Kim. Convolutional Neural Networks for Sentence Classification. In *EMNLP 2014 - Empirical Methods in Natural Language Processing*, pages 1746–1751, 2014. 1

Jamie Kiros and William Chan. Inferlite: Simple universal sentence representations from natural language inference data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4868–4874. Association for Computational Linguistics, 2018. URL http://aclweb.org/anthology/D18-1524. 6

Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Skip-thought vectors. In *Advances in neural information processing systems*, pages 3294–3302, 2015. 6

Lili Kotlerman, Ido Dagan, Idan Szpektor, and Maayan Zhitomirsky-Geffet. Directional distributional similarity for lexical inference. *Natural Language Engineering*, 16(4):359–389, 2010. 7

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. In *International Conference on Machine Learning*, pages 957–966, 2015. 2, 3

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11): 2278–2324, 1998. 1

Omer Levy and Yoav Goldberg. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 302–308, 2014a. 2

Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. In *Advances in neural information processing systems*, pages 2177–2185, 2014b. 5

Omer Levy, Yoav Goldberg, and Ido Dagan. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225, 2015. 5

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008. 5

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*, pages 3111–3119, 2013. 1, 2

Jeff Mitchell and Mirella Lapata. Vector-based models of semantic composition. *proceedings of ACL-08: HLT*, pages 236–244, 2008. 5

Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris*, 1781. 1

Boris Muzellec and Marco Cuturi. Generalizing Point Embeddings using the Wasserstein Space of Elliptical Distributions . *arXiv preprint arXiv:1805.07594*, 2018. 2

Matteo Pagliardini, Prakhar Gupta, and Martin Jaggi. Unsupervised learning of sentence embeddings using compositional n-gram features. *arXiv preprint arXiv:1703.02507*, 2017. 5, 6

Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 1, 2, 5

Ye Qi, Devendra Sachan, Matthieu Felix, Sarguna Padmanabhan, and Graham Neubig. When and why are pre-trained word embeddings useful for neural machine translation? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 529–535. Association for Computational Linguistics, 2018. doi: 10.18653/v1/N18-2084. URL http://aclweb.org/anthology/N18-2084. 1

Laura Rimell. Distributional lexical entailment by topic coherence. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 511–519, 2014. 7

Antoine Rolet, Marco Cuturi, and Gabriel Peyré. Fast dictionary learning with a smoothed wasserstein loss. In Arthur Gretton and Christian C. Robert, editors, *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 630–638, Cadiz, Spain, 09–11 May 2016. PMLR. 3

Herbert Rubenstein and John B Goodenough. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633, 1965. 4

Enrico Santus, Alessandro Lenci, Qin Lu, and S Schulte im Walde. Chasing hypernyms in vector spaces with entropy. In *14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 38–42. EACL (European chapter of the Association for Computational Linguistics), 2014. 7

Enrico Santus, Frances Yung, Alessandro Lenci, and Chu-Ren Huang. Evalution 1.0: an evolving semantic dataset for training and evaluation of distributional semantic models. In *Proceedings of the 4th Workshop on Linked Data in Linguistics: Resources and Applications*, pages 64–69, 2015. 8

Aliaksei Severyn and Alessandro Moschitti. Twitter Sentiment Analysis with Deep Convolutional Neural Networks. In *38th International ACM SIGIR Conference*, pages 959–962, 2015. 1

Vered Shwartz, Yoav Goldberg, and Ido Dagan. Improving hypernymy detection with an integrated path-based and distributional method. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2389–2398. Association for Computational Linguistics, 2016. 8

Richard Sinkhorn. A relationship between arbitrary positive matrices and doubly stochastic matrices. *The Annals of Mathematical Statistics*, 35(2):876–879, 1964. ISSN 00034851. 3

Sandeep Subramanian, Adam Trischler, Yoshua Bengio, and Christopher J Pal. Learning general purpose distributed sentence representations via large scale multi-task learning, 2018. 2, 6

Chi Sun, Hang Yan, Xipeng Qiu, and Xuanjing Huang. Gaussian Word Embedding with a Wasserstein Distance Loss. *arXiv preprint arXiv:1808.07016v7*, 2018. 2

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In *NIPS*, 2014. 1

Alexandru Tifrea, Gary Bécigneul, and Octavian-Eugen Ganea. Poincar\'e glove: Hyperbolic word embeddings. *arXiv preprint arXiv:1810.06546*, 2018. 8

Peter D Turney and Saif M Mohammad. Experiments with three approaches to recognizing lexical entailment. *Natural Language Engineering*, 21(3):437–476, 2015. 8

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008. 1

Luke Vilnis and Andrew McCallum. Word representations via gaussian embedding. *arXiv preprint arXiv:1412.6623*, 2014. 2, 7

Julie Weeds and David Weir. A general framework for distributional similarity. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 81–88. Association for Computational Linguistics, 2003. 7

Julie Weeds, Daoud Clarke, Jeremy Reffin, David Weir, and Bill Keller. Learning to distinguish hypernyms and co-hyponyms. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2249–2259. Dublin City University and Association for Computational Linguistics, 2014. 8

Ledell Wu, Adam Fisch, Sumit Chopra, Keith Adams, Antoine Bordes, and Jason Weston. Starspace: Embed all the things! *arXiv preprint arXiv:1709.03856*, 2017. 2

Lingfei Wu, Ian En-Hsu Yen, Kun Xu, Fangli Xu, Avinash Balakrishnan, Pin-Yu Chen, Pradeep Ravikumar, and Michael J. Witbrock. Word mover's embedding: From word2vec to document embedding. In *EMNLP*, 2018. 6

Hongteng Xu, Wenlin Wang, Wei Liu, and Lawrence Carin. Distilled wasserstein learning for word embedding and topic modeling. *arXiv preprint arXiv:1809.04705*, 2018. 3

Jianbo Ye, Yanran Li, Zhaohui Wu, James Z Wang, Wenjie Li, and Jia Li. Determining gains acquired from word embedding quantitatively using discrete distribution clustering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1847–1856, 2017. 3

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. Earth mover's distance minimization for unsupervised bilingual lexicon induction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1934–1945. Association for Computational Linguistics, 2017. 3

Yukun Zhu, Ryan Kiros, Richard Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. *arXiv preprint arXiv:1506.06724*, 2015. 6