On the Compositionality Prediction of Noun Phrases using Poincaré Embeddings

Abhik Jana[†], **Dmitry Puzyrev**[‡], **Alexander Panchenko**[⋆,§], **Pawan Goyal**[†], **Chris Biemann**[§], and **Animesh Mukherjee**[†]

[†]Indian Institute of Technology Kharagpur, Kharagpur, India
[‡]National Research University Higher School of Economics, Moscow, Russia
[⋆]Skolkovo Institute of Science and Technology, Moscow, Russia
[§]Universität Hamburg, Hamburg, Germany
abhik.jana@iitkgp.ac.in, {pawang,animeshm}@cse.iitkgp.ac.in
dapuzyrev@edu.hse.ru
{panchenko,biemann}@informatik.uni-hamburg.de

Abstract

The compositionality degree of multiword expressions indicates to what extent the meaning of a phrase can be derived from the meaning of its constituents and their grammatical relations. Prediction of (non)-compositionality is a task that has been frequently addressed with distributional semantic models. We introduce a novel technique to blend hierarchical information with distributional information for predicting compositionality. In particular, we use hypernymy information of the multiword and its constituents encoded in the form of the recently introduced Poincaré embeddings in addition to the distributional information to detect compositionality for noun phrases. Using a weighted average of the distributional similarity and a Poincaré similarity function, we obtain consistent and substantial, statistically significant improvement across three gold standard datasets over state-of-the-art models based on distributional information only. Unlike traditional approaches that solely use an unsupervised setting, we have also framed the problem as a supervised task, obtaining comparable improvements. Further, we publicly release our Poincaré embeddings, which are trained on the output of handcrafted lexical-syntactic patterns on a large corpus.

1 Introduction

An important challenge in Natural Language Processing is to represent words, phrases, and larger spans in a way that reflects their meaning. Compositionality is one of the strongest assumptions in semantics, stating that the meaning of larger units can be derived from their smaller parts and their contextual relation. However, for idiomatic phrases, this assumption does not hold true as the meaning of the whole phrase may not be related to their parts in a straightforward fashion. The meaning of the phrases like 'data format', 'head teacher', 'green tree' can easily be understood from the constituent words whereas the semantics of the idiomatic phrases like 'couch potato', 'rat race', 'nut case' are non-compositional, i.e., refer to a different meaning than their parts suggest.

In this work, we address compositionality prediction, which is the task of assigning a numerical score to a phrase indicating the extent to which the meaning of the phrase can be derived from the meanings of its constituent words. To motivate its importance, e.g., in machine translation, non-compositional phrases must be translated as a unit; in word sense disambiguation, assigning one of the constituent word's senses to the whole phrase should be avoided for idiomatic phrases; semantic parsing also requires to correctly identify complex predicates and their arguments in this way.

A significant amount of effort has gone into operationalizing dense-vector distributional semantic models (DSMs) of different flavors such as count-based models (Baldwin et al. (2003); Venkatapathy and Joshi (2005); McCarthy et al. (2007)), word embeddings based on word2vec (both CBOW and SkipGram) and similar (Reddy et al. (2011); Salehi et al. (2014); Cordeiro et al. (2016, 2019)), and multi-sense skip-gram models for compositionality prediction (Salehi et al., 2015). All these attempts are based on the hypothesis that the composition of the representation of constituent words will be closer to the representation of the entire phrase in case of compositional phrases as compared to the non-compositional ones (Choueka, 1988).

Observing that the distributional information

alone is not enough for precise compositionality prediction, we propose to utilize hypernymy information, hypothesizing that, for compositional phrases, the hypernym of the whole phrase is semantically closer to the hypernyms of one of the constituent words (head words) as compared to the non-compositional phrases. For example, 'art school' and 'school' have one common hypernym 'educational institution' whereas 'hot dog' has no common hypernym with 'hot' or 'dog', apart from very abstract concepts such as 'physical entity'. Of course, this only holds for noun phrases, where taxonomic relations between nouns apply.

To represent hypernymy information we use Poincaré embeddings (Nickel and Kiela, 2017) for learning hierarchical representations of symbolic data by embedding them into a hyperbolic space. To this end, we extract hyponym-hypernym pairs by applying well-known lexical-syntactic patterns proposed by Hearst (1992) on a large corpus and train Poincaré embeddings on a list of hyponym-hypernym pairs.

Relying on two types of representations, i.e., dense vectors in the Euclidean space and the novel hyperbolic Poincaré embeddings, we interpolate their similarity predictions in a novel compositionality score metric that takes both distributional and hypernymy information into account. We evaluate our proposed metric on three well-accepted English datasets, i.e., Reddy (Reddy et al., 2011), Reddy++ (Ramisch et al., 2016) and Farahmand (Farahmand et al., 2015), demonstrating a performance boost when including hyperbolic embeddings by 2-4% absolute points across all datasets.

In particular, our work contains the three following **contributions**:

1. We devise a straightforward and efficient approach for combining distributional and hypernymy information for the task of noun phrase compositionality prediction. As far as we are aware, this is the first application of Poincaré embeddings to this task.

2. We demonstrate consistent and significant improvements on benchmark datasets in unsupervised and supervised settings.

3. We publicly release our Poincaré embeddings trained on pattern extractions on a very large corpus.

2 Related Work

Some of the initial efforts on compositionality prediction were undertaken by Baldwin et al. (2003), who use LSA to calculate the similarity between a phrase and its components, whereas Venkatapathy and Joshi (2005) extend this idea with collocation features (e.g., phrase frequency, point-wise mutual information). Researchers also tried to identify non-compositionality in verb-noun phrases using syntax (Cook et al., 2007) and selectional preferences (McCarthy et al., 2007). Attempts to examine the possibility to derive the semantics of a compound or multiword expression from its parts have been researched extensively (McCarthy et al., 2003; Mitchell and Lapata, 2008; Tratz and Hovy, 2010). Reddy et al. (2011) define a compositionality score and use different vector operations to estimate the semantic distance between a phrase and its individual components. Some of the investigations are made for compositionality detection using representation learning of word embeddings (Socher et al., 2012; Salehi et al., 2015). Salehi et al. (2014) also show that distributional similarity over multiple languages can help in improving the quality of compositionality prediction.

In a recent attempt, Yazdani et al. (2015) tries to learn semantic composition and finds that complex functions such as polynomial projection and neural networks can model semantic composition more effectively than the commonly used additive and multiplicative functions. Kiela and Clark (2013) detect non-compositionality using concepts of mutual information. Lioma et al. (2015) replace the context vectors with language models and compute their Kullback–Leibler divergence to approximate their semantic distance. In another stream, researchers have also attempted to classify idiomatic vs. non-idiomatic expressions in different languages considering the context of the expressions (Flor and Klebanov, 2018; Bizzoni et al., 2018; Peng et al., 2018), see also a respective shared task (Biemann and Giesbrecht, 2011). In one of the recent attempts, Cordeiro et al. (2016) conduct an analysis of several DSMs (word2vec, GloVe, PPMI) with variations of hyper-parameters and produce the state-of-the-art results in the compositionality prediction task, which is extended further for different languages by Cordeiro et al. (2019). We take their work as our baseline and carry forward our investigation to improve the state-of-the-art performance by introducing the

hyponymy-hypernymy information in the form of Poincaré embeddings.

Le et al. (2019) and Aly et al. (2019) also showed usefulness the use of Poincaré embeddings: in their case for inducing taxonomies from the text. In both works, hyperbolic embeddings are trained using relations harvested using Hearst patterns, like in our work. The usefulness of hyperbolic embeddings was also shown beyond text processing: Khrulkov et al. (2019) successfully applied them for hierarchical relations in image classification tasks.

3 Methodology

Our aim is to produce a compositionality score for a given two-word noun phrase w_1w_2. As per our hypothesis, the proposed compositionality score metric has two components: one component takes care of the extent of the distributional similarity between the phrase and the composition of constituent words. The second component captures hypernymy-based similarity obtained through Poincaré embeddings (Nickel and Kiela, 2017). The rationale behind this is that replacing a word with its hypernym should yield phrases with similar meaning for compositional cases, dissimilar phrases otherwise (e.g., a 'red herring' is not similar to 'red fish').

Distributional component: For the first component, we follow the scheme prescribed by Cordeiro et al. (2016), relying on the state-of-the-art DSM model and the score metric ($Score_D$) proposed in that work. The metric $Score_D$ is defined as,

$$Score_D(w_1w_2) = \cos(v(w_1w_2), v(w_1 + w_2)), \quad (1)$$

where

$$v(w_1 + w_2) = \frac{v(w_1)}{\|v(w_1)\|} + \frac{v(w_2)}{\|v(w_2)\|}, \quad (2)$$

and $v(w)$ is the vector representation of w obtained from the DSM, $||.||$ is the L2-norm. For the composition of two component word vectors, we use the additive model, which is well-accepted in the literature (Mitchell and Lapata, 2010).

Hypernymy component: For the second component, we prepare Poincaré embeddings. The Poincaré embedding as introduced by Nickel and Kiela (2017) is a very recent approach to learn hierarchical representations of symbolic data by embedding them into the hyperbolic space. The underlying hyperbolic geometry helps to learn parsimonious representations of symbolic data by simultaneously capturing hierarchy and similarity. As per this proposed Poincaré ball model, let

$$\beta^d = \{x \in \mathbb{R} \quad : \quad \|x\| < 1\} \quad (3)$$

be the open d-dimensional unit ball, where $\|.\|$ denotes the Euclidean norm.

The list of hyponym-hypernym pairs was obtained by applying lexical-syntactic patterns described by Hearst (1992) on the corpus prepared by Panchenko et al. (2016). This corpus is a concatenation of the English Wikipedia (2016 dump), Gigaword (Parker et al., 2009), ukWaC (Ferraresi et al., 2008) and English news corpora from the Leipzig Corpora Collection (Goldhahn et al., 2012). The lexical-syntactic patterns proposed by Hearst (1992) and further extended and implemented in the form of FSTs by Panchenko et al. (2012)[1] for extracting (noisy) hyponym-hypernym pairs are given as follows – (i) *such* NP *as* NP, NP[,] *and/or* NP; (ii) NP *such as* NP, NP[,] *and/or* NP; (iii) NP, NP [,] *or other* NP; (iv) NP, NP [,] *and other* NP; (v) NP, *including* NP, NP [,] *and/or* NP; (vi) NP, *especially* NP, NP [,] *and/or* NP.

Pattern extraction on the corpus yields a list of 27.6 million hyponym-hypernym pairs along with the frequency of their occurrence in the corpus. We normalize the frequency of each hyponym-hypernym pair by dividing it by the logarithm of the global frequency of the hypernym in the list, which realizes a TF-IDF (Sparck Jones, 1972) weighting, to downrank noisy extractions with frequent pattern-extracted 'hypernyms' such as 'problem, issue, bit'.

Further, we sort the list of hyponym-hypernym pairs with respect to their the normalized frequency. As the Poincaré embedding method takes as input a list of hyponym-hypernym pairs, we first prepare a list by adding top k pairs (based on normalized frequency) where the noun phrases or component words present in the gold-standard dataset exist as hyponym or hypernym. Note that we embed noun phrases as extracted by the patterns as units, i.e. a term like "educational institution" will get its own embedding if it appears in the pattern extractions as an NP. This list is quite sparse and therefore the hyperbolic space is

[1] https://zenodo.org/record/3234817

not rich enough to produce good results (see Section 5).

In order to circumvent this problem, we further populate the above list by appending the top m percent pairs from the complete sorted list of hyponym-hypernym pairs we prepared earlier. Next, we use this expanded list as input to prepare Poincaré embeddings.

Hyperparameters for training Poincaré model: For both the unsupervised and the supervised setup we maintain the following settings for the training of the Poincaré model unless otherwise stated: vector dimensionality d = 50, number of negative samples = 2, learning rate = 0.1, coefficient used for L2-regularization while training = 1, and number of epochs to use for burn-in initialization = 10.

3.1 Unsupervised Setup

The Poincaré distance between points $x, y \in \beta^d$ is defined in the following way:

$$d(x,y) = \text{arcosh}\left(1 + 2\frac{||x-y||^2}{(1-\|x\|^2)(1-\|y\|^2)}\right). \tag{4}$$

Poincaré similarity score $Score_P$ is derived from the Poincaré distance as

$$Score_P(x,y) = \frac{1}{1+d(x,y)}. \tag{5}$$

Let w_1w_2 be the noun phrase for which we compute the compositionality score. Further let $H_{w_1w_2}$ be the set of top k hypernyms of the phrase w_1w_2 and H_{w_1}, H_{w_2} be the set of top k hypernyms of the constituent words w_1 and w_2, respectively. Our proposed compositionality score metric $Score(w_1w_2)$ is defined as follows:

$$\begin{aligned} Score(w_1w_2) &= (1-\alpha)Score_D(w_1w_2) + \\ &\alpha \max_{\substack{a\in H_{w_1w_2} \\ b\in H_{w_1} \\ c\in H_{w_2}}} (Score_P(v(a), v(b)+v(c))), \end{aligned} \tag{6}$$

where $v(w)$ indicates the vector representation of the word w and α is used to set the relative weight of the two components.

3.2 Supervised Setup

We explore the utility of hierarchical information encoded in Poincaré embeddings for the task of compositionality prediction in a supervised setup as well. As our aim is to predict a compositionality score, we employ several regression techniques like Support Vector Regression (Drucker et al., 1997), Kernel Ridge Regression (Vovk, 2013), k-Nearest Neighbours Regression (Altman, 1992), Partial Least Squares Regression (PLS) (Abdi, 2007) etc. We randomly split the full dataset into a 75% training set and a 25% test set, and experiment on 25 such random splits. For each split, we plugin the concatenation of the vector representation of the noun phrase as well as the component words. The supervised predicted score is

$$\begin{aligned} Score_S(w_1w_2) = (1-\alpha)\cdot Score_{DS}(w_1w_2) + \\ \alpha\cdot Score_{PS}(w_1w_2), \end{aligned} \tag{7}$$

where $Score_{DS}(w_1w_2)$ is the predicted score when we plugin the vectors from DSMs into the regression model and $Score_{PS}(w_1w_2)$ is the predicted score when Poincaré embeddings are used as input. Thus, $Score_S$ indicates the weighted (weight = α) mixed prediction score from the supervised model. We measure the performance of our supervised model for each of the 25 random splits and report the mean and standard deviation of the performance metric.

3.3 Hyperparameters of the Model

Apart from the hyperparameters used to train the Poincaré model, our proposed model has three hyperparameters: k, m and α. k indicates the number of top hypernyms or hyponyms per target word to be used for training the Poincaré model. Since only considering hyponym-hypernym pairs containing target words does not lead to sufficient training samples for the Poincaré model, we add top m% hyponym-hypernym pairs extracted by using Hearst pattern to the training set. Note that we consider the top hyponym-hypernym pairs on the basis of normalized frequency. α indicates the relative weight between Poincaré similarity and distributional similarity. We have optimized these three hyperparameters by grid search.

4 Evaluation

4.1 Datasets

To evaluate our proposed models (both supervised and unsupervised) we use three gold standard datasets for English on compositionality detection and describe them in the following.

Reddy (RD): This dataset contains compositionality judgments for 90 compounds in a scale of literality from 0 (idiomatic) to 5 (compositional), obtained by averaging crowdsourced judgments on these pairs (Reddy et al., 2011). For evaluation, we use only the global compositionality score, ignoring individual word judgments.

Reddy++ (RD++): This is a recently introduced resource created for evaluation (Ramisch et al., 2016) that extends the Reddy dataset with an additional 90 English nominal compounds, amounting to a total of 180 nominal compounds. Consistent with RD, the scores range from 0 (idiomatic) to 5 (compositional) and are annotated through Mechanical Turk and averaged over the annotators. The additional 90 entries are adjective-noun pairs, balanced with respect to compositionality.

Farahmand (FD): This dataset contains 1042 English compounds extracted from Wikipedia with binary non-compositionality judgments by four experts (Farahmand et al., 2015). In evaluations we use the sum of all the judgments to have a single numeral compositionality score, ranging from 0 (compositional) to 4 (idiomatic).

We optimize our method on subsets of the datasets for pairs and constituents with available Poincaré embeddings in order to measure the direct impact of our method, which comprises 79, 146 and 780 datapoints for the three sets RD-R, RD++-R and FD-R, respectively.

We subsequently report scores on the full datasets RD-F (90), RD++-F (180) and FD-F (1042) for the sake of fair comparison to previous works. In cases where no Poincaré embeddings are available, we use the fallback strategy of only relying on the distributional model, i.e. $Score_{DS}$.

For the supervised setup, we experiment on the FD dataset (on the reduced version and the full version) since for the other two datasets, the number of instances are not enough for supervision.

4.2 Baselines

We use the recent work by Cordeiro et al. (2016) as the baseline, where authors apply several distributional semantic models and their variants by tuning hyperparameters like the dimension of vectors, the window-size during training and others. We resort to PPMI-SVD, two variants of word2vec (CBOW and SkipGram) and GloVe as our baselines. We use these models as provided, with the vector dimension size of 750 (PPMI-SVD, W2V) and 500 (GloVe)[2].

PPMI-SVD baseline: For each word, its neighboring nouns and verbs in a symmetric sliding window of w words in both directions, using a linear decay weighting scheme with respect to its distance d to the target (Levy et al., 2015) are extracted. The representation of a word is a vector containing the positive pointwise mutual information (PPMI) association scores between the word and its contexts. Note that, for each target word, contexts that appear less than 1000 times are discarded. The Dissect toolkit (Dinu et al., 2013) is then used in order to build a PPMI matrix and its dimensionality is reduced using singular value decomposition (SVD) to factorize the matrix.

word2vec baseline: This DSM is prepared using the well-known word2vec (Mikolov et al., 2013) in both variants CBOW (W2V-CBOW) and Skip-Gram (W2V-SG), using default configurations except for the following: no hierarchical softmax; negative sampling of 25; frequent-word downsampling weight of 10^{-6}; runs 15 training iterations; minimum word count threshold of 5.

GloVe baseline: The count-based DSM of Pennington et al. (2014), implementing a factorization of the co-occurrence count matrix is used for the task. The configurations are the default ones, except for the following: internal cutoff parameter $x_{max} = 75$; builds co-occurrence matrix in 15 iterations; minimum word count threshold of 5.

Other baseline models proposed by Reddy et al. (2011), Salehi et al. (2014), Salehi et al. (2015) report results only on Reddy dataset (since the other two datasets have been introduced later) whereas Yazdani et al. (2015) perform their evaluation only on the Farahmand dataset for their supervised model. In addition, this supervised approach requires an additional resource of $\sim 70k$ known noun phrases from Wikipedia for training. However, Cordeiro et al. (2016) compare their best models with all these baseline models and show that their models outperform across all the respective datasets. Hence we execute all our evaluations by considering only the best models proposed by Cordeiro et al. (2016) as our baselines.

[2]These pre-trained DSMs were provided by Cordeiro et al. (2016); on re-computation we get slightly different results than those reported in their paper.

4.3 Evaluation Setup

Quantitative evaluation is usually done by comparing model outcomes against the gold standard datasets. For all the three datasets (RD-R, RD++-R, FD-R), we report Spearman's rank correlation (ρ) between the scores provided by the humans and the compositionality score obtained from the models. Note that for the nominal compounds in FD-R dataset, higher human scores indicate a higher degree of idiomaticity, which is opposite to the scoring in the RD-R and RD++-R datasets. We therefore always report the absolute correlation values ($|\rho|$) for all the datasets.

5 Experimental Results

In this section, we report the results obtained from the baseline models and the unsupervised and supervised variants of our model.

5.1 Unsupervised Baseline Results

We compare the performance of the baseline models (Cordeiro et al., 2016) and Poincaré embeddings as a single signal on the reduced version of the three gold standard datasets: RD-R (79 instances), RD++-R (146 instances), FD-R (780 instances) in order to closely examine the influence of Poincaré embeddings. Table 1 shows the performance for all the baselines in terms of Spearman's rank correlation ρ. We observe that W2V-CBOW model produces the best performance across all the three datasets and W2V-SG achieves the second-best performance. As noted in the table, the Poincaré embeddings on their own perform worse than all the other baselines. Further, since our final model is based on an interpolation between Poincaré embeddings and W2V-CBOW, we also attempted interpolation between other four baseline models, but the best results were always close to the better of the two models, and are not reported here.

Base. Model	RD-R	RD++-R	FD-R
W2V-CBOW	**0.8045**	**0.6964**	**0.3405**
W2V-SG	0.8034	0.6963	0.3396
GloVe	0.7604	0.6487	0.2620
PPMI-SVD	0.7484	0.6468	0.2428
Poincaré	0.6023	0.4765	0.2007

Table 1: Baseline (Cordeiro et al., 2016) results on the reduced version of three gold-standard datasets ordered in decreasing overall performance along with the results of using only Poincaré embedding.

5.2 Results of Proposed Unsupervised Model

We report the effect of tuning hyper-parameters introduced in Section 3, e.g. k, m, or α.

Fixed k neighbours: We start by fixing $k = 5$ and obtain the correlations by varying m and α. The results are presented in Table 2. We experiment with values of m ranging from 0 to 10 and report results for $m = 0, 1, 5, 10$. Note that here $m = 0$ indicates the case where we use the Poincaré embeddings of the target word's top k hypernyms and hyponyms only with no additional highly frequent hyponym-hypernym pairs. Values of $m > 10$ degrade the quality, as too many noisy pattern extractions would be used in training.

Key observations: For certain values of α we obtain considerable improvements over the baseline Spearman's correlation when introducing Poincaré embeddings. The addition of top hyponym-hypernym pairs (i.e., $m > 0$) improves the performance of the model. Finally, note that for $m > 0$, $\alpha = 0.4$ generally produces better results across the three datasets.

m(%)	α	RD-R	RD++-R	FD-R
0	0.2	0.8160	0.7102	0.3536
	0.4	0.8117	0.7012	0.3532
	0.6	0.7844	0.6581	0.3278
1	0.2	0.8274	0.7155	0.3482
	0.4	**0.8391**	**0.7165**	**0.3373**
	0.6	0.8136	0.6817	0.3036
5	0.2	0.8362	0.7268	0.3501
	0.4	**0.8578**	**0.7389**	**0.3432**
	0.6	0.8467	0.7279	0.3126
10	0.2	0.8346	0.7250	0.3513
	0.4	**0.8421**	**0.7461**	**0.3469**
	0.6	0.8299	0.7372	0.3204

Table 2: Effect of the introduction of the Poincaré embeddings for varying values of m and α. Here W2V-CBOW is used as distributional model.

MODEL-DP with W2V-CBOW			
α	**RD-R**	**RD++-R**	**FD-R**
0.2	0.8265	0.7177	0.3594
0.4	**0.8324**	**0.7321**	**0.3646**
0.6	0.8082	0.7077	0.3450
MODEL-DP with W2V-SG			
α	**RD-R**	**RD++-R**	**FD-R**
0.2	0.8244	0.7215	0.3603
0.4	**0.8330**	**0.7337**	**0.3673**
0.6	0.8152	0.7101	0.3461

Table 3: Performance of MODEL-DP using W2V-CBOW as well as W2V-SG as distributional models: Effect of removal of top 1% hypernym-hyponym pairs from the top 10% pairs ($k = 5$).

Effect of the top m pairs: Since the extraction of the hypernyms from the corpus is completely unsupervised and based on handcrafted lexical-syntactic patterns, we investigate whether the most frequent hyponym-hypernym pairs are affecting the quality of Poincaré embeddings, having noted many erroneous extractions for very frequent pairs. We fix the value of $m = 10$, but drop the most frequent 1% hyponym-hypernym pairs and retrain the Poincaré model with the rest of the pairs. We call this variant MODEL-DP. The upper half of Table 3 shows the performance of this model while using W2V-CBOW as the distributional models ($k = 5$, which was the optimal k also in this setting). We compare the result of MODEL-DP for $\alpha = 0.4$ with Table 2, row corresponding to $m = 10\%, \alpha = 0.4$.

k	α	**RD-R**	**RD++-R**	**FD-R**
3	0.2	0.8269	0.7228	0.3563
	0.4	0.8275	0.7382	0.3557
	0.6	0.8089	0.7188	0.3278
5	0.2	0.8265	0.7177	0.3594
	0.4	**0.8324**	**0.7321**	**0.3646**
	0.6	0.8082	0.7077	0.3450
10	0.2	0.8123	0.7103	0.3534
	0.4	0.8168	0.7248	0.3589
	0.6	0.7700	0.6957	0.3484

Table 4: Results obtained for MODEL-DP ($m = 10$, top 1% hypernym-hyponym pairs removed) by varying the values of k.

Key observations: We mainly observe that discarding the most frequent 1% hyponym-hypernym pairs improves the results for the largest dataset FD-R considerably while making the results from the other two datasets a little worse. We also produce results on MODEL-DP by varying the value of k. We try with $k = 3, 5, 10$, the results of which is presented in Table 4. Clearly, $k = 5$ gives the best performance. If we consider very few hypernyms per target word, it results in lack of sufficient information for the Poincaré model, while training with too many hypernyms per target word dilutes the useful hierarchy information because it adds noise.

Other DSM models: We use W2V-CBOW as the DSM for MODEL-DP. Keeping all the other parameters of MODEL-DP the same (i.e., $m = 10$, $k = 5$, $\alpha = 0.4$) we replace the DSM by the W2V-SG vectors, which was performing the second best among the baselines. We are interested in observing whether the Poincaré embeddings also benefit other DSM models as well.

Key observations: The performance of this variant of our model is presented in the lower half of Table 3. We indeed observe the same effect of the Poincaré embeddings improving the overall performance by 3-4% on all datasets.

Other hyperparameters: In a series of experiments that we do not report in detail for brevity, we could make the following observations: For our task, the vector dimensionality of Poincaré embeddings of $d = 50$ shows better results than higher or lower values, as tested with $d \in \{20, 100\}$. Similarly, we tried with several vector dimensions of DSMs with $d \in 50, 100, 300$ but 750 gives the best performance for the best models reported by Cordeiro et al. (2016) and our model in the unsupervised setup. We further tried varying the relative weight of single word vectors for the sum in Equation 1, which did not have positive effects.

Performance for reduced dataset			
Model	**RD-R**	**RD++-R**	**FD-R**
W2V-CBOW	0.8045	0.6964	0.3405
MODEL-DP	**0.8324**	**0.7321**	**0.3646**
Performance for full dataset			
Model	**RD-F**	**RD++-F**	**FD-F**
W2V-CBOW	0.7867	0.7022	0.2688
MODEL-DP	**0.8095**	**0.7302**	**0.2958**

Table 5: Performance of our model (MODEL-DP) and most competitive baseline (W2V-CBOW) for both the reduced datasets and the whole datasets (using the fallback strategy).

Fallback strategy to encompass the whole dataset: In all the above experiments we consider the reduced version of the three gold-standard datasets due to lack of the Poincaré embeddings for certain target words. We suggest a fallback strategy to incorporate the target words that do not have Poincaré embeddings. In cases where the Poincaré embeddings are not present, we fall back to the distributional similarity score. In cases, where the Poincaré embeddings are available we use the combined score as discussed in Section 3. Note that, the distributions of distributional similarity scores and proposed combined scores are significantly different (according to the z-test (Fisher, 1932)). Therefore while falling back to the distributional similarity scores we scale up the scores by the proportion of normalized means of the two distributions.

Key observations: The results for this fall back strategy is noted in the lower half of Table 5. We observe that for all three datasets we perform significantly better than the baselines. To be consistent with the literature, we compare our performance even with the supervised model proposed by Yazdani et al. (2015) for the FD-F dataset. For this dataset, the supervised model proposed by the authors produces a Spearman's rank correlation (ρ) of 0.41 whereas the unsupervised MODEL-DP produces 0.29. However, our supervised approach, as we shall see later, beats this number reported by Yazdani et al. (2015) by a considerable margin.

Significance test: From the extensive evaluation of our model by tuning several hyper-parameters, we obtain MODEL-DP (Table 3), which gives the best performance for all the three datasets outperforming the baselines (Table 1). We perform Wilcoxon's sign-rank test (Rey and Neuhäuser, 2011) for all the three datasets separately. We obtain $p < 0.05$ while comparing MODEL-DP and the best baseline model (W2V-CBOW) indicating that the difference between their compositionality predictions is statistically significant.

Error analysis: We investigate the erroneous cases for which the annotators give a high compositionality score while our model produces a very low compositional score, e.g. 'area director', 'discussion page', and 'emergency transportation'. We observe that the number of hypernyms extracted for these target noun phrases is very low (1 or 2), which leads to a less informative hierarchical representation in the Poincaré model; this is either caused by a low frequency of terms overall, or by a low occurrence in hypernym pattern contexts. We also analyzed the non-compositional cases for which the annotators give a low compositionality score but our model produces a high score, e.g. 'hard disk', 'hard drive' and 'soft drink'. In these cases even though they are non-compositional, the hypernyms of the noun phrases match with the hypernyms of the head constituent words. For example, 'hard disk' and 'disk' have the same hypernym 'storage device'; similarly 'soft drink' and 'drink' have 'product'; 'hard drive' and 'drive' have 'device'. Thus, these non-compositional cases are different from entirely opaque expressions like 'couch potato', 'hot dog' where none of the hypernyms of the noun phrases match with the hypernyms of any of the constituent words. Categorizing the non-compositional words based on the above observation and dealing with such cases is left for future work.

Model	**RD-RL**	**RD++-RL**	**FD-RL**
W2V-CBOW	0.8111	0.7256	0.4198
MODEL-DP-L	0.8223	0.7451	0.4179
MODEL-DP	**0.8288**	**0.7592**	**0.4790**

Table 6: Comparisons of the results produced by MODEL-DP-L from lexical resources vs. MODEL-DP along with the baselines for the reduced dataset.

Training using lexical resources: We further investigated the use of hyponym-hypernym pairs extracted from lexical resources like WordNet (Miller, 1995) or ConceptNet (Speer et al., 2017) for training the Poincaré model. Even though the quality of the hyponym-hypernym pairs from lexical resources is better compared to the pairs extracted using Hearst patterns, the coverage of target words is very low. Therefore, for a fair comparison, we prepare a reduced version of the three gold standard datasets (RD-RL, RD++-RL, FD-RL), where all the target words are present in lexical resources as well as hyponym-hypernym pairs extracted using Hearst patterns. RD-RL, RD++-RL, and FD-RL contain 74, 131, 380 target words, respectively. MODEL-DP-L uses the same compositionality score metric as MODEL-DP but in the case of MODEL-DP-L, the Poincaré embedding is learned using the hyponym-hypernym pairs extracted only from WordNet and ConceptNet combined. The results are presented in Table 6. We see that even though MODEL-DP-L performs better than the baselines for two of the datasets, MODEL-DP gives the best result. We attribute this to the relative sparsity of lexical resources, which are seemingly not sufficient for training reliable Poincaré embeddings.

5.3 Results of Proposed Supervised Model

For the supervised setup we present our results on the reduced FD-R dataset (780 instances) and the full Farhamand FD-F dataset (1042 instances). We do not use the other two datasets for the supervised setup since the number of instances in both these datasets are too small to produce a reasonable training-test split required for supervision.

As discussed in Section 3.2, we use various regression models; 75% of the dataset is used for training and the remaining 25% is used for testing; we experiment on 25 such random splits and

FD-R				
	Kernel Regression		PLS Regression	
	$\mu(\lvert\rho\rvert)$	$\sigma(\lvert\rho\rvert)$	$\mu(\lvert\rho\rvert)$	$\sigma(\lvert\rho\rvert)$
CBOW-S (750)	0.4017	0.0599	0.3972	0.0590
α	MODEL-DP-S			
0.2	0.4294	0.0591	0.4078	0.0566
0.4	**0.4347**	**0.0563**	**0.4096**	**0.0525**
0.6	0.4221	0.0540	0.3959	0.0497
CBOW-S (50)	0.4339	0.0570	0.4227	0.0584
α	MODEL-DP-S, CBOW vectors of dim. 50			
0.2	0.4487	0.0547	0.4361	0.0561
0.4	**0.4520**	**0.0528**	**0.4372**	**0.0518**
0.6	0.4410	0.0510	0.4196	0.0491
FD-F				
	Kernel Regression		PLS Regression	
	$\mu(\lvert\rho\rvert)$	$\sigma(\lvert\rho\rvert)$	$\mu(\lvert\rho\rvert)$	$\sigma(\lvert\rho\rvert)$
CBOW-S (750)	0.3822	0.0471	0.3910	0.0434
α	MODEL-DP-S			
0.2	0.4030	0.0446	0.3984	0.0450
0.4	**0.4083**	**0.0425**	**0.3941**	**0.0459**
0.6	0.3986	0.0418	0.3747	0.0471
CBOW-S (50)	0.4212	0.0502	0.4201	0.0470
α	MODEL-DP-S, CBOW vectors of dim. 50			
0.2	0.4329	0.0500	0.4270	0.0467
0.4	**0.4340**	**0.0488**	**0.4211**	**0.0469**
0.6	0.4213	0.0478	0.3943	0.0499

Table 7: Mean (μ) and Standard Deviation (σ) of Spearman's rank correlation (ρ) of the supervised approach for FD-R and FD-F datasets over 25 random splits. We compare best baseline model (CBOW - 750 and 50 dimension) and our model (MODEL-DP-S) using both 750 and 50 dimension of CBOW vectors.

report mean and standard deviation of Spearman's rank correlation (ρ). Among all the regression models (respective to the best choice of the hyperparameters), Kernel Ridge regression gives the best performance while PLS regression is the second best for both the FD-R and FD-F dataset. We compare the performance of the best baseline supervised model (CBOW-S) where only $Score_{DS}$ from Equation 7 is used as the predicted score with our proposed supervised model (MODEL-DPS) where $Score_S$ from Equation 7 is used as the predicted score. The performance of these two best regression models for the baseline and our model (for $\alpha = 0.4$)[3] are noted in Table 7. In the same table, we also report the results of the evaluation on FD-F dataset using a fallback strategy for the supervised setup: here, we use a 50-dimensional zero vector of the target word or compound for which the Poincaré embedding is absent. We observe that for both the datasets (reduced and full) our approach outperforms the baseline results by a large margin. As discussed earlier, the CBOW vectors used for experiments consist of 750 dimensions. Since the number of data points in the training set is small, we also experiment with CBOW vector dimension of 50 (MODEL-DPS-50) in the supervised setup to avoid overfitting due to a large number of parameters. The results presented in Table 7 show that with the reduced number of dimensions, our model yields even better results and outperforms the correlations 0.41 and 0.34 reported respectively in (Yazdani et al., 2015) and (Cordeiro et al., 2016).

6 Conclusion

In this paper, we present a novel straightforward method for estimating degrees of compositionality in noun phrases. The method is mixing hypernymy and distributional information of the noun phrases and their constituent words. To encode hypernymy information, we use Poincaré embeddings, which – to the best of our knowledge – are used for the first time to accomplish the task of compositionality prediction. While these hyperbolic embeddings trained on hypernym pattern extractions are not a good signal on their own for this task, we observe that mixing distributional and hypernymy information via Euclidean and hyperbolic embeddings helps to substantially and significantly improve the performance of compositionality prediction, outperforming previous state-of-the-art models. Our pretrained embeddings and the source codes are publicly available.[4]

Two directions for future work are (i) to extend our approach to other languages by using multilingual resources or translation data; and (ii) to explore various compositionality functions to combine the words' representation on the basis of their grammatical function within a phrase.

Acknowledgments

We acknowledge the support of the DFG under the "JOIN-T" (BI 1544/4) and "ACQuA" (BI 1544/7) projects, Humboldt Foundation for providing scholarship as well as the DAAD and the Indian Department of Science and Technology via a DAAD-DST PPP grant.

[3] $\alpha = 0.4$ produces the best results per grid search.

[4] https://github.com/uhh-lt/poincare

References

Hervé Abdi. 2007. Partial least squares regression. *Encyclopedia of measurement and statistics*, 2:740–744.

Naomi S. Altman. 1992. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185.

Rami Aly, Alexander Ossa, Arne Köhn, Chris Biemann, and Alexander Panchenko. 2019. Every child should have parents: a taxonomy refinement algorithm based on hyperbolic term embeddings. In *Proceedings of the 57th Annual Meeting of the Association of Computational Linguistics (Volume 2: Short Papers)*, Florence, Italy.

Timothy Baldwin, Colin Bannard, Takaaki Tanaka, and Dominic Widdows. 2003. An empirical model of multiword expression decomposability. In *Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment*, pages 89–96, Sapporo, Japan.

Chris Biemann and Eugenie Giesbrecht. 2011. Distributional semantics and compositionality 2011: Shared task description and results. In *Proceedings of the Workshop on Distributional Semantics and Compositionality*, pages 21–28, Portland, OR, USA.

Yuri Bizzoni, Marco S. G. Senaldi, and Alessandro Lenci. 2018. Finding the neural net: Deep-learning idiom type identification from distributional vectors. *Italian Journal of Computational Linguistics*, 4(1):27–41.

Yaacov Choueka. 1988. Looking for needles in a haystack or locating interesting collocational expressions in large textual databases. In *RIAO 88:(Recherche d'Information Assistée par Ordinateur). Conference*, pages 609–623, Cambridge, MA, USA.

Paul Cook, Afsaneh Fazly, and Suzanne Stevenson. 2007. Pulling their weight: Exploiting syntactic forms for the automatic identification of idiomatic expressions in context. In *Proceedings of the workshop on a broader perspective on multiword expressions*, pages 41–48, Prague, Czech Republic.

Silvio Cordeiro, Carlos Ramisch, Marco Idiart, and Aline Villavicencio. 2016. Predicting the compositionality of nominal compounds: Giving word embeddings a hard time. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1986–1997, Berlin, Germany.

Silvio Cordeiro, Aline Villavicencio, Marco Idiart, and Carlos Ramisch. 2019. Unsupervised compositionality prediction of nominal compounds. *Computational Linguistics*, 45(1):1–57.

Georgiana Dinu, Nghia The Pham, and Marco Baroni. 2013. Dissect - distributional semantics composition toolkit. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 31–36, Sofia, Bulgaria.

Harris Drucker, Christopher J. C. Burges, Linda Kaufman, Alex J. Smola, and Vladimir Vapnik. 1997. Support vector regression machines. In *Advances in Neural Information Processing Systems 9*, pages 155–161, Denver, CO, USA.

Meghdad Farahmand, Aaron Smith, and Joakim Nivre. 2015. A multiword expression data set: Annotating non-compositionality and conventionalization for English noun compounds. In *Proceedings of the 11th Workshop on Multiword Expressions*, pages 29–33, Denver, CO, USA.

Adriano Ferraresi, Eros Zanchetta, Marco Baroni, and Silvia Bernardini. 2008. Introducing and evaluating ukWaC, a very large web-derived corpus of English. In *Proceedings of the 4th Web as Corpus Workshop (WAC-4) Can we beat Google*, pages 47–54, Marrakech, Morocco.

Ronald A. Fisher. 1932. *Statistical methods for research workers*. Oliver and Boyd, Edinburgh.

Michael Flor and Beata Beigman Klebanov. 2018. Catching idiomatic expressions in EFL essays. In *Proceedings of the Workshop on Figurative Language Processing*, pages 34–44, New Orleans, LA, USA.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the Leipzig Corpora Collection: From 100 to 200 languages. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 759–765, Istanbul, Turkey.

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th Conference on Computational Linguistics - Volume 2*, COLING '92, pages 539–545, Nantes, France.

Valentin Khrulkov, Leyla Mirvakhabova, Evgeniya Ustinova, Ivan Oseledets, and Victor Lempitsky. 2019. Hyperbolic image embeddings. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA.

Douwe Kiela and Stephen Clark. 2013. Detecting compositionality of multi-word expressions using nearest neighbours in vector space models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1427–1432, Seattle, WA, USA.

Matt Le, Stephen Roller, Laetitia Papaxanthos, Douwe Kiela, and Maximilian Nickel. 2019. Inferring concept hierarchies from text corpora via hyperbolic embeddings. *arXiv preprint arXiv:1902.00913*.

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

Christina Lioma, Jakob G. Simonsen, Birger Larsen, and Niels D. Hansen. 2015. Non-compositional term dependence for information retrieval. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 595–604, Santiago, Chile.

Diana McCarthy, Bill Keller, and John Carroll. 2003. Detecting a continuum of compositionality in phrasal verbs. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 73–80, Sapporo, Japan.

Diana McCarthy, Sriram Venkatapathy, and Aravind K. Joshi. 2007. Detecting compositionality of verb-object combinations using selectional preferences. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 369–379, Prague, Czech Republic.

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

George A. Miller. 1995. WordNet: A lexical database for English. *Communications of the ACM*, 38(11):39–41.

Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In *Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL-08: HLT)*, pages 236–244, Columbus, OH, USA.

Jeff Mitchell and Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive science*, 34(8):1388–1429.

Maximillian Nickel and Douwe Kiela. 2017. Poincaré embeddings for learning hierarchical representations. In *Advances in Neural Information Processing Systems 30*, pages 6338–6347, Long Tail Beach, CA, USA.

Alexander Panchenko, Stefano Faralli, Eugen Ruppert, Steffen Remus, Hubert Naets, Cédrick Fairon, Simone P. Ponzetto, and Chris Biemann. 2016. TAXI at SemEval-2016 Task 13: a taxonomy induction method based on lexico-syntactic patterns, substrings and focused crawling. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1320–1327, San Diego, CA, USA.

Alexander Panchenko, Olga Morozova, and Hubert Naets. 2012. A semantic similarity measure based on lexico-syntactic patterns. In *KONVENS*, pages 174–178, Vienna, Austria.

Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2009. English gigaword forth edition. In *Linguistic Data Consortium*, Philadelphia, PA, USA.

Jing Peng, Katsiaryna Aharodnik, and Anna Feldman. 2018. A distributional semantics model for idiom detection - the case of english and russian. In *Proceedings of the 10th International Conference on Agents and Artificial Intelligence, ICAART 2018, Volume 2*, pages 675–682, Funchal, Madeira, Portugal.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar.

Carlos Ramisch, Silvio Cordeiro, Leonardo Zilio, Marco Idiart, and Aline Villavicencio. 2016. How naked is the naked truth? a multilingual lexicon of nominal compound compositionality. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 156–161, Berlin, Germany.

Siva Reddy, Diana McCarthy, and Suresh Manandhar. 2011. An empirical study on compositionality in compound nouns. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*, pages 210–218, Chiang Mai, Thailand.

Denise Rey and Markus Neuhäuser. 2011. Wilcoxon-signed-rank test. In Miodrag Lovric, editor, *International Encyclopedia of Statistical Science*, pages 1658–1659. Springer, Berlin, Heidelberg.

Bahar Salehi, Paul Cook, and Timothy Baldwin. 2014. Using distributional similarity of multi-way translations to predict multiword expression compositionality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 472–481, Gothenburg, Sweden.

Bahar Salehi, Paul Cook, and Timothy Baldwin. 2015. A word embedding approach to predicting the compositionality of multiword expressions. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 977–983, Denver, CO, USA.

Richard Socher, Brody Huval, Christopher D. Manning, and Andrew Y. Ng. 2012. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and*

Computational Natural Language Learning, pages 1201–1211, Jeju Island, Korea.

Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, pages 4444–4451, San Francisco, CA, USA.

Stephen Tratz and Eduard Hovy. 2010. ISI: Automatic classification of relations between nominals using a maximum entropy classifier. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 222–225, Uppsala, Sweden.

Sriram Venkatapathy and Aravind K. Joshi. 2005. Measuring the relative compositionality of verb-noun (V-N) collocations by integrating features. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 899–906, Vancouver, BC, Canada.

Vladimir Vovk. 2013. Kernel ridge regression. In *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik*, pages 105–116. Springer.

Majid Yazdani, Meghdad Farahmand, and James Henderson. 2015. Learning semantic composition to detect non-compositionality of multiword expressions. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1733–1742, Lisbon, Portugal.