Cross-lingual Word Embeddings in Hyperbolic Space

Anonymous ACL submission

Abstract

Cross-lingual word embeddings can be applied to several natural language processing applications across multiple languages. Unlike prior works that use word embeddings based on the Euclidean space, this short paper presents a simple and effective cross-lingual Word2Vec model that adapts to the Poincaré ball model of hyperbolic space to learn unsupervised cross-lingual word representations from a German-English parallel corpus. It has been shown that hyperbolic embeddings can capture and preserve hierarchical relationships. We evaluate the model on both hypernymy and analogy tasks. The proposed model achieves comparable performance with the vanilla Word2Vec model on the cross-lingual analogy task, the hypernymy task shows that the cross-lingual Poincaré Word2Vec model can capture latent hierarchical structure from free text across languages, which are absent from the Euclidean-based Word2Vec representations. Our results show that by preserving the latent hierarchical information, hyperbolic spaces can offer better representations for cross-lingual embeddings.

1 Introduction

In Natural Language Processing (NLP), cross-lingual word embeddings refer to the representations of words from two or more languages in a joint feature space. Prior works have demonstrated the use of these continuous representations in a variety of NLP tasks such as information retrieval (Zoph et al., 2016), semantic textual similarity (Cer et al., 2017), knowledge transfer (Gu et al., 2018), lexical analysis (Dong and De Melo, 2018), plagiarism detection (Alzahrani and Aljuaid, 2020), etc. across different languages.

Natural language data possesses latent tree-like hierarchies in linguistic ontologies (e.g., hypernyms, hyponyms) (Dhingra et al., 2018; Asteefanoaei and Collignon, 2020) such as the taxonomy of WordNet (Miller, 1998) for a language. From the statistics of word co-occurrence in training text, word embeddings models in Euclidean space can capture associations of words and their semantic relatedness. However, they fail to capture asymmetric word relations, including the latent hierarchical structure of words such as specificity (Dhingra et al., 2018). For example, ‘bulldog’ is more specific than ‘dog.’ The use of non-Euclidean spaces has recently been advocated as alternatives to the conventional Euclidean space to infer latent hierarchy from the language data (Nickel and Kiela, 2017, 2018; Dhingra et al., 2018; Tifrea et al., 2018). Learning cross-lingual hierarchies such as cross-lingual types-sub types and hypernyms-hyponyms, is useful for tasks like cross-lingual lexical entailment, textual entailment, machine translation, etc. (Vulić et al., 2019).

This paper builds upon previous work in monolingual hyperbolic Word2Vec\(^1\) modeling from Tifrea et al. (2018) by learning cross-lingual hyperbolic embeddings from a parallel corpus. As a first step, we adopt the German-English parallel corpus from Wolk and Marasek (2014). We summarize the main contributions as follows: (1) To the best of our knowledge, we are the first to attempt at learning cross-lingual embeddings of natural language data using non-Euclidean geometry; (2) we evaluate the hyperbolic embeddings on cross-lingual HyperLex hypernym task to evaluate its performance in learning latent hierarchies from free text and how a word’s specificity correlates to its embedding’s norm. We also compare the hyperbolic Word2Vec embeddings with the vanilla Word2Vec embeddings in the cross lingual analogy task.

2 Related Work

2.1 Cross-lingual Word Embeddings

Cross-lingual word representations have been a subject of extensive research (Upadhyay et al., 2016;\(^1\)The hyperbolic Word2Vec model is not described in Tifrea et al. (2018)’s paper, but available in the corresponding codebase)
Ruder et al., 2019). Recent advances in the field can be grouped into unsupervised, supervised, and joint learning algorithms. Unsupervised models (Lample et al., 2017; Artetxe and Schwenk, 2019; Chen et al., 2018) exploit existing monolingual word embeddings, followed by various cross-lingual alignment procedures. Supervised models (Mikolov et al., 2013; Smith et al., 2017; Grave et al., 2018) learn a mapping function from a source embedding space to the target embedding space based on different objective criteria. Joint learning models (Coulmance et al., 2015; Josifoski et al., 2019; Sabet et al., 2019; Lachraf et al., 2019) use parallel corpora to train bilingual embeddings in the same space jointly. This work adopts the settings similar to the joint learning model for embedding alignments by Lachraf et al. (2019).

2.2 Hyperbolic Word Embeddings

Hyperbolic spaces offer a continuous representation for embedding tree-like structures with arbitrarily low distortion (Sala et al., 2018; Chami et al., 2020). Word embeddings in hyperbolic spaces have been applied to diverse NLP applications such as text classification (Zhu et al., 2020), learning taxonomy (Asteananaei and Collignon, 2020), and concept hierarchy (Le et al., 2019). By using hyperbolic space these applications were able to outperform their euclidean counterparts by exploiting the benefits of hierarchical structure of the text data with high quality embedding which capture similarity and generality of concept together enforce transitivity of the is-a-relations in a smaller embedding space (Le et al., 2019). Some recent work use supervised models (Nickel and Kiela, 2017, 2018; Ganea et al., 2018) that require external information on word relations such as WordNet or ConceptNet in addition to free text corpora to learn word and sentence embeddings in the hyperbolic space. Nickel and Kiela (2017) consider a non-parametric method to learn hierarchical representation from a lookup table for symbolic data. Ganea et al. (2018) propose a supervised method to learn embeddings for an acyclic graph structure of words. Unsupervised word embedding models (Leimeister and Wilson, 2018; Dhingra et al., 2018; Tifrea et al., 2018) which can directly learn from text corpora have been recently applied in the hyperbolic spaces. Leimeister and Wilson (2018) employ the skip-gram with negative sampling architecture of the Word2Vec model for learning word embeddings from free text. Dhingra et al. (2018) present a two-step model to embed a co-occurrence graph of words and map the output of the encoder to the Poincaré ball using the algorithm from Nickel and Kiela (2017). Tifrea et al. (2018) remodel the GloVe algorithm to learn unsupervised word representation in hyperbolic spaces.

3 Methodology

3.1 Hyperbolic Space

Hyperbolic space in Riemannian geometry is a homogeneous space of constant negative curvature with special geometric properties. Hyperbolic space can endow infinite trees to have nearly isometric embeddings. We embed words using the Poincaré ball model of the hyperbolic space.

The Poincaré Ball. The Poincaré ball model $\mathbb{B}^n$ of n-dimensional hyperbolic geometry is a manifold equipped with a Riemannian metric $g^B$. Formally, an n-dimensional Poincaré unit ball is defined as $(\mathbb{B}^n, g^B)$ and the metric $g^B$ is conformal to the Euclidean metric $g^E$ as $g^B = \lambda_x^2 g^E$. Where $\lambda_x = \frac{2}{1-||x||^2}$, $x \in \mathbb{B}^n$, and $||.||$ stands for the Euclidean norm. Notably, the hyperbolic distance $d_{\mathbb{B}^n}$ between n-dimensional points $(x, y) \in \mathbb{B}^n$ in the Poincaré ball is defined as:

$$d_{\mathbb{B}^n}(x, y) = \text{arcosh} \left(1 + \frac{||x - y||^2}{(1-||x||^2)(1-||y||^2)}\right)$$

where $\text{arcosh}(w) = \ln(w + \sqrt{w^2 - 1})$ is the inverse of hyperbolic cosine function. Using ambient Euclidean geometry, the geodesic distance between points $(x, y)$ can be induced using Equation (1) as $d_{\mathbb{B}^n}(x, y) = \text{arcosh} \left(1 + \frac{1}{2\lambda_x \lambda_y}||x - y||^2\right)$. This indicates that the distance changes evenly w.r.t. $||x||$ and $||y||$, which is a key point to learning continuous representation for hierarchical structures (Chen et al., 2020).

3.2 Hyperbolic Cross-lingual Word Embedding

We first adopt the mono-lingual hyperbolic word embedding from a model defined in the work by Tifrea et al. (2018). We extend it to cross-lingual hyperbolic word embedding by using parallel text corpora input to capture word relationships through bilingual word co-occurrence statistics. Tifrea et al. (2018) added a hyperparameter function $h$ on the distance between word and context pairs in the hyperbolic Word2Vec’s objective.
function. Hence, the effective distance function in
the objective function becomes $h(d_{GB}(x, y))$.

Hyperbolic word embeddings have shown to
embed general words near the origin and specific
words towards the edges — we attempt to exploit
this property to identify latent hierarchies and in
hypernym evaluation task by using the Poincaré
norms of the words to determine their hierarchy as
words with higher norm will be more specific, i.e.,
lower in hierarchy (Nickel and Kiela, 2017; Dhingra et al., 2018; Linzhao et al., 2020). We evaluate
the hyperbolic model on the cross-lingual analogy
task to compare it with its Euclidean counterpart.

### 3.3 Cross-lingual Alignment

To train the cross-lingual Word2Vec model in the
hyperbolic space, we perform a pre-processing step
of word-to-word alignment as defined by Lachraf et al. (2019) using parallel sentences from a biling-
ual parallel corpus. We generate word-to-word
alignment by matching the indices of tokens from
both languages in parallel sentences.

### 3.4 Evaluation Methodology

#### Hypernymy Evaluation.

We perform hypernymy
evaluation to assess performance of the proposed
model based on learning the latent hierarchical
structure from free text. In the hypernymy eval-
uation task, given a word pair $(u, v)$, we evaluate
is-a $(u, v)$ i.e., to what degree $u$ is of type $v$.

For English, German and cross-lingual German-
English hypernymy evaluation, we use the Hyper-
Lex benchmark Vulić et al. (2017, 2019), which
contains word pairs $(u, v)$ and a corresponding
degree to which $u$ is of type $v$ i.e. the is-
a score. This score has been obtained by hu-
man annotators, scored by the degree of typi-
cality and semantic category membership (Vulić et al., 2017). For example, in the HyperLex
dataset, is-a(chemistry, science) = 6.00 and is-
a(chemistry, knife) = 0.50 as chemistry is a
type of science but not a type of knife.

To generate the is-a score we follow the same
approach as used by Nickel and Kiela (2017):

$$is-a(u, v) = -(1 + \alpha (||v|| - ||u||))d_{GB}(u, v) \quad (2)$$

The evaluation is performed by calculating the
Spearman correlation between the ground-truth
score and the predicted score. Note that our model
is not trained on any hypernymy detection task but
tries to learn latent hierarchy from free text.

<table>
<thead>
<tr>
<th>Word</th>
<th>Closest Children</th>
</tr>
</thead>
<tbody>
<tr>
<td>Species</td>
<td>arten, gattung, subspecies, unterfamilie</td>
</tr>
<tr>
<td>Physics</td>
<td>astrophysik, astrophysics, mechanik</td>
</tr>
<tr>
<td>Molekülen</td>
<td>atomen, protonen, elektronen, ionen</td>
</tr>
<tr>
<td>Orchestra</td>
<td>symphony, philharmonic, concerto</td>
</tr>
<tr>
<td>Regierung</td>
<td>governments, regierungen, bundesregierung</td>
</tr>
</tbody>
</table>

Table 1: For a given word in the left column, this table
shows the top closest children using a 100Dim with bias
hyperbolic Word2Vec model. Note that the children
consist of both English and German words.

<table>
<thead>
<tr>
<th>Hyperbolic Model</th>
<th>English Rank</th>
<th>German Rank</th>
<th>Cross Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>100D w/ bias</td>
<td>0.166</td>
<td>0.130</td>
<td>0.150</td>
</tr>
<tr>
<td>100D</td>
<td>0.175</td>
<td>0.104</td>
<td>0.162</td>
</tr>
<tr>
<td>120D w/ bias</td>
<td>0.192</td>
<td>0.120</td>
<td>0.179</td>
</tr>
<tr>
<td>300D w/ bias</td>
<td>0.183</td>
<td>0.125</td>
<td>0.155</td>
</tr>
</tbody>
</table>

Table 2: Spearman correlations from different hyper-
boric Word2Vec models on the English, German and
German-English HyperLex dataset for hypernymy eval-
uation. Best results are in bold.

#### Cross-lingual Analogy Evaluation.

The anal-
ogy evaluation task is one of the standard intrinsic
evaluations for word embeddings. In cross-lingual
analogy evaluation task, given a word pair $(w_1$, $w_2)$
in one language, and a word $w_3$ in the other lan-
guage, the goal is to predict the word $w_4^*$ such
that $w_4^*$ is related to $w_3$ same way $w_2$ is related
to $w_1$. For example, as prince ($w_1$) is to princess
($w_2$), prinz ($w_3$; German equivalent for prince) is
to prinzzessin ($w_4^*$; German equivalent for princess).

For evaluating cross-lingual analogy for the Ger-
man and English language, we use the cross-lingual
analogy dataset provided by Brychcin et al. (2018).

### 4 Experiments & Results

#### 4.1 Dataset

This paper uses the Wikipedia corpus of parallel
sentences extracted by Wolk and Marasek (2014)
to train the model. The dataset is accessed through
OPUS (Tiedemann, 2012). The corpus consists of
~2.5 million parallel aligned German-English
sentence pairs with 43.5 million German tokens
and 58.4 million English tokens.
We reference Tifrea et al. (2018)’s Poincaré implementation and extended it to learn cross-lingual word embeddings. We set the minimum frequency of words in the vocabulary to 100, and a window size of 5. The models use Negative-Log-Likelihood loss. The non-hyperbolic vanilla Word2Vec uses Stochastic Gradient Descent optimizer (Bonnabel, 2013). For hyperbolic embeddings, the hyperparameter $h$ is set to $\cosh^2(x)$. During the analogy evaluation, we use the cosine distance instead of Poincaré distance for hyperbolic models. We use the hypernymysuite for hypernymy evaluation (Roller et al., 2018).

### 4.2 Experimental Settings

We reference Tifrea et al. (2018)’s Poincaré Word2Vec implementation and extended it to learn cross-lingual word embeddings. The models use Negative-Log-Likelihood loss. The non-hyperbolic vanilla Word2Vec uses Stochastic Gradient Descent optimizer, whereas hyperbolic Word2Vec uses Weighted Full Riemannian Stochastic Gradient Descent optimizer (Bonnabel, 2013). For hyperbolic embeddings, the hyperparameter $h$ is set to $\cosh^2(x)$. During the analogy evaluation, we use the cosine distance instead of Poincaré distance for hyperbolic models. We use the hypernymysuite for hypernymy evaluation (Roller et al., 2018).

### 4.3 Evaluation Results

#### Hypernymy Evaluation

We present the top closest children of selected words in Table 1. As described in Section 3.2, the closest children are calculated by finding the target word’s ($t$) nearest neighbors ($N$) and extracting the neighbour $n \in N$ such that $||n||_p > ||t||_p$, where $||.||_p$ is the Poincaré norm. We observe that the model is able to find the hyponyms of the words using the closest children across languages. For example, the children of ‘Physics’ are its subtypes – ‘astrophysik’ (astrophysics), ‘mechanik’ (mechanics), and ‘biophysik’.

Table 2 reports the results on the hypernymy evaluation task. Although the models were not trained on hypernymy tasks, we observe that they could still learn some latent hierarchies from the free text across languages. Word pairs with out-of-vocabulary words were ignored during evaluation. Table 3 shows lists of related words in order of increasing hyperbolic norm and specificity, similar to Dhingra et al. (2018)’s evaluation. We show counts of these words in the corpus. Higher the count, more generic the word, and has a smaller hyperbolic norm. The Spearman correlation between $1/f$, where $f$ is the frequency of a word in the corpus, and its embedding’s hyperbolic norm is 0.747 using a 300D w/bias Poincaré model.

#### Cross-lingual Analogy Evaluation

Table 4 reports the results on the cross-lingual analogy task. We observe that for 20D models, hyperbolic model outperformed the vanilla model. For higher dimension models, hyperbolic Word2Vec performed on par with its Euclidean counterpart. Similar to hypernymy evaluation, analogy pairs with out-of-vocabulary words were ignored during evaluation.

### 5 Conclusion and Future Work

This work adapts a monolingual hyperbolic Word2Vec model and extend to cross-lingual embeddings. We observe that the hyperbolic Word2Vec embeddings are competent on cross-lingual analogy task. The hypernymy evaluation show that it also captures some latent hierarchies across languages without being trained on a hypernymy task. Future work will include extrinsic evaluation of hyperbolic cross-lingual word embeddings on downstream tasks such as machine translation, cross-lingual textual entailment detection, cross-lingual taxonomy learning, etc.

---

**Table 3:** Words in order of increasing hyperbolic norm which are related to the word indicated in the top row along with their counts in the corpus. General words have a lower norm and specific words have a higher norm.

<table>
<thead>
<tr>
<th>Word</th>
<th>Count</th>
<th>Norm</th>
<th>Word</th>
<th>Count</th>
<th>Norm</th>
<th>Word</th>
<th>Count</th>
<th>Norm</th>
<th>Word</th>
<th>Count</th>
<th>Norm</th>
</tr>
</thead>
<tbody>
<tr>
<td>music</td>
<td>33167</td>
<td>0.607</td>
<td>art</td>
<td>28551</td>
<td>0.606</td>
<td>film</td>
<td>61682</td>
<td>0.606</td>
<td>chemistry</td>
<td>3165</td>
<td>0.628</td>
</tr>
<tr>
<td>musik</td>
<td>10637</td>
<td>0.608</td>
<td>arts</td>
<td>13888</td>
<td>0.623</td>
<td>films</td>
<td>7185</td>
<td>0.607</td>
<td>chemie</td>
<td>2530</td>
<td>0.629</td>
</tr>
<tr>
<td>musical</td>
<td>6585</td>
<td>0.612</td>
<td>design</td>
<td>11558</td>
<td>0.624</td>
<td>drama</td>
<td>4948</td>
<td>0.617</td>
<td>chemiker</td>
<td>908</td>
<td>0.620</td>
</tr>
<tr>
<td>musicians</td>
<td>1955</td>
<td>0.628</td>
<td>comedy</td>
<td>3937</td>
<td>0.630</td>
<td>chemisches</td>
<td>628</td>
<td>0.647</td>
<td>organischen</td>
<td>344</td>
<td>0.651</td>
</tr>
<tr>
<td>filmmusik</td>
<td>278</td>
<td>0.640</td>
<td>kunstgalerie</td>
<td>102</td>
<td>0.665</td>
<td>stummfilm</td>
<td>179</td>
<td>0.648</td>
<td>orga</td>
<td>344</td>
<td>0.651</td>
</tr>
</tbody>
</table>

**Table 4:** Accuracy on the cross-lingual analogy task.

<table>
<thead>
<tr>
<th>Model Type</th>
<th>Dim</th>
<th>Bias term</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vanilla</td>
<td>20D</td>
<td>✗</td>
<td>16.8</td>
</tr>
<tr>
<td>Poincaré</td>
<td>20D</td>
<td>✗</td>
<td>20.5</td>
</tr>
<tr>
<td>Vanilla</td>
<td>40D</td>
<td>✗</td>
<td>25.4</td>
</tr>
<tr>
<td>Poincaré</td>
<td>40D</td>
<td>✗</td>
<td>26.5</td>
</tr>
<tr>
<td>Vanilla</td>
<td>80D</td>
<td>✗</td>
<td>30.8</td>
</tr>
<tr>
<td>Poincaré</td>
<td>80D</td>
<td>✗</td>
<td>28.7</td>
</tr>
<tr>
<td>Vanilla</td>
<td>180D</td>
<td>✓</td>
<td>36.1</td>
</tr>
<tr>
<td>Poincaré</td>
<td>180D</td>
<td>✓</td>
<td>29.3</td>
</tr>
</tbody>
</table>

---

1https://github.com/alex-tifrea/poincare_glove
2https://github.com/facebookresearch/hypernymysuite

References


