Exploring Supervised Learning of Hierarchical Event Embedding with Poincaré Embeddings

Pride Kavumba† Naoya Inoue†,‡ Kentaro Inui†,‡

†Tohoku University ‡RIKEN Center for Advanced Intelligence Project (AIP)
kavumba.pride.q2@dc.tohoku.ac.jp
{inoue, inui}@ecei.tohoku.ac.jp

1 Introduction

Understanding events expressed in text is important in many natural language understanding tasks such as dialogue systems, question answering, discourse understanding, and information extraction. Natural language sentences and events exhibit hierarchical structure [8, 1]. This hierarchy can be defined in terms of specificity. General events can be considered as parent events to the more specific events. For instance, the event "*person eats food*" can be considered as the parent event to "*John ate an apple*" (see Figure 1 for more examples).

Capturing this kind of hierarchy is important in many applications such as causality recognition. For example, if we have the hierarchy for these events; "eat something", "eat food", and "eat apple". It is enough to only know that *eat something* causes *someone to be full*.

Previous work in event understanding [18, 23, 17, 9, 14, 5] use distributed representational of events in Euclidean space. In the recent work, it has been demonstrated that hyperbolic space is more suitable for learning representations for data which exhibit some kind of hierarchical structure such as nouns, social, semantic and complex networks [3, 15, 1, 12].

However, embedding events into non-Euclidean space has not yet been explored very well. Some previous work explored embedding *words* into other spaces to represent specificity of concepts [15]. Dhingra et al. [6] extend Nickel and Kiela's 2017 [15] work to learn sentence encoder that can embed sentences into hyperbolic space by using an unsupervised Skip Thought-based objective [20]. However, the extrinsic evaluation, e.g. sentiment classification, results do not show significant improvement over Euclidean space-based encoder. In addition, they do not analyze the learned embeddings deeply. They try to learn hierarchical structure exhibited in the data implicitly. However, before going into fully unsupervised approaches, we believe that we should explore the properties of learned event embeddings with explicit supervision of hierarchical structure. Specifically we explore the following research questions:

1. Can hyperbolic embedding, proposed for non-structured entity (i.e. word), be adopted for structured entity (i.e. event)?
2. Does hyperbolic event embedding capture specificity of events?

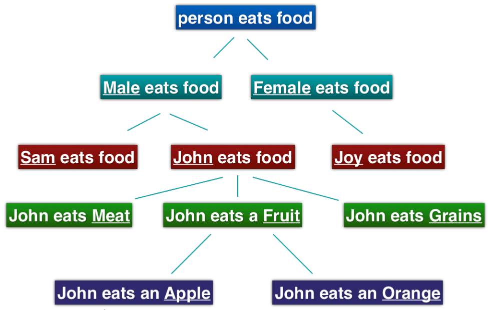

Figure 1: Event hierarchy in-terms of specificity

We explore [15]-like approach, where the model is explicitly informed what concepts form a hierarchy. The contribution of our work can be summarized as follows:

- This is the first study to explore event embeddings learned with explicit supervision of event hierarchy.
- Our experiments demonstrate that hyperbolic event embeddings learned with explicit supervision capture the hierarchical nature of events.
- We show that, even without explicit supervision of word-based conceptual hierarchy, the learned embedding captures the hierarchy of words.
- We also show that learned hyperbolic event embeddings generalize well to unseen events.

2 Preliminaries

2.1 Word embeddings in hyperbolic space

Next, we describe word embeddings in hyperbolic space as presented by [15]. Hyperbolic geometry is the geometry you obtain by assuming all the postulates of Euclid, except the fifth one, which is replaced by its negation. That is, in hyperbolic geometry there exist a line l and a point P not on l such that at least two distinct lines parallel to l pass through P.

In a regular tree, the number of children at each node grows exponentially with the distance from the node. In hyperbolic space, the circumference and the area of the circle is proportional to the $sinh$ of the radius and the $cosh$ of the radius, respectively. This exponential relationship with the radius makes hyperbolic space more suitable for embedding trees. A regular tree can be easily embedded in 2-dimensional hyperbolic space. However, previous work [15, 21, 6], has shown that as the dimension of the hyperbolic embedding increases, the performance on downstream tasks improves.

Of the many hyperbolic space models, Poincarés ball model is the more suitable model for representational learning in neural networks as it is more suitable for gradient-based optimization. Poincarés ball model can be defined as unit ball, $\mathcal{B}^d = \{\boldsymbol{x} \in R^d | \|\boldsymbol{x}\| < 1\}$, were the distance between two points u and v is given by;

$$d(\boldsymbol{u}, \boldsymbol{v}) = \operatorname{arcosh}\left(1 + 2\frac{\|\boldsymbol{u} - \boldsymbol{v}\|^2}{(1 - \|\boldsymbol{u}\|^2)(1 - \|\boldsymbol{v}\|^2)}\right) \quad (1)$$

To learn Poincaré embeddings, $\Theta = \{\boldsymbol{\theta}_i\}_{i=1}^n$, for a set of symbols $\mathcal{E} = \{e_i\}_{i=1}^n$, we need to define an objective function, $\mathcal{L}(\Theta)$, that minimizes the hyperbolic distance between semantically similar objects. Then, we need to optimize:

$$\Theta' \leftarrow \arg\min_{\Theta} \mathcal{L}(\Theta) \quad \text{s.t. } \forall \boldsymbol{\theta}_i \in \Theta : \|\boldsymbol{\theta}_i\| < 1 \quad (2)$$

[15] optimizes this equation using Riemannian Stochastic Gradient Descent (RSGD) [2]. In this work, we use the re-parameterization technique proposed by [6], described in section 2.2. In RSGD-based optimization, used by [15], it is possible that some embeddings can lie outside the Poincarés ball. Therefore, it is necessary to project such embeddings back in the Poincarés ball during each update. However, with re-parameterization technique [6], the projection is not necessary as the resulting embedding vectors always lie in the Poincarés ball. As a result of this, we can make use of any available optimizer such as Adam [11]. In addition, it was shown that training using re-parameterization converges faster while offering comparable results, in a similar task setting, to the work by [15].

2.2 Parametric Poincaré Embedding

Given event e_i and its embedding $\mathbf{e}(s)$, we compute:

$$\overline{\mathbf{v}} = \phi_{\text{dir}}(\mathbf{e}(s)), \quad \mathbf{v} = \frac{\overline{\mathbf{v}}}{\|\overline{\mathbf{v}}\|},$$
$$\overline{p} = \phi_{norm}(\mathbf{e}(s)), \quad p = \sigma(\overline{p}),$$

where ϕ_{dir} and ϕ_{norm} are arbitrary parametric functions, whose parameters are learned during training. We then obtain hyperbolic embedding $\theta = p\mathbf{v}$.

3 Hyperbolic Event Embeddings

3.1 Model

We use the re-parametrization technique described in section 2.2. Our approach is event encoder-agnostic. For $\mathbf{e}(s)$, we can employ any kind of sentence encoder that outputs fixed-length vector. We use LSTM as an event encoder. For projections, we use the following parametric function: $\phi_{\text{dir}}(\mathbf{x}) = W_1^T\mathbf{x}, \quad \phi_{norm}(\mathbf{x}) = W_2^T\mathbf{x}$ We expect that event embeddings will be organized in hierarchical manner such that more general events will appear closer to the origin and more specific events will appear towards the edge.

3.2 Training Objective

To learn representations $\Theta = \{\boldsymbol{\theta}_i\}_{i=1}^n$ for a set of events $\mathcal{E} = \{e_i\}_{i=1}^n$, we define a loss function $L(\Theta, d)$ that minimizes the hyperbolic distance (1) between embeddings of related events.

$$\mathcal{L}(\Theta, d) = \sum_{(u,v)\in\mathcal{D}} \log \frac{e^{-d(\boldsymbol{u},\boldsymbol{v})}}{\sum_{\boldsymbol{v}'\in\mathcal{N}(u)} e^{-d(\boldsymbol{u},\boldsymbol{v}')}}$$

where **u** and **v** are composition vectors from the sentence encoder. It is worth noting that because the hyperbolic distance is symmetrical, the loss function does not use any directions of edges between **u** and **v**.

4 Experiments & Results

4.1 Training Data

In our experiment, we use part of the entailment datasets, (KS2016), introduced by Kartsaklis and Sadrzadeh [10] which consist of 70 subject-verb-object (SVO) pairs, (u, v) were v is a more general event of u. From each event u in the dataset, we use WordNet hyponyms to get more specific events. In addition, we use WordNet hypernyms to get more general events.

Using the aforementioned method we get 12,803 positive SVO pairs, (u, v), and a vocabulary of 6027 unique words. This dataset, $\mathcal{D}$, is arranged in the form $\mathcal{D} = \{(u, v) | v \text{ is the general event of } u\}$. For each positive SVO pair (u, v), we generate negative example by pairing u with randomly sampled SVO triplet. We split the dataset into train/test in the ratio 4:1.

4.2 Model Settings

Our model consists of three layers. The first layer of our event encoder is an Embedding layer. This layer uses pretrained 100-dimensional glove vectors [16]. The second layer is an LSTM layer with 64 units, and the third and final layer is the Projection Layer. The projection layers projects to 128-dimensional Poincaré embedding space. We use the Adam optimizer [11] for optimization.

Norm	Event
0.262 - 0.271	physical object accent abstract entity
0.271 - 0.279	organism see abstract entity, somebody show abstract entity, living thing determine abstract entity, abstract entity show abstract entity, abstract entity verbalise noesis, physical object transfer abstract entity
0.279 - 0.288	woman experience abstract entity, idea fit abstract entity, person accept abstract entity, group verbalize abstract entity
0.288 - 0.296	human action give abstract entity, event show abstract entity, physical object show cognition
0.296 - 0.305	physical object change abstract entity, artifact represent abstract entity, psychological feature cerebrate abstract entity, physical entity interact substance, psychological feature move abstract entity, animate thing obtain abstract entity

Table 1: Events with the lowest norms

Norm	Event
0.99621 - 0.99623	police catch ripper, term tell law of reciprocal proportions, mutawa'een skirmish statutory offence
0.99623 - 0.99626	term tell first law of motion, police catch jailbird, police collect suffering, term tell third law of motion, acute glossitis indicate cause
0.99626 - 0.99628	flush indicate cause, police catch collaborator, dowager wholesale legal jointure, police torpedo crime
0.99628 - 0.9963	police lump evidence, fever indicate cause, police catch rustler
0.9963 - 0.99632	hypoglycemia indicate cause, term slip in concept, police catch stickup man

Table 2: Events with the largest norms

4.3 Intrinsic Qualitative Evaluation

We compare the norms of events from the resulting embedding. More general events are supposed to have lower norms than more specific events because general events are embedded closer to the origin. The results of this experiment are show in table 1 and table 2. Additionally, figure 2 shows the visualization of the 2D Poincaré event Embedding. For visual clarity, we manually picked only a few events from the KS2016 dataset. The resulting embedding capture the hierarchical nature of events, it places more general events closer to the origin.

In addition, we also compare the word level norms. Words which express more general concepts are supposed to lie closer to the origin than words the express specific concepts. The results for this experiment are shown in table 3 and table 4. Figure 3, shows the visualization of the word embedding obtained from 2D Poincaré event Embedding. For visual clarity, we manually picked a few related words from the training set. Even without explicit supervision of word-based conceptual hierarchy, the learned embedding captures the hierarchy of words.

4.4 Intrinsic Quantitative Evaluation

In the dataset, $\mathcal{D}$, for each positive pair, $P(u_i, v_i)$ and its negative counterpart $N(u_i, v_i')$, we calculate *is-a* score [15].

$$\text{score}(\text{is-a}(x, y)) = -(1 + \alpha(\|y\| - \|x\|))d(x, y) \quad (3)$$

Norm	Words
0.473 - 0.533	fauna, vertebrate, abstract, art, entity, interact
0.533 - 0.553	cognition, department, europol, organism, host, aspect, landscape, intelligence, biological, nestle, reality, defense, complex, interior, index, germany
0.553 - 0.573	culture, raw, chemical, flora, abstraction, fleischer, agency, language, nature, food, speak, flavor, rubin, situation, affairs, kraft, agriculture

Table 3: Words with the lowest norms

Norm	Words
0.94 - 0.945	femoral, election, archimedes, teargas, voter, elected, quadruple, bus, cause
0.945 - 0.95	suffrage, systolic, venous, puncture, hail, levitation, string, lumbar, encephalitis, highway, cantus
0.95 - 0.955	incumbent, rail, edema, livery, spur, atrial, hemorrhagic, fibrillation, booster, siphon, ferry, railway
0.955 - 0.965	pulmonary, maiden, republish, arterial, torpedo

Table 4: Words with the largest norms

We count the number of instances for which the *is-a(P)* score is greater than the *is-a(N)* score on the held out test set. In this evaluation, we obtained an accuracy of 0.845763. Therefore, the resulting embedding generalizes well to unseen events. The obtained *is-a* score is well above random guessing.

5 Related Work

Research on event understanding ranges from inferring intent and emotional reaction [18], sentiment classification [7], and script knowledge [19] modeling [4, 9, 14, 5, 17, 23, 13]. Previous work in event understanding [18, 23, 17, 9, 14, 5] use distributed representational of events in Euclidean space. However, in recent work, it has been demonstrated that hyperbolic space is more suitable for learning representations for data which exhibit some kind of hierarchical structure.

A variety of approaches have been proposed to capture the hierarchical structure of datasets. Vilnis et al. [22] proposed Gaussian Embeddings to capture uncertainty and asymmetry. Nickel and Kiela [15] learned word embeddings on Poincarés ball, while our work focuses on event embeddings. Tay et al. [21] learned question and answer embeddings on the Poincarés ball for question-answer retrieval. Dhingra et al. [6] extended [15] work and [6] showed a method to embed words and sentences into hyperbolic space. However, the extrinsic evaluation results do not show significant improvement over Euclidean space-based encoder, and they do not analyze the learned embeddings deeply. They learned hierarchical structure exhibited in the data implicitly using an unsupervised Skip Thought-based objective [20]. However, before going into fully unsupervised approaches, we believe that we should explore the properties of learned event embeddings with explicit supervision of hierarchical structure.

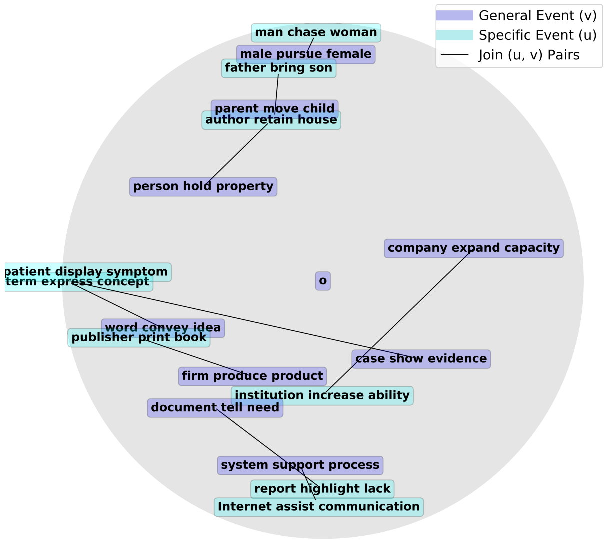

Figure 2: 2D Poincaré event embedding

Figure 3: Wording embedding from 2D Poincaré event encoder

6 Conclusion and future work

In this paper, we presented the first study that explores learning event embeddings with explicit supervision of event hierarchy. We demonstrated that hyperbolic event embeddings learned with explicit supervision capture the hierarchical nature of events. We also showed that, even without explicit supervision of word-based conceptual hierarchy, the learned embedding also captures the hierarchy of words. Finally, we showed that the learned hyperbolic event embeddings generalize well to unseen events.

In future we intend to perform extrinsic evaluation of the hyperbolic event embedding and to attempt script knowledge modeling. In addition, we intend to explore if conventional point-based embedding capture general-specific relations of events. We also intend to compare our method with Skip-Thought [6] objective. Finally, we intend to consider monotonicity for real-world situation like Multi-NLI.

Acknowledgement

This work was partially supported by JSPS KAKENHI Grant Number 15H01702 and JST CREST Grant Number JPMJCR1513, including AIP challenge.

References

[1] Aaron B. Adcock, Blair D. Sullivan, and Michael W. Mahoney. Tree-like structure in large social and information networks. *Proceedings - IEEE International Conference on Data Mining, ICDM*, pages 1–10, 2013.
[2] Silvere Bonnabel. Stochastic gradient descent on riemannian manifolds. *IEEE Transactions on Automatic Control*, 2013.
[3] Benjamin Paul Chamberlain, James Clough, and Marc Peter Deisenroth. Neural Embeddings of Graphs in Hyperbolic Space. Technical report, 2017.
[4] Nathanael Chambers and Dan Jurafsky. Unsupervised learning of narrative event chains. *Proceedings of the Association of Computational Linguistics*, 31(14):789–797, 2008.
[5] Kevin Clark and Christopher D. Manning. Improving Coreference Resolution by Learning Entity-Level Distributed Representations. 2016.
[6] Bhuwan Dhingra, Christopher Shallue, Mohammad Norouzi, Andrew Dai, and George Dahl. Embedding Text in Hyperbolic Spaces. In *Proceedings of the Twelfth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-12)*, pages 59–69, New Orleans, Louisiana, USA, 2018. Association for Computational Linguistics.
[7] Haibo Ding and Ellen Riloff. Weakly Supervised Induction of Affective Events by Optimizing Semantic Consistency. In *AAAI*, 2018.
[8] Martin B.H. Everaert, Marinus A.C. Huybregts, Noam Chomsky, Robert C. Berwick, and Johan J. Bolhuis. Structures, Not Strings: Linguistics as Part of the Cognitive Sciences. *Trends in Cognitive Sciences*, pages 729–743, 2015.
[9] Mark Granroth-Wilding and Stephen Clark. What Happens Next ? Event Prediction Using a Compositional Neural Network Model. In *In Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2727–2733, Gothen burg, Sweden, 2016.
[10] Dimitri Kartsaklis and Mehrnoosh Sadrzadeh. A Compositional Distributional Inclusion Hypothesis. In *Proceedings of the 9th International Conference on Logical Aspects of Computational Linguistics. Volume 10054*, LACL 2016, pages 116–133, Berlin, Heidelberg, 2016. Springer-Verlag.
[11] Diederik P Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. In *Proceed- ings of the 3rd International Conference on Learning Representations (ICLR)*, 2014.
[12] Dmitri Krioukov, Fragkiskos Papadopoulos, Maksim Kitsak, Amin Vahdat, and Marian Boguna. Hyperbolic Geometry of Complex Networks. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 82(3), jun 2010.
[13] Zhongyang Li, Xiao Ding, and Ting Liu. Constructing Narrative Event Evolutionary Graph for Script Event Prediction. *EMNLP*, may 2018.
[14] Ashutosh Modi. Event Embeddings for Semantic Script Modeling. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 75–83, Berlin, Germany, 2016.
[15] Maximilian Nickel and Douwe Kiela. Poincar\'e Embeddings for Learning Hierarchical Representations. may 2017.
[16] 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.
[17] Karl Pichotta and Raymond J. Mooney. Using Sentence-Level LSTM Language Models for Script Inference. *Arxiv*, pages 279–289, 2016.
[18] Hannah Rashkin, Maarten Sap, Emily Allaway, Noah A. Smith, and Yejin Choi. Event2Mind: Commonsense Inference on Events, Intents, and Reactions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, volume 1 Long Pap, pages 463–473, Melbourne, Australia, oct 2018. Association for Computational Linguistics.
[19] Roger C Schank and Robert P Abelson. *Scripts, plans, goals, and understanding : an inquiry into human knowledge structures.* Hillsdale, N.J. : Lawrence Erlbaum Associates, 1977.
[20] Shuai Tang, Hailin Jin, Chen Fang, Zhaowen Wang, and Virginia R. de Sa. Trimming and Improving Skip-thought Vectors. jun 2017.
[21] Yi Tay, Luu Anh Tuan, and Siu Cheung Hui. Hyperbolic Representation Learning for Fast and Efficient Neural Question Answering. 2018.
[22] Luke Vilnis and Andrew McCallum. Word Representations via Gaussian Embedding. dec 2014.
[23] Noah Weber, Niranjan Balasubramanian, and Nathanael Chambers. Event Representations with Tensor-based Compositions. 2017.