Representing and Reconstructing PhySH: Which Embedding Competent?

Xiaoli Chen  
National Science Library  
University of Chinese Academy of Sciences  
Chinese Academy of Sciences  
chenxl@mail.las.ac.cn

Zhixiong Zhang  
National Science Library  
Wuhan Library  
University of Chinese Academy of Sciences  
Chinese Academy of Sciences  
Zhangzhx@mail.las.ac.cn

Abstract

In this paper, we conduct a comprehensive comparison of well-known embeddings’ capability in capturing the hierarchical Physics knowledge. Several key findings are: (i) Poincaré embeddings do outperform if trained on PhySH, but it fails if trained on co-occurrence pairs which are extracted from raw text. (ii) No algorithm can properly learn hierarchies from the more realistic case of co-occurrence pairs, which contains more noisy relations other than hierarchical relations. (iii) Our statistic analysis of Poincaré embedding’s representation of PhySH shows successful hierarchical representation share two characteristics: firstly, upper-level terms have a smaller semantic distance to root; secondly, upper-level hypernym-hyponym pairs should be further apart than lower-level hypernym-hyponym pairs.

1 Introduction

Concept hierarchy or taxonomy\(^1\) is highly organized and expertly curated hierarchical hypernym-hyponym sets. How to effectively represent these terms with the hierarchical relation is the main hurdle for automatically taxonomy construction and other downstream applications.

Though embeddings have been taken for granted in most NLP pipelines, none of the previous work has fully explored which embeddings can capture hierarchical scientific knowledge. Even though Poincaré embedding is proved to have a better ability to capture hierarchical relations, it is learned based on existing WordNet hypernym-hyponym pairs. It is never been tested in the scientific domain. In this paper, we conduct a comprehensive comparison of well-known embeddings’ performance in reconstructing Physical Subject Headings (PhySH) from raw APS datasets.

Our main contributions are mainly three-fold: Firstly, for the first time, we compare mainstream embeddings’ capability to represent and reconstruct Physical Subject Headings (PhySH) both from raw text and PhySH. Secondly, our experiment shows Poincaré embedding is not sufficient for taxonomy induction from raw text. Thirdly, we explore the characteristics of successful representation of PhySH, which might be the inspiration for better taxonomy construction algorithms.

2 Related Work

Representations for Concept Hierarchy. Representations for concept hierarchy has been receiving quite growing interests in recent years (Kozareva et al., 2008; Carlson et al., 2010; Shen et al., 2018). It is the basis of automatically taxonomy construction. In the survey study of (Wang et al., 2017), there are Pattern-based (Hearst, 1992; Wu et al., 2012; Kozareva and Hovy, 2010) methods and distributional (Navigli and Velardi, 2004; Luu et al., 2014; Pado and Lapata, 2003; Baroni and Lenci, 2010; Nguyen et al., 2017) methods use hand-crafted rule-based, co-occurrence features, syntactic features or graph features to learn representations of hierarchical pairs. They also apply pretrained neural language models such as Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014).

Recently, Poincaré embedding (Nickel and Kiela, 2017) is proposed to better represent hierarchical relations. Following works like (Law et al., 2019) use Lorentzian distance to replace the Poincaré metric, (Dhingra et al., 2018) extends Poincaré embedding to apply in raw text with re-parameterization technique, (Leimeister and Wilson, 2019) and (Tifrea et al., 2019) introduce hyperbolic embeddings in word embeddings like Skipgram and GloVe. Effectively in reconstructing WordNet though, the Poincaré embedding is not quite perfect yet (De Sa et al., 2018). It has only been tested on WordNet re-
construction and Hyperlex entailment (Nickel and Kiela, 2017). Whether it is an effective tool in representing hierarchical relations from raw domain text need to be further explored.

**Embeddings Analysis.** With the fast pacing of text representation technology, it is also important to revisit existing embedding methods for different downstream tasks. Several previous works have explicitly done this work based on their unique perspectives. Gladkova et al. (2016) explores GloVe’s ability to encode different morphological and semantic relations. (Zuccon et al., 2015) analyze word embeddings for information retrieval. Nooralahzadeh et al. (2019) compared COW and Skipgram by using Gensim implementation (Rehurek and Sojka, 2010) with several different hyper-parameters settings and different domain corpus. Sanchez and Riedel (2017) explored different datasets in evaluation hypernyms identification by using GloVe. (Lastra-Dámaz et al., 2019) surveys main word embeddings for word similarity.

Despite the above-mentioned work, there is still a missing part describing which embedding is the optimal choice for taxonomy induction. In this paper, we design our evaluation pipeline to choose the optimal embedding scheme for taxonomy learning and construction. In our paper, we consider two perspectives to represent and construct concept hierarchy: (i) Learn and construct from raw texts by word embeddings; (ii) Learn and construct from extracted co-occurrence pairs from raw texts by graph embeddings and Poincaré embeddings.

### 3 Method

In our pipeline (Figure 1), we follow three steps: raw text and PhySH preprocessing; learn various embeddings with different hyperparameters; evaluate embeddings by reconstructing PhySH.

We evaluate the following embeddings:

- **Word embeddings**: CBOW and Skipgram (Mikolov et al., 2013), fastText (Joulin et al., 2017), GloVe (Pennington et al., 2014)².
- **Graph embeddings**: deepWalk (Perozzi et al., 2014), node2vec (Grover and Leskovec, 2016), LINE (Tang et al., 2015), LLE (Roweis and Saul, 2000), HOPE (Ou et al., 2016), GF(Ahmed et al., 2013), SDNE(Wang et al., 2016)³.
- **Poincaré embeddings**: Poincaré-gensim⁴, Poincaré-cpp ⁵, Poincaré-pytorch⁶, Poincaré-numpy⁷, Poincaré-glove⁸ (Tifrea et al., 2019).

**Word embeddings** are trained on **title and abstract** of APS publications. The PhySH terms’ embedding vectors will be extracted for taxonomy construction.

---

²CBOW, Skipgram and fastText are trained by https://github.com/NHOPA/word2vec_pipeline. GloVe is trained by https://github.com/stanfordnlp/GloVe

³Graph embeddings are implemented by OpenNE repository https://github.com/thunlp/OpenNE

⁴https://radimrehurek.com/gensim/models/poincare.html

⁵https://github.com/TatsuyaShirakawa/poincare-embedding.git

⁶https://github.com/facebookresearch/poincare-embddings

⁷https://github.com/nishnik/poincare_embeddings.git

⁸https://github.com/alex-tifrea/poincare_glove
Graph embeddings are trained on the co-occurrence of PhySH terms in each of APS publications. As with (Nickel and Kiela, 2017), we also train graph embeddings and Poincaré embeddings on PhySH hypernym-hyponym pairs.

**Taxonomy Reconstruction:** We follow (Nickel and Kiela, 2017) to reconstruct taxonomy based on embedding vectors. For each embedding vector in Poincaré disk space, which is denoted as $B^d = \{ x \in R^d, \|x\|_2 \leq 1 \}$. The norm of each vector can measure the radius of each vector, while the hyperbolic distance can measure the closeness of two vectors. The closest two are assigned as hypernym-hyponym pairs. The hyperbolic distance of two vector points $u, v \in B^d$ is calculated as (Nickel and Kiela, 2017).

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

The distance could only tell how semantically close are the node pairs $(u, v)$. But which one is the parent node is not answered. One property that makes hyperbolic space outstanding for the hierarchical structure is that the hyperbolic disc area and circle length grow exponentially with its radius. Node with the smaller norm is the higher-level term.

### Table 2: PhySH reconstruction from PhySH hypernym-hyponym pairs. Since there is no context information, word embeddings are not applicable here.

| Model Name | Metric       | Dimensions |
|------------|--------------|------------|
|            | low          | 7          | 10         | 20         | 50         | 100        | 200        |
| deepWalk   | mean rank    | 517.22     | 496.36     | 496.67     | 497.64     | 525.74     | 539.92     |
|            | MAP          | 0.22       | 0.19       | 0.21       | 0.22       | 0.22       | 0.23       |
| GF         | mean rank    | 277.24     | 125.80     | 50.67      | 8.90       | 2.93       | 9.79       |
|            | MAP          | 0.40       | 0.35       | 0.58       | 0.65       | 0.66       | 0.66       |
| GloRep     | mean rank    | 78.87      | 34.45      | 22.19      | 13.45      | 8.21       | 8.18       |
|            | MAP          | 0.49       | 0.53       | 0.56       | 0.58       | 0.57       |
| HOPE       | mean rank    | 561.03     | 758.32     | 690.95     | 615.45     | 513.47     |
|            | MAP          | 0.64       | 0.47       | 0.43       | 0.43       | 0.45       |
| LINE       | mean rank    | 439.49     | 344.14     | 141.35     | 34.32      | 15.84      |
|            | MAP          | 0.57       | 0.73       | 0.52       | 0.60       | 0.62       |
| node2vec   | mean rank    | 265.60     | 264.94     | 265.60     | 264.81     | 269.52     | 265.28     |
|            | MAP          | 0.35       | 0.34       | 0.35       | 0.34       | 0.35       |
| SDINE      | mean rank    | 72.58      | 33.18      | 17.87      | 5.37       | 1.96       |
|            | MAP          | 0.37       | 0.54       | 0.12       | 0.04       | 0.02       | 0.02       |

### 4.2 Evaluation Metrics

**mean rank** and **MAP** metrics are used to measure taxonomy reconstruction performance. **mean rank** is calculated for each node’s distance of ground truth children against all other nodes. **MAP** is the mean average precision at the threshold of each correctly retrieved child.

$$mean\_rank(u) = \frac{sp(u)}{sp(u) + lp(u)} \in [0, 1]$$

$l_p(u)$ is the furthest length from node $u$ to its descendants. $sp(u)$ is the shortest length from node $u$ to root node. The optimal embedding should score a low $mean\_rank$ and a high $MAP$.

### 4.3 PhySH Reconstruction Evaluation

**PhySH Reconstruction From Raw Text.** In this experiment, we extract co-occurrence of PhySH terms in each APS publication. The graph embeddings are trained on the co-occurrence graph. Poincaré embeddings are trained on the noisy co-occurrence pairs. Word embeddings are trained on APS publication raw texts. PhySH terms’ representation vectors are extracted from word vectors in the postprocessing step.

### Table 1 is the performance of PhySH reconstruction by learning representation from raw APS datasets. None of the embeddings get the best result in both metrics. Word embeddings like CBOW achieve better $MAP$, while graph embeddings like deepWalk outperform in $mean\_rank$. Poincaré embeddings did not show any superior. Learn PhySH

9[https://journals.aps.org/datasets](https://journals.aps.org/datasets)

10[https://github.com/physh-org/PhySH](https://github.com/physh-org/PhySH)

11Web of Science is a commercial database of Clarivate Analytics, it can be accessed by most universities and institutions

12We experiment each embedding with different hyperparameters by grid search, we present the optimal performance of each embedding in the tables.
from noisy co-occurrence pairs are much more complicated than the mammal tree of the WordNet described in the origin paper (Nickel and Kiela, 2017). We can conclude Poincaré embeddings are not sufficient for learning and representing from the co-occurrence pairs.

**PhySH Reconstruction From PhySH.** Table 2 is the performance of PhySH reconstruction by learning representation from PhySH hypernym-hyponym pairs. The graph embeddings are trained on the PhySH hypernym-hyponym graph. Poincaré embeddings are trained on the PhySH hypernym-hyponym pairs.

In this experiment, Poincaré’s official implementation Poincaré-Pytorch wins with far better results than other algorithms. This is because Poincaré is trained with the loss function designed to learn hierarchies, while graph embeddings are trained to learn from neighbors and global graph structure. However, GF at dimension 100 and LINE at dimension 200 also get very good performance.

### 4.4 The Hierarchical Characteristics of PhySH Poincaré embedding

If we understand the successful representation characteristics of taxonomy hierarchical relations, it will be the help of taxonomy construction. We will analyze what are the hierarchical characteristics of PhySH preserved by Poincaré embeddings in this section.

In Figure 2, we visualize how the norm value varies in different PhySH level. There is a clear pattern from taxonomy level 2 to level 6: lower-level terms have bigger norm values. It means lower terms are further from the root term. The pace of the decrease of the norm in lower levels seems to decelerate, which needs to be further validated. However, the norm of level 1 terms is rather distributed, which we think is the points where Poincaré embedding fails.

In Figure 3, we compare the distance of terms over different PhySH levels. The ancestor nodes are further than parent nodes. For each node, its distance to the child is smaller than the distance to parent, and the distance to the child is nearly half as the distance to parent. These patterns are important for a successful representation of taxonomy.

### 5 Conclusion and Future Work

we compare word embeddings, graph embeddings, and Poincaré embeddings by reconstructing PhySH. We consider two scenario case: reconstructing from raw texts and reconstructing from existing PhySH. The experiment shows even though Poincaré embeddings far outweigh other embeddings in reconstructing PhySH from PhySH, it is also not competent as other embeddings in reconstructing PhySH from raw APS texts.

We further demystify what is the success of Poincaré embeddings in reconstructing PhySH from PhySH. The future work would be how to design a powerful taxonomy induction algorithm which could benefit from the characteristics of our paper.

### Acknowledgments

The authors thank anonymous reviewers for their helpful comments that improved the manuscript. The authors also thank American Physical Society for providing access to the APS Datasets.

This work is supported in part by the National Natural Science Foundation of China under contracts No 71950003, and the project “Design and Research on a Next Generation of Open Knowledge Services System and Key Technologies” (2019XM55).
6 References

References

Amr Ahmed, Nino Shervashidze, Shravan Narayanamurthy, Vanja Josifovski, and Alexander J. Smola. 2013. Distributed large-scale natural graph factorization. In Proceedings of the 22Nd International Conference on World Wide Web, WWW ’13, pages 37–48, New York, NY, USA. ACM.

Marco Baroni and Alessandro Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. Computational Linguistics, 36(4):673–721.

Andrew Carlson, Justin Betteridge, Richard C. Wang, Estevam R. Hruschka, Jr., and Tom M. Mitchell. 2010. Coupled semi-supervised learning for information extraction. In Proceedings of the Third ACM International Conference on Web Search and Data Mining, WSDM ’10, pages 101–110, New York, NY, USA. ACM.

Christopher De Sa, Albert Gu, Christopher RÃ­al’, and Frederic Sala. 2018. Representation tradeoffs for hyperbolic embeddings. Proceedings of machine learning research, 80.

Bhuwan Dhingra, Christopher Shallue, Mohammad Norouzi, Andrew Dai, and George Dahl. 2018. 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. Association for Computational Linguistics.

Anna Gladkova, Aleksandr Drozd, and Satoshi Matsuoka. 2016. Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn’t. In Proceedings of the NAACL Student Research Workshop, pages 8–15, San Diego, California. Association for Computational Linguistics.

Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In COLING 1992 Volume 2: The 15th International Conference on Computational Linguistics.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pages 427–431. Association for Computational Linguistics.

Zornitsa Kozareva and Eduard Hovy. 2010. A semi-supervised method to learn and construct taxonomies using the web. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 1110–1118, Cambridge, MA. Association for Computational Linguistics.

Zornitsa Kozareva, Ellen Riloff, and Eduard Hovy. 2008. Semantic class learning from the web with hyponym pattern linkage graphs. In Proceedings of ACL-08: HLT, pages 1048–1056, Columbus, Ohio. Association for Computational Linguistics.

Juan J. Lastra-DÁ­naez, Josu Goikoetxea, Mohamed Ali Hadij Taleb, Ana GarcÃ­a-Serrano, Mohamed Ben Aouicha, and Eneko Agirre. 2019. A reproducible survey on word embeddings and ontology-based methods for word similarity: Linear combinations outperform the state of the art. Engineering Applications of Artificial Intelligence, 85:645 – 665.

Marc T Law, Jake Snell, and Richard S Zemel. 2019. Lorentzian distance learning.

Matthias Leimeister and Benjamin J. Wilson. 2019. Skip-gram word embeddings in hyperbolic space.

Anh Tuan Luu, Jung-jae Kim, and See Kiong Ng. 2014. Taxonomy construction using syntactic contextual evidence. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 810–819, Doha, Qatar. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2, NIPS’13, pages 3111–3119, USA. Curran Associates Inc.

Roberto Navigli and Paola Velardi. 2004. Learning domain ontologies from document warehouses and dedicated web sites. Comput. Linguist., 30(2):151–179.

Kim Anh Nguyen, Maximilian KĂ­uõeper, Sabine Schulte im Walde, and Ngoc Thang Vu. 2017. Hierarchical Embeddings for Hyponymy Detection and Directionality. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pages 233–243, Copenhagen, Denmark.

Maximilian Nickel and Douwe Kiela. 2017. Poincaré embeddings for learning hierarchical representations. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, Advances in Neural Information Processing Systems 30, pages 6341–6350, Curran Associates, Inc.

Farhad Nooralahzadeh, Lilja Óvrelid, and Jan Tore Lønning. 2019. Evaluation of domain-specific word embeddings using knowledge resources. LREC 2018 - 11th International Conference on Language Resources and Evaluation, pages 1438–1445.
