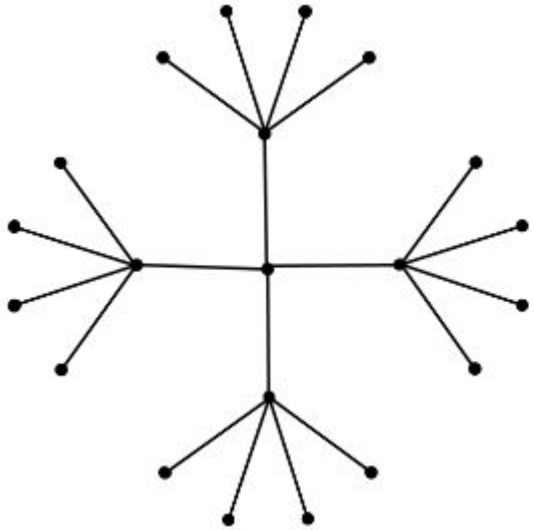
On hierarchical representations using Poincaré Embeddings and its applications for noun compositionality

Puzyrev Dmitri
HSE
16.11.2019

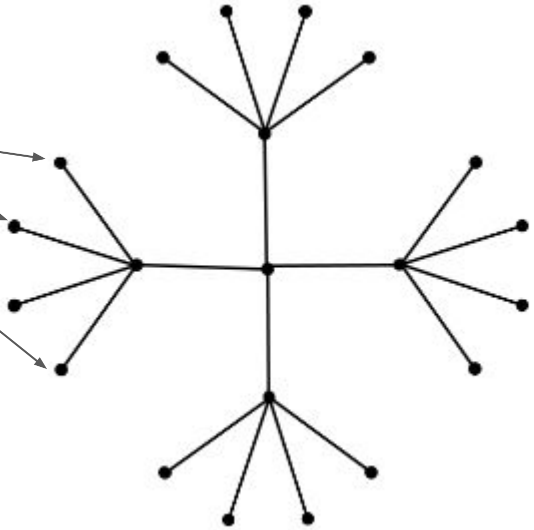
Problem of hierarchical representations



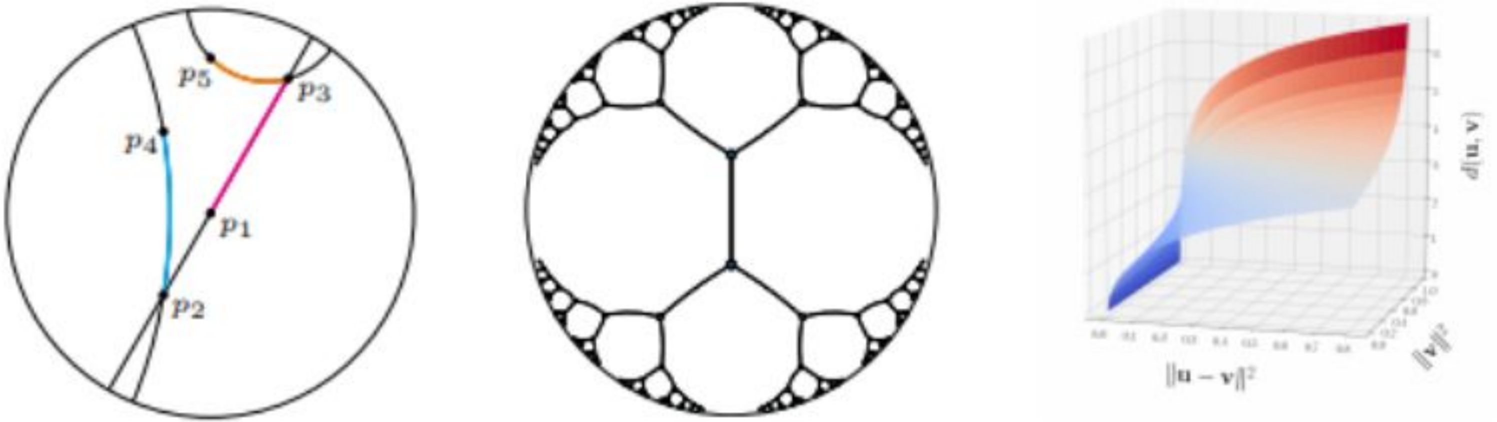
suppose we want to project tree on euclidean space

Problem of hierarchical representations

the deeper we go, the closer siblings get!



Solution: projection onto Poincaré unit space



(a) Geodesics of the Poincaré disk (b) Embedding of a tree in $\mathcal{B}^2$ (c) Growth of Poincaré distance

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

How to construct this space?

Define d-dimensional unit sphere

$$\mathcal{B}^d = \{\boldsymbol{x} \in \mathbb{R}^d \mid \|\boldsymbol{x}\| < 1\}$$

Our model corresponds to Riemannian manifold with its metric defined as

$$g_{\boldsymbol{x}} = \left(\frac{2}{1 - \|\boldsymbol{x}\|^2}\right)^2 g^E,$$

How to construct this space? (cont.)

$$d(x, y) = \operatorname{arccos}(1 + 2 * \frac{||x - y||^2}{(1 - ||x||^2)(1 - ||y||^2)})$$

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

How to get embeddings in this space?

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

standard optimization problem with constraint of keeping embeddings inside the ball

How to get embeddings in this space? (optimize)

RSGD (Riemannian Stochastic Gradient Descent)

$$\boldsymbol{\theta}_{t+1} = \mathfrak{R}_{\theta_t}\left(-\eta_t \nabla_R \mathcal{L}(\boldsymbol{\theta}_t)\right)$$

Gradient in Euclidean case

$$\nabla_E = \frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial d(\boldsymbol{\theta}, \boldsymbol{x})} \frac{\partial d(\boldsymbol{\theta}, \boldsymbol{x})}{\partial \boldsymbol{\theta}}$$

Then Gradient in Riemannian case

$$\frac{(1 - \|\boldsymbol{\theta}_t\|^2)^2}{4} \nabla_E$$

How to get embeddings in this space? (optimize)

$$\operatorname{proj}(\boldsymbol{\theta}) = \begin{cases} \boldsymbol{\theta}/\|\boldsymbol{\theta}\| - \varepsilon & \text{if } \|\boldsymbol{\theta}\| \geq 1 \\ \boldsymbol{\theta} & \text{otherwise ,} \end{cases}$$

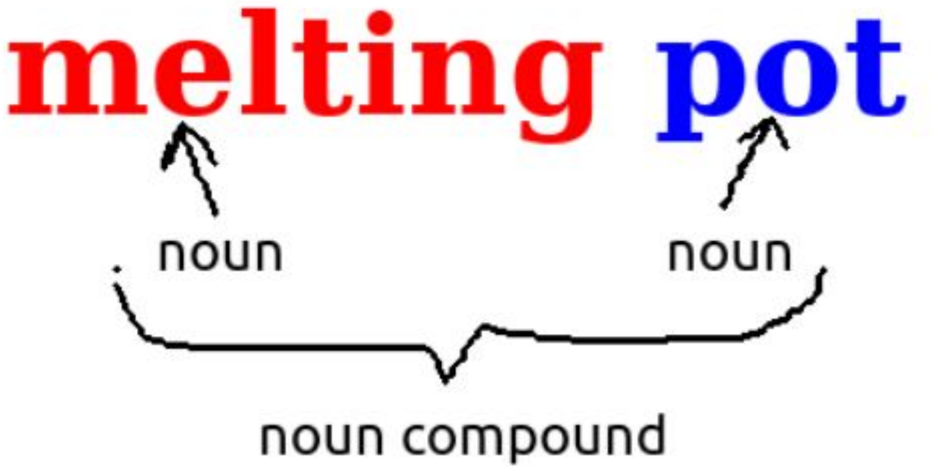
$$\boldsymbol{\theta}_{t+1} \leftarrow \operatorname{proj}\left(\boldsymbol{\theta}_t - \eta_t \frac{(1 - \|\boldsymbol{\theta}_t\|^2)^2}{4} \nabla_E\right)$$

Some applications

Compositionality of noun compounds

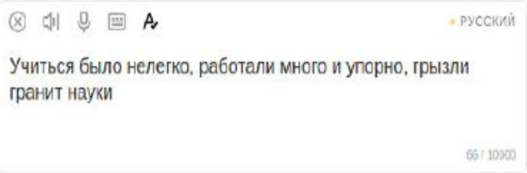
melting pot

Compositionality of noun compounds



Where it can be important?

machine translation

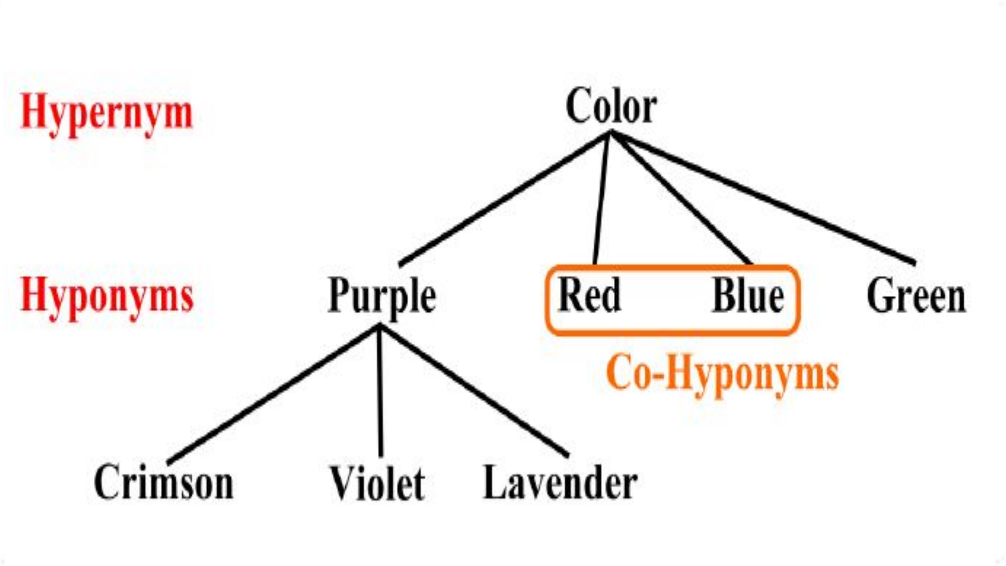


Our approach

- Two setups: “unsupervised” and “supervised”
- Each setup uses two types of embeddings: usual CBOW from FastText and Poincare
- In Poincare part, *hyponym and hyperonym* information is stored

On the Compositionality Prediction of Noun Phrases using Poincare Embeddings
Abhik Jana, Dmitry Puzyrev, Alexander Panchenko, Pawan Goyal, Chris Biemann, and Animesh Mukherjee
In proceedings of ACL, 2019

Hyperonymy and hyponymy



Fruit

hypernym_of

hyponym_of

Banana

“Unsupervised” approach

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

Scores are based on euclidean distance between sum of word vectors and word vector of a compound

Results for “unsupervised” approach

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

Baselines with different embedding models

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

Our approach for different amount of hyperonymy pairs learned

“Supervised” approach

$$Score_S(w_1 w_2) = (1 - \alpha) * Score_{DS}(w_1 w_2) + \alpha * Score_{HS}(w_1 w_2)$$

Results for “supervised” approach

FD-R				
	Kernel Regression		PLS Regression	
	Mean ($\lvert\rho\rvert$)	SD ($\lvert\rho\rvert$)	Mean ($\lvert\rho\rvert$)	SD ($\lvert\rho\rvert$)
α	MODEL-DP-S, CBOW vectors of dim. 50			
0.2	0.45	0.05	0.44	0.05
0.3	**0.45**	0.05	0.44	0.05
0.4	**0.45**	0.05	0.44	0.05
0.5	**0.45**	0.05	0.43	0.05
0.6	0.44	0.05	0.42	0.05

FD-F				
	Kernel Regression		PLS Regression	
	Mean ($\lvert\rho\rvert$)	SD ($\lvert\rho\rvert$)	Mean ($\lvert\rho\rvert$)	SD ($\lvert\rho\rvert$)
α	MODEL-DP-S, CBOW vectors of dim. 50			
0.2	0.43	0.05	0.43	0.05
0.3	**0.43**	0.05	0.43	0.05
0.4	**0.43**	0.05	0.42	0.05
0.5	**0.43**	0.05	0.41	0.05
0.6	0.42	0.05	0.39	0.05

“Reduced” dataset (only examples with existing hypernymy are present)

Full dataset (zero vector is passed where Poincare representation is not available)

Another interesting application: music recommendations

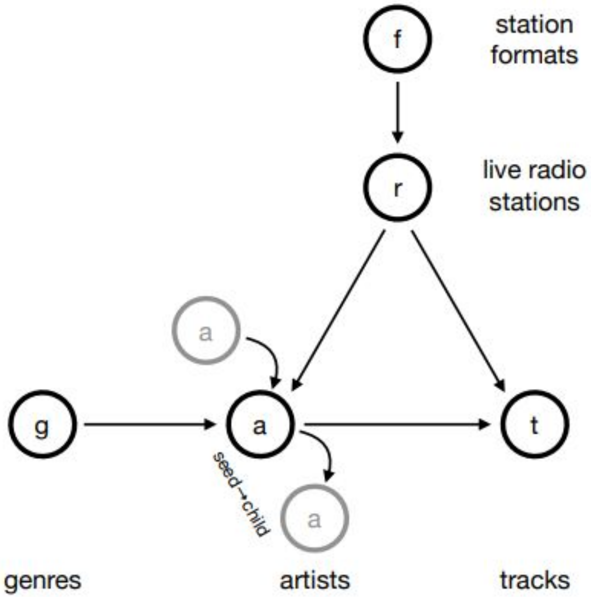


Table 1: Entity counts

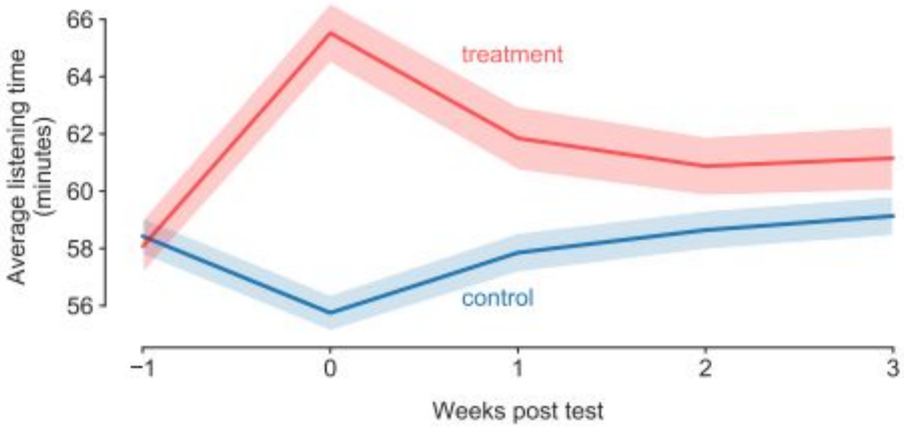
Entity type	Order of magnitude
Station formats and genres	10-100
Live radio stations	1,000
Artists	10,000
Tracks	1,000,000
Users	1,000,000

Music Recommendations in Hyperbolic Space: An Application of Empirical Bayes and Hierarchical Poincaré Embeddings
Timothy Schmeier, Sam Garrett, Joseph Chisari, and Brett Vintch
RecSys ’19, September 16–20, 2019, Copenhagen, Denmark

Another interesting application: music recommendations

- Child-parent relations between artists
- Emperical Bayes approach to determine link importance
- Poincare embeddings for representing hierarchy

Results


treatment
control
Average listening time (minutes)
66
64
62
60
58
56
−1
0
1
2
3
Weeks post test

Thanks for listening!