SphereRE: Distinguishing Lexical Relations with Hyperspherical Relation Embeddings

Chengyu Wang[1], Xiaofeng He[1*], Aoying Zhou[2]

[1] School of Computer Science and Software Engineering,
[2] School of Data Science and Engineering,
East China Normal University
Shanghai, China

Outline

- Introduction
- The SphereRE Model
  - Learning Objective
  - Relation-aware Semantic Projection
  - Relation Representation Learning
  - Lexical Relation Classification
- Experiments
- Conclusion

DaSE Data Science & Engineering

Introduction (1)

- Lexical Relation Classification
  - Task: Classifying a word pair into a finite set of relation types (e.g., synonymy, antonymy)

Relation	Tag	Template	Example
Synonymy	SYN	W2 can be used with the same meaning as W1	*candy-sweet, apartment-flat*
Antonymy	ANT	W2 can be used as the opposite of W1	*clean-dirty, add-take*
Hypernymy	HYPER	W1 is a kind of W2	*cannabis-plant, actress-human*
Part-whole meronymy	PART_OF	W1 is a part of W2	*calf-leg, aisle-store*
Random word	RANDOM	None of the above relations apply	*accident-fish, actor-mild*

Examples taken from the CogALex-V shared task

DaSE Data Science & Engineering

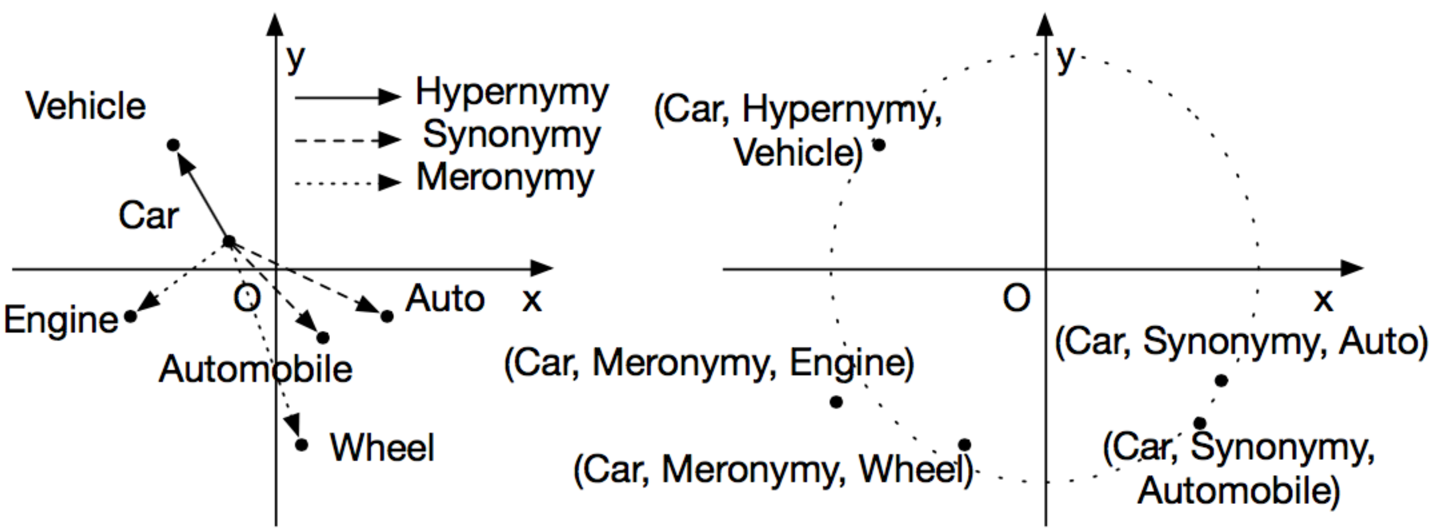
Introduction (2)

- Existing Approaches
  - **Path-based approaches:** use dependency paths connecting two terms to infer lexical relations
    - **"Low coverage" problem**
  - **Distributional approaches:** consider the global contexts of terms to predict lexical relations using word embeddings
    - **"Lexical memorization" problem**

DaSE Data Science & Engineering

Introduction (3)

- Our Idea
  - Learning relation embeddings for term pairs (in the hyperspherical embedding space)
  - Term pairs with similar lexical relation types share similar embeddings



(a) Term Embedding Space (b) Relation Embedding Space

SphereRE: Learning Objective (1)

- Basic Notations
  - Training data (term pairs): $(x_i, y_i) \in D$
  - Testing data (term pairs): $(x_i, y_i) \in U$
  - Lexical relation types (e.g., synonymy, antonymy): $r_i \in R$
- Learning Objective in the Word Embedding Space
  - $f_m(\vec{x}_i)$: maps the relation subject x_i to the relation object y_i in the embedding space, where x_i and y_i have the lexical relation type $r_m \in R$
  - Objective function:

$$J_f = \sum_{i=1}^{|D|} \sum_{r_m \in R} I(r_i = r_m) \| f_m(\vec{x}_i) - \vec{y}_i \|^2$$

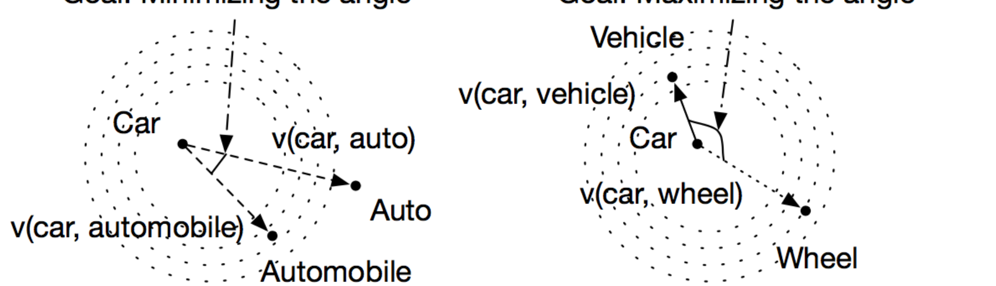
DaSE Data Science & Engineering

SphereRE: Learning Objective (2)

- Learning Objective in the Hyperspherical Relation Space

$$J_g = \sum_{i,j}^{D \cup U} \delta(r_i, r_j) g(f_i(\vec{x}_i) - \vec{x}_i, f_j(\vec{x}_j) - \vec{x}_j)$$

- $\delta(r_i, r_j) = \begin{cases} 1, (x_i, y_i), (x_j, y_j)\ have\ the\ same\ relation\ type \\ -1, (x_i, y_i), (x_j, y_j)\ have\ different\ relation\ types \end{cases}$



Case i) Same Lexical Relation Type   Case ii) Different Lexical Relation Types

- General Learning Objective of SphereRE

$$J(\Theta) = J_f + \lambda_1 J_g + \lambda_2 ||\Theta||^2$$

DaSE Data Science & Engineering

SphereRE: Relation-aware Semantic Projection

- Learning J_f
  - For each lexical relation type $r_m \in R$

$$J_m = \sum_{i=1}^{|D|} I(r_i = r_m)\|M_m\vec{x}_i - \vec{y}_i\|^2 + \mu\|M_m\|_F^2$$

  - Closed-form solution

$$M_m^* = \arg\min_{M_m} J_m = (X_m^T X_m + \mu E)^{-1} X_m^T Y_m$$

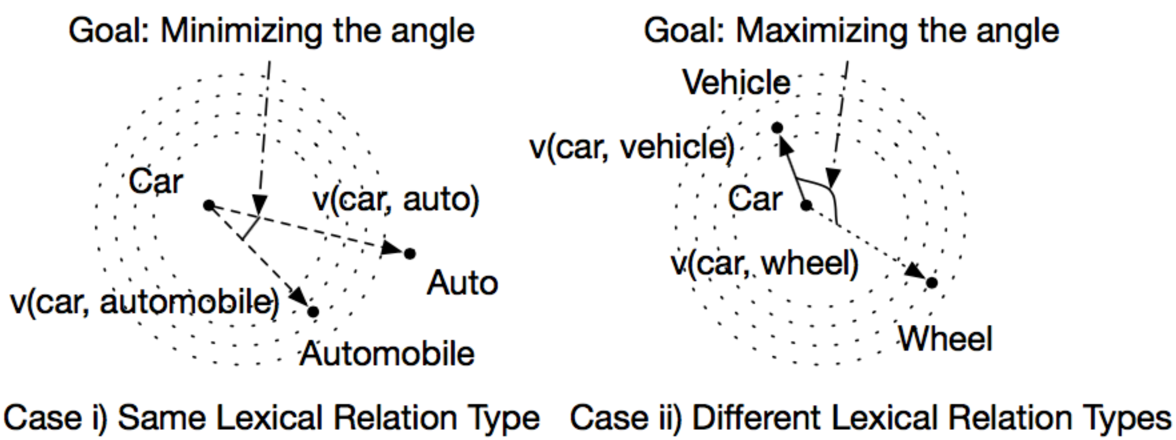
- Approximating the probabilistic distribution over all lexical relation types R w.r.t. $(x_i, y_i) \in U$
  - Train a logistic regression classifier using the feature set

$$\mathcal{F}(x_i, y_i) = (M_1\vec{x}_i - \vec{y}_i) \oplus \cdots \oplus (M_{|R|}\vec{x}_i - \vec{y}_i)$$

DaSE Data Science & Engineering

SphereRE: Relation Representation Learning (1)

- Approximating J_g
  - Learning a SphereRE vector $\vec{r}_i$ for each $(x_i, y_i) \in D \cup U$



  - Re-writing J_g via negative log likelihood

$$J_g' = - \sum_{(x_i,y_i)\in D\cup U} \sum_{(x_j,y_j)\in Nb(x_i,y_i)} \log \Pr((x_j, y_j)|\vec{r}_i)$$

Similar to node2vec!

DaSE Data Science & Engineering

SphereRE: Relation Representation Learning (2)

- Minimizing J'_g by random walk based sampling
  - Sampling probability

$$\Pr((x_j, y_j)|(x_i, y_i)) = \frac{w_{i,j}}{\sum_{(x'_j, y'_j) \in D_{mini}} w_{i,j'}}$$

Condition	Value of $w_{i,j}$
$(x_i, y_i) \in D, (x_j, y_j) \in D, r_i = r_j$	1
$(x_i, y_i) \in D, (x_j, y_j) \in D, r_i \neq r_j$	0
$(x_i, y_i) \in D, (x_j, y_j) \in U, r_i = r_m$	$\frac{1}{2} p_{j,m} (\cos(M_m \vec{x}_i - \vec{x}_i, M_m \vec{x}_j - \vec{x}_j) + 1)$
$(x_i, y_i) \in U, (x_j, y_j) \in D, r_j = r_m$	$\frac{1}{2} p_{i,m} (\cos(M_m \vec{x}_i - \vec{x}_i, M_m \vec{x}_j - \vec{x}_j) + 1)$
$(x_i, y_i) \in U, (x_j, y_j) \in U$	$\frac{1}{2} \sum_{r_m \in R} p_{i,m} p_{j,m} \cdot (\cos(M_m \vec{x}_i - \vec{x}_i, M_m \vec{x}_j - \vec{x}_j) + 1)$

DaSE Data Science & Engineering

SphereRE: Relation Representation Learning (3)

- Overall Procedure of Learning SphereRE Vectors

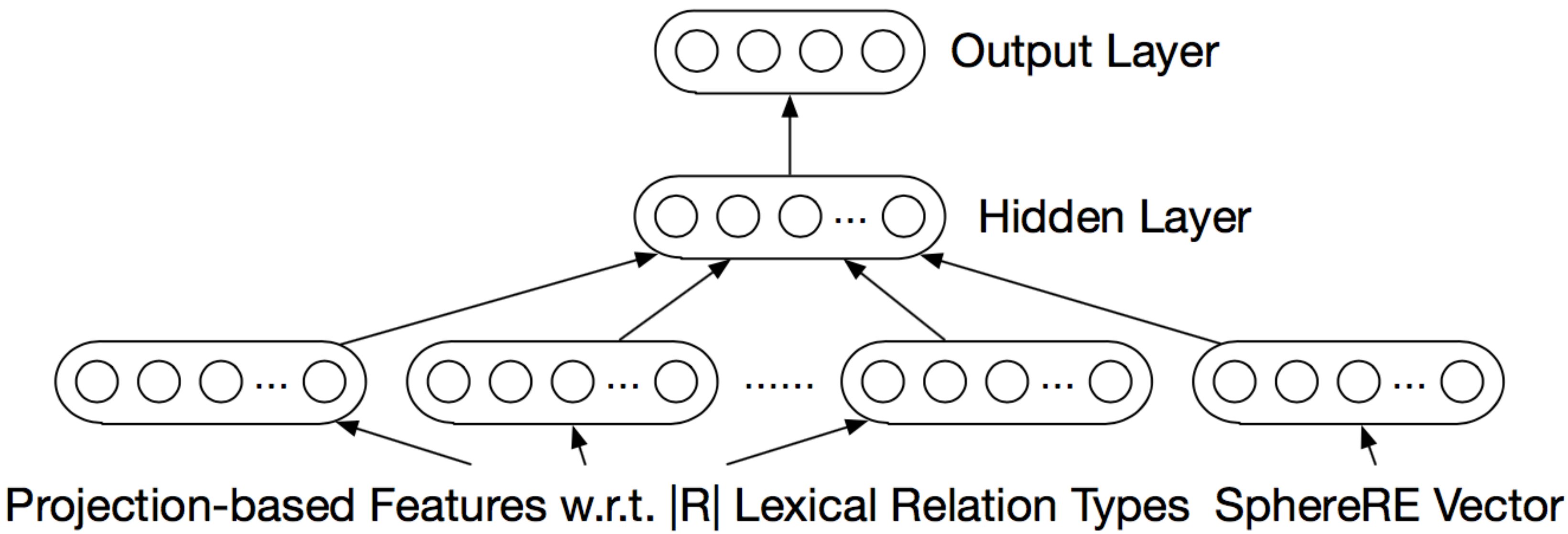
Algorithm 1 SphereRE Learning

1: **for** each $(x_i, y_i) \in D \cup U$ **do**
2: Randomly initialize SphereRE vector $\vec{r}_i$;
3: **end for**
4: **for** $i = 1$ to max iteration **do**
5: Sample a sequence based on Eq. (3): $\mathcal{S} = \{(x_1, y_1), (x_2, y_2), \cdots, (x_{|\mathcal{S}|}, y_{|\mathcal{S}|})\}$;
6: Update all SphereRE vectors $\vec{r}_i$ by minimizing $-\sum_{(x_i, y_i) \in \mathcal{S}} \sum_{j=i-l(j \neq i)}^{i+l} \log \Pr((x_j, y_j)|\vec{r}_i)$;
7: **end for**

DaSE Data Science & Engineering

SphereRE: Lexical Relation Classification

- Train a feed-forward neural network over all the features to predict lexical relations



Experiments (1)

- Datasets and Experimental Settings
  - Word embeddings: fastText embeddings, $d = 300$
  - Default parameters settings:
    - $\mu = 0.001$, $d_r = 300$, $|D_{mini}| = 20$, $|S| = 100$, $\gamma = 2$, $l = 3$
  - Five datasets:

Relation	K&H+N	BLESS	ROOT09	EVALution	CogALex
Antonym	-	-	-	1,600	601
Attribute	-	2,731	-	1,297	-
Co-hyponym	25,796	3,565	3,200	-	-
Event	-	3,824	-	-	-
Holonym	-	-	-	544	-
Hypernym	4,292	1,337	3,190	1,880	637
Meronym	1,043	2,943	-	654	387
Random	26,378	12,146	6,372	-	5,287
Substance meronym	-	-	-	317	-
Synonym	-	-	-	1,086	402
All	57,509	26,546	12,762	7,378	7,314

Experiments (2)

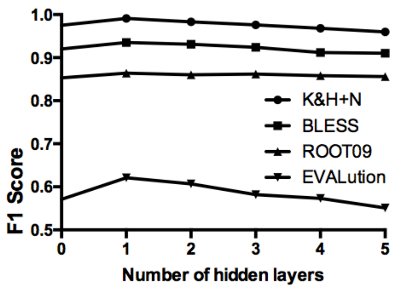
- General Performance over Four Public Datasets
  - SphereRE outperforms all the baselines in terms of F1 scores.

Method↓ Dataset→	K&H+N			BLESS			ROOT09			EVALution		
	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1
Concat	0.909	0.906	0.904	0.811	0.812	0.811	0.636	0.675	0.646	0.531	0.544	0.525
Concat$_h$	0.983	0.984	0.983	0.891	0.889	0.889	0.712	0.721	0.716	0.57	0.573	0.571
Diff	0.888	0.886	0.885	0.801	0.803	0.802	0.627	0.655	0.638	0.521	0.531	0.528
Diff$_h$	0.941	0.942	0.941	0.861	0.859	0.860	0.683	0.692	0.686	0.536	0.54	0.539
NPB	0.713	0.604	0.55	0.759	0.756	0.755	0.788	0.789	0.788	0.53	0.537	0.503
LexNET	0.985	0.986	0.985	0.894	0.893	0.893	0.813	0.814	0.813	0.601	0.607	0.6
LexNET$_h$	0.984	0.985	0.984	0.895	0.892	0.893	0.812	0.816	0.814	0.589	0.587	0.583
NPB+Aug	-	-	0.897	-	-	0.842	-	-	0.778	-	-	0.489
LexNET+Aug	-	-	0.970	-	-	0.927	-	-	0.806	-	-	0.545
SphereRE	**0.990**	**0.989**	**0.990**	**0.938**	**0.938**	**0.938**	**0.860**	**0.862**	**0.861**	**0.62**	**0.621**	**0.62**
Improvement	-	-	**0.5%↑**	-	-	**1.1%↑**	-	-	**4.7%↑**	-	-	**2.0%↑**

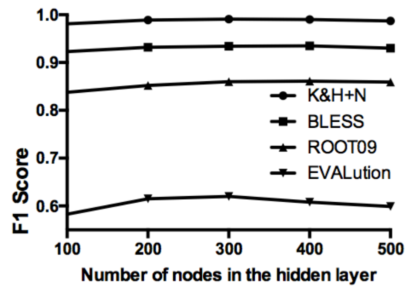
DaSE Data Science & Engineering

Experiments (3)

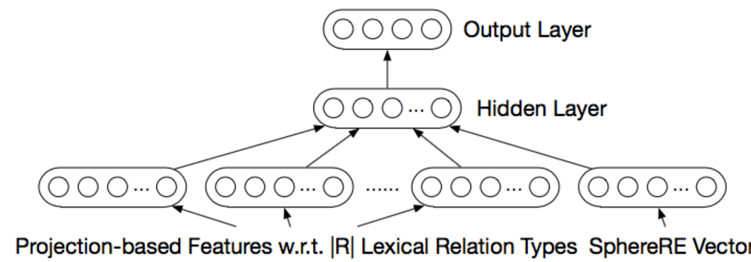
- Detailed analysis of SphereRE
  - Network structure analysis



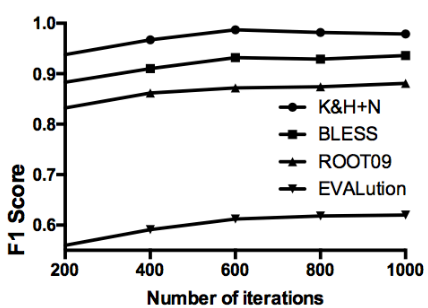
(a) Varying #hidden layers



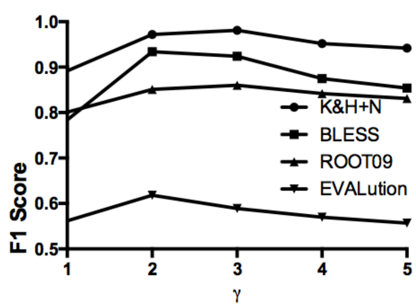
(b) Varying #nodes



  - MC sampling analysis



(a) Varying #iterations



(b) Varying γ

Algorithm 1 SphereRE Learning

for each $(x_i, y_i) \in D \cup U$ **do**
  Randomly initialize SphereRE vector $\vec{r}_i$;
end for
for $i = 1$ to max iteration **do**
  Sample a sequence based on Eq. (3): $\mathcal{S} = \{(x_1, y_1), (x_2, y_2), \cdots, (x_{|\mathcal{S}|}, y_{|\mathcal{S}|})\}$;
  Update all SphereRE vectors $\vec{r}_i$ by minimizing $-\sum_{(x_i,y_i)\in\mathcal{S}} \sum_{j=i-l(j\neq i)}^{i+l} \log \Pr((x_j, y_j)|\vec{r}_i)$;
end for

Experiments (4)

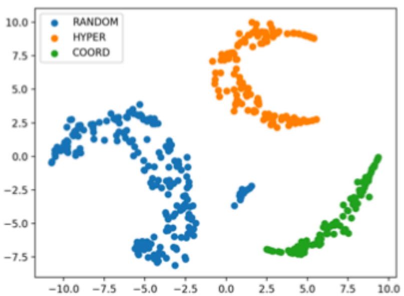
- Experiments over the CogALex-V Shared Task (Subtask 2)
  - Consider random relations as noise, discarding it from the averaged F1 score.
  - Enforce the lexical spilt of the training and testing sets.
  - SphereRE outperforms previous systems in the shared task.

Method↓ Relation→	SYN	ANT	HYP	MER	All
Attia et al. (2016)	0.204	0.448	0.491	0.497	0.423
Shwartz and Dagan (2016)	**0.297**	0.425	0.526	0.493	0.445
Glavas and Vulic (2018)	0.221	**0.504**	0.498	0.504	0.453
SphereRE	0.286	0.479	**0.538**	**0.539**	**0.471**

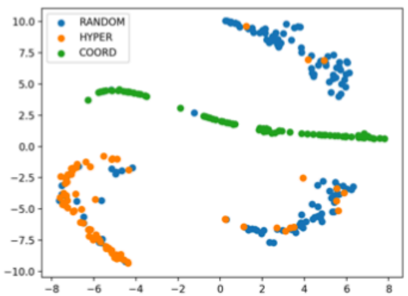
DaSE Data Science & Engineering

Experiments (5)

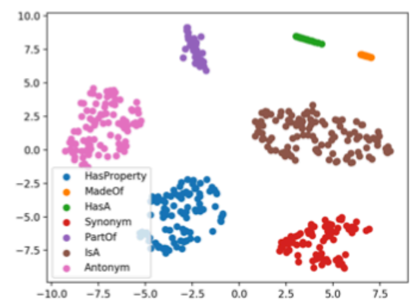
- Visualization of SphereRE Vectors



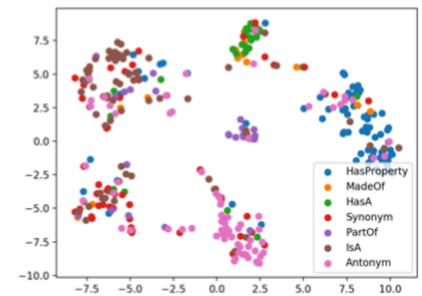
(a) ROOT09 (Training)



(b) ROOT09 (Testing)



(c) EVALution (Training)



(d) EVALution (Testing)

Conclusion

- Model
  - SphereRE: A distributional model for lexical relation classification based on hyperspherical relation embeddings
- Result
  - Outperforming previous baselines on four public datasets and the CogALex-V shared task
- Future Work
  - Dealing with datasets containing a relatively large number of lexical relation types and random term pairs
  - Improving the the mapping technique used for relation-aware semantic projection

Thank You!

Questions & Answers