SphereRE: Distinguishing Lexical Relations with Hyperspherical Relation Embeddings

Chengyu Wang¹, Xiaofeng He¹*, Aoying Zhou²

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
Introduction (1)

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

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

Examples taken from the CogALex-V shared task
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
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
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**
  - \(f_m(\bar{x}_i)\): 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:
    \[
    J_f = \sum_{i=1}^{\lvert D \rvert} \sum_{r_m \in R} I(r_i = r_m) \| f_m(\bar{x}_i) - \bar{y}_i \|^2
    \]
SphereRE: Learning Objective (2)

• **Learning Objective in the Hyperspherical Relation Space**

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

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

- **General Learning Objective of SphereRE**

\[ J(\Theta) = J_f + \lambda_1 J_g + \lambda_2 \|\Theta\|^2 \]
SphereRE: Relation-aware Semantic Projection

• **Learning** $J_f$
  
  – For each lexical relation type $r_m \in R$
  
  \[
  J_m = \sum_{i=1}^{\left|D\right|} I(r_i = r_m) \| M_m \vec{x}_i - \vec{y}_i \|^2 + \mu \| M_m \|_F^2
  \]

  – 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

  \[
  \mathcal{F}(x_i, y_i) = (M_1 \vec{x}_i - \vec{y}_i) \oplus \cdots \oplus (M_{|R|} \vec{x}_i - \vec{y}_i)
  \]
SphereRE: Relation Representation Learning (1)

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

- Re-writing \( J_g \) via negative log likelihood
  \[
  J'_g = \sum_{(x_i, y_i) \in D \cup U} \sum_{(x_j, y_j) \in Nb(x_i, y_i)} \log \Pr((x_j, y_j) | \tilde{r}_i)
  \]

Similar to node2vec!
SphereRE: Relation Representation Learning (2)

• Minimizing $J'_g$ by random walk based sampling
  – Sampling probability

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

<table>
<thead>
<tr>
<th>Condition</th>
<th>Value of $w_{i,j}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$(x_i, y_i) \in D, (x_j, y_j) \in D, r_i = r_j$</td>
<td>1</td>
</tr>
<tr>
<td>$(x_i, y_i) \in D, (x_j, y_j) \in D, r_i \neq r_j$</td>
<td>0</td>
</tr>
<tr>
<td>$(x_i, y_i) \in D, (x_j, y_j) \in U, r_i = r_m$</td>
<td>$\frac{1}{2}p_{i,m}(\cos(M_m \bar{x}_i - \bar{x}_i, M_m \bar{x}_j - \bar{x}_j) + 1)$</td>
</tr>
<tr>
<td>$(x_i, y_i) \in U, (x_j, y_j) \in D, r_j = r_m$</td>
<td>$\frac{1}{2}p_{i,m}(\cos(M_m \bar{x}_i - \bar{x}_i, M_m \bar{x}_j - \bar{x}_j) + 1)$</td>
</tr>
<tr>
<td>$(x_i, y_i) \in U, (x_j, y_j) \in U$</td>
<td>$\frac{1}{2} \sum_{r_m \in R} p_{i,m}p_{j,m} \cdot (\cos(M_m \bar{x}_i - \bar{x}_i, M_m \bar{x}_j - \bar{x}_j) + 1)$</td>
</tr>
</tbody>
</table>
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 \(\bar{r}_i\);
3: **end for**
4: **for** \(i = 1\) to max iteration **do**
5: Sample a sequence based on Eq. (3):
   \[
   S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_{|S|}, y_{|S|})\}:
   \]
6: Update all SphereRE vectors \(\bar{r}_i\) by minimizing
   \[
   - \sum_{(x_i, y_i) \in S} \sum_{j=i-l}^{i+l} \log \Pr((x_j, y_j) | \bar{r}_i);
   \]
7: **end for**
SphereRE: Lexical Relation Classification

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

![Diagram of a feed-forward neural network](image-url)
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:

<table>
<thead>
<tr>
<th>Relation</th>
<th>K&amp;H+N</th>
<th>BLESS</th>
<th>ROOT09</th>
<th>EVALution</th>
<th>CogALex</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antonym</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1,600</td>
<td>601</td>
</tr>
<tr>
<td>Attribute</td>
<td>-</td>
<td>2,731</td>
<td>-</td>
<td>1,297</td>
<td>-</td>
</tr>
<tr>
<td>Co-hyponym</td>
<td>25,796</td>
<td>3,565</td>
<td>3,200</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Event</td>
<td>-</td>
<td>3,824</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Holonym</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>544</td>
<td>-</td>
</tr>
<tr>
<td>Hyponym</td>
<td>4,292</td>
<td>1,337</td>
<td>3,190</td>
<td>1,880</td>
<td>637</td>
</tr>
<tr>
<td>Meronym</td>
<td>1,043</td>
<td>2,943</td>
<td>-</td>
<td>654</td>
<td>387</td>
</tr>
<tr>
<td>Random</td>
<td>26,378</td>
<td>12,146</td>
<td>6,372</td>
<td>-</td>
<td>5,287</td>
</tr>
<tr>
<td>Substance meronym</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>317</td>
<td>-</td>
</tr>
<tr>
<td>Synonym</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1,086</td>
<td>402</td>
</tr>
<tr>
<td><strong>All</strong></td>
<td>57,509</td>
<td>26,546</td>
<td>12,762</td>
<td>7,378</td>
<td>7,314</td>
</tr>
</tbody>
</table>
Experiments (2)

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

<table>
<thead>
<tr>
<th>Method/Dataset</th>
<th>K&amp;H+N Pre</th>
<th>K&amp;H+N Rec</th>
<th>K&amp;H+N F1</th>
<th>BLESS Pre</th>
<th>BLESS Rec</th>
<th>BLESS F1</th>
<th>ROOT09 Pre</th>
<th>ROOT09 Rec</th>
<th>ROOT09 F1</th>
<th>EVALution Pre</th>
<th>EVALution Rec</th>
<th>EVALution F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concat</td>
<td>0.909</td>
<td>0.906</td>
<td>0.904</td>
<td>0.811</td>
<td>0.812</td>
<td>0.811</td>
<td>0.636</td>
<td>0.675</td>
<td>0.646</td>
<td>0.531</td>
<td>0.544</td>
<td>0.525</td>
</tr>
<tr>
<td>Concat$_h$</td>
<td>0.983</td>
<td>0.984</td>
<td>0.983</td>
<td>0.891</td>
<td>0.889</td>
<td>0.889</td>
<td>0.712</td>
<td>0.721</td>
<td>0.716</td>
<td>0.57</td>
<td>0.573</td>
<td>0.571</td>
</tr>
<tr>
<td>Diff</td>
<td>0.888</td>
<td>0.886</td>
<td>0.885</td>
<td>0.801</td>
<td>0.803</td>
<td>0.802</td>
<td>0.627</td>
<td>0.655</td>
<td>0.638</td>
<td>0.521</td>
<td>0.531</td>
<td>0.528</td>
</tr>
<tr>
<td>Diff$_h$</td>
<td>0.941</td>
<td>0.942</td>
<td>0.941</td>
<td>0.861</td>
<td>0.859</td>
<td>0.860</td>
<td>0.683</td>
<td>0.692</td>
<td>0.686</td>
<td>0.536</td>
<td>0.54</td>
<td>0.539</td>
</tr>
<tr>
<td>NPB</td>
<td>0.713</td>
<td>0.604</td>
<td>0.55</td>
<td>0.759</td>
<td>0.756</td>
<td>0.755</td>
<td>0.788</td>
<td>0.789</td>
<td>0.788</td>
<td>0.53</td>
<td>0.537</td>
<td>0.503</td>
</tr>
<tr>
<td>LexNET</td>
<td>0.985</td>
<td>0.986</td>
<td>0.985</td>
<td>0.894</td>
<td>0.893</td>
<td>0.893</td>
<td>0.813</td>
<td>0.814</td>
<td>0.813</td>
<td>0.601</td>
<td>0.607</td>
<td>0.6</td>
</tr>
<tr>
<td>LexNET$_h$</td>
<td>0.984</td>
<td>0.985</td>
<td>0.984</td>
<td>0.895</td>
<td>0.892</td>
<td>0.893</td>
<td>0.812</td>
<td>0.816</td>
<td>0.814</td>
<td>0.589</td>
<td>0.587</td>
<td>0.583</td>
</tr>
<tr>
<td>NPB+Aug</td>
<td>-</td>
<td>-</td>
<td>0.897</td>
<td>-</td>
<td>-</td>
<td>0.842</td>
<td>-</td>
<td>-</td>
<td>0.778</td>
<td>-</td>
<td>-</td>
<td>0.489</td>
</tr>
<tr>
<td>LexNET+Aug</td>
<td>-</td>
<td>-</td>
<td>0.970</td>
<td>-</td>
<td>-</td>
<td>0.927</td>
<td>-</td>
<td>-</td>
<td>0.806</td>
<td>-</td>
<td>-</td>
<td>0.545</td>
</tr>
<tr>
<td>SphereRE</td>
<td>0.990</td>
<td>0.989</td>
<td>0.990</td>
<td>0.938</td>
<td>0.938</td>
<td>0.938</td>
<td>0.860</td>
<td>0.862</td>
<td>0.861</td>
<td>0.62</td>
<td>0.621</td>
<td>0.62</td>
</tr>
<tr>
<td>Improvement</td>
<td>-</td>
<td>-</td>
<td>0.5%↑</td>
<td>-</td>
<td>-</td>
<td>1.1%↑</td>
<td>-</td>
<td>-</td>
<td>4.7%↑</td>
<td>-</td>
<td>-</td>
<td>2.0%↑</td>
</tr>
</tbody>
</table>
Experiments (3)

- **Detailed analysis of SphereRE**
  - Network structure analysis
    - [Graphs showing F1 score vs. number of hidden layers and nodes in the hidden layer.]
  - MC sampling analysis
    - [Graphs showing F1 score vs. number of iterations and γ.]

---

**Algorithm 1 SphereRE Learning**

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

<table>
<thead>
<tr>
<th>Method</th>
<th>Relation</th>
<th>SYN</th>
<th>ANT</th>
<th>HYP</th>
<th>MER</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attia et al. (2016)</td>
<td></td>
<td>0.204</td>
<td>0.448</td>
<td>0.491</td>
<td>0.497</td>
<td>0.423</td>
</tr>
<tr>
<td>Shwartz and Dagan (2016)</td>
<td></td>
<td>0.297</td>
<td>0.425</td>
<td>0.526</td>
<td>0.493</td>
<td>0.445</td>
</tr>
<tr>
<td>Glavas and Vulic (2018)</td>
<td></td>
<td>0.221</td>
<td>0.504</td>
<td>0.498</td>
<td>0.504</td>
<td>0.453</td>
</tr>
<tr>
<td>SphereRE</td>
<td></td>
<td>0.286</td>
<td>0.479</td>
<td>0.538</td>
<td>0.539</td>
<td>0.471</td>
</tr>
</tbody>
</table>
Experiments (5)

- **Top-k Similar Relation Retrieval based on SphereRE Vectors**
  - Evaluation metric: Average Precision@k
  - Near prefect performance over the training sets
  - Not very satisfying for unbalanced datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>AP@1</th>
<th>AP@5</th>
<th>AP@10</th>
<th>AP@1</th>
<th>AP@5</th>
<th>AP@10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Training Set</td>
<td></td>
<td>Testing Set</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>K&amp;H+N</td>
<td>0.972</td>
<td>0.954</td>
<td>0.951</td>
<td>0.862</td>
<td>0.844</td>
<td>0.839</td>
</tr>
<tr>
<td>BLESS</td>
<td>0.962</td>
<td>0.950</td>
<td>0.948</td>
<td>0.868</td>
<td>0.830</td>
<td>0.825</td>
</tr>
<tr>
<td>ROOT09</td>
<td>0.987</td>
<td>0.993</td>
<td>0.989</td>
<td>0.814</td>
<td>0.789</td>
<td>0.828</td>
</tr>
<tr>
<td>EVALution</td>
<td>0.988</td>
<td>0.987</td>
<td>0.982</td>
<td>0.653</td>
<td>0.650</td>
<td>0.697</td>
</tr>
<tr>
<td>CogALex</td>
<td>0.953</td>
<td>0.904</td>
<td>0.918</td>
<td>0.631</td>
<td>0.628</td>
<td>0.649</td>
</tr>
</tbody>
</table>
Experiments (6)

• Visualization of SphereRE Vectors
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 mapping technique used for relation-aware semantic projection
Thank You!

Questions & Answers