A Family of Fuzzy Orthogonal Projection Models for Monolingual and Cross-lingual Hypernymy Prediction

Chengyu Wang¹, Yan Fan¹, 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
• Related Work
• Monolingual Model
  – Multi-Wahba Projection (MWP)
• Cross-lingual Models
  – Transfer MWP (TMWP)
  – Iterative Transfer MWP (ITMWP)
• Experiments
  – Monolingual Experiments
  – Cross-lingual Experiments
• Conclusion and Future Work
Introduction (1)

- Hypernymy ("is-a") relations are important for NLP and Web applications.
  - Semantic resource construction: semantic hierarchies, taxonomies, knowledge graphs, etc.
  - Web-based applications: query understanding, post-search navigation, personalized recommendation, etc.

A simple example of taxonomy
Introduction (2)

• Research challenges for predicting hypernymy relations between words:
  – Monolingual hypernymy prediction
    • Pattern-based approaches: have low recall
    • Distributional classifiers: suffer from the “lexical memorization” problem
  – Cross-lingual hypernymy prediction
    • The small size of training sets for lower-resourced languages
    • Not sufficient research in this area
Task 1: Monolingual Hypernymy Prediction

Training Set

(animal, fruit) ✓
(dog, animal) ✓
(tree, flower) ×
(actor, desk) ×

Training

Testing Set

(professor, job) ?
(coffee, juice) ?
(food, sandwich) ?

Task 2: Cross-lingual Hypernymy Prediction

Training Set (en, large)

(animal, fruit) ✓
(dog, animal) ✓
(tree, flower) ×
(actor, desk) ×

Training

Testing Set (fr)

(thé, boisson) ?
(chien, ours) ?
(mars, mois) ?

Translation (fr-en)

poulet: chicken
viande: meat
maison: house
eau: water
thé: tea
boisson: drink
chien: dog
ours: bear
mars: March
mois: month

* For human reference only, unseen by algorithm
Related Work (1)

• **Monolingual hypernymy prediction**
  
  – Pattern based approaches:
    
    • Handcraft patterns: high accuracy, low coverage
      
      – Hearst Patterns: NP1 such as NP2
    
    • Automatic generated patterns: higher coverage, lower accuracy
    
    • High language dependency
  
  – Distributional approaches:
    
    • Unsupervised distributional measures: relatively low precision
    
    • Supervised distributional classifiers: suffer from the “lexical memorization” problem
Related Work (2)

• **Cross-lingual hypernymy prediction**
  
  – Learning multi-lingual taxonomies based on existing knowledge sources
    
    • YAGO3: Multi-lingual Wikipedia + WordNet
    
    • More precise but have limited scope constrained by sources

  – This task has not been extensively studied for lower-resourced languages.
Monolingual Model (1)

- **Basic Notations**
  - Hypernymy training set $D^{(+)} = \{(x_i, y_i^{(+)})\}$
  - Non-hypernymy training set $D^{(-)} = \{(x_i, y_i^{(-)})\}$

- **Orthogonal Projection Model for Hypernymy Relations**
  - Objective function
    \[
    \min \sum_{i=1}^{\left|D^{(+)}\right|} \|M\tilde{x}_i - \tilde{y}_i^{(+)}\|^2 \quad \text{s. t.} \quad M^TM = I
    \]
    
    - Normalized embeddings
    - Adding orthogonal constraints to guarantee normalization!
    
    - It does not consider the complicated linguistic regularities of hypernymy relations.
Monolingual Model (2)

- **Fuzzy Orthogonal Projection Model for Hypernymy Relations**
  
  - Apply K-means to $D^{(+)}$ over the features $\tilde{x}_i - \tilde{y}_i^{(+)}$ with cluster centroids as $\tilde{c}_1^{(+)}, \tilde{c}_2^{(+)}, \ldots, \tilde{c}_K^{(+)}$.
  
  - Compute the weight of $(x_i, y_i^{(+)})$ in $D^{(+)}$ w.r.t. the $j$th cluster.

  $$a_{i,j}^{(+)} = \frac{\cos(\tilde{x}_i - \tilde{y}_i^{(+)}, \tilde{c}_j^{(+)})}{\sum_{i' = 1}^{|D^{(+)}|} \cos(\tilde{x}_{i'} - \tilde{y}_{i'}^{(+)}, \tilde{c}_j^{(+)})}$$

  - Objective function

  $$\min \tilde{J}(M^{(+)}) = \frac{1}{2} \sum_{j=1}^K \sum_{i=1}^{|D^{(+)}|} a_{i,j}^{(+)} \| M_j^{(+)} \tilde{x}_i - \tilde{y}_i^{(+)} \|^2$$

  | Multi-Wahba Projection (MWP) |

  s. t. $M_j^{(+)^T} M_j^{(+)} = I$, $\sum_{i=1}^{|D^{(+)}|} a_{i,j}^{(+)} = 1, j = 1, \ldots, K$
Some Observations

- **Objective Function**

\[
\min \hat{f}(M^{(+)}) = \frac{1}{2} \sum_{j=1}^{K} \sum_{i=1}^{\lvert D^{(+)} \rvert} a_{i,j}^{(+)} \| M_{j}^{(+)} \hat{x}_{i} - \hat{y}_{i}^{(+)} \|^2
\]

s. t. \( M_{j}^{(+)^T} M_{j}^{(+)} = I, \sum_{i=1}^{\lvert D^{(+)} \rvert} a_{i,j}^{(+)} = 1, j = 1, \ldots, K \)

- The optimization of different matrices is independent from each other!

\[
\min J(M_{j}^{(+)}) = \frac{1}{2} \sum_{i=1}^{\lvert D^{(+)} \rvert} a_{i,j}^{(+)} \| M_{j}^{(+)} \hat{x}_{i} - \hat{y}_{i}^{(+)} \|^2
\]

s. t. \( M_{j}^{(+)^T} M_{j}^{(+)} = I, \sum_{i=1}^{\lvert D^{(+)} \rvert} a_{i,j}^{(+)} = 1 \)
Monolingual Model (3)

• Solving the MWP Problem
  – Consider the $j$th cluster only:

  \[
  \min J(M_j^{(+)}) = \frac{1}{2} \sum_{i=1}^{|D^{(+)}|} a_{i,j}^{(+)} \| M_j^{(+)} \vec{x}_i - \vec{y}_i^{(+)} \|^2
  \]

  s. t. $M_j^{(+)}^T M_j^{(+)} = I$, $\sum_{i=1}^{|D^{(+)}|} a_{i,j}^{(+)} = 1$

  – An SVD-based closed-form solution:

  (1) $B_j = \sum_{i=1}^{|D^{(+)}|} a_{i,j}^{(+)} \vec{y}_i^{(+)} \vec{x}_i^T$; Refer to the paper for the proof of correctness.

  (2) $\text{SVD}(B_j) = U_j \Sigma_j V_j^T$;

  (3) $R_j = \text{diag}(1, \ldots, 1, \text{det}(U_j) \text{det}(V_j))$;

  (4) $M_j^{(+)} = U_j R_j V_j^T$;
Monolingual Model (4)

- **Overall Procedure**
  - Learning hypernymy projections
    \[
    \min \tilde{j}(\mathcal{M}^{(+)}) = \frac{1}{2} \sum_{j=1}^{K} \sum_{i=1}^{|D^{(+)}|} a_{i,j}^{(+)} \| \mathbf{M}_j^{(+)} \mathbf{x}_i - \mathbf{y}_i^{(+)} \|^2
    \]
    \[
    \text{s. t. } \mathbf{M}_j^{(+)^T} \mathbf{M}_j^{(+)} = \mathbf{I}, \sum_{i=1}^{|D^{(+)}|} a_{i,j}^{(+)} = 1, j = 1, \ldots, K
    \]
  - Learning non-hypernymy projections
    \[
    \min \tilde{j}(\mathcal{M}^{(-)}) = \frac{1}{2} \sum_{j=1}^{K} \sum_{i=1}^{|D^{(-)}|} a_{i,j}^{(-)} \| \mathbf{M}_j^{(-)} \mathbf{x}_i - \mathbf{y}_i^{(-)} \|^2
    \]
    \[
    \text{s. t. } \mathbf{M}_j^{(-)^T} \mathbf{M}_j^{(-)} = \mathbf{I}, \sum_{i=1}^{|D^{(-)}|} a_{i,j}^{(-)} = 1, j = 1, \ldots, K
    \]
Monolingual Model (5)

- **Overall Procedure**
  - Training the projection-based neural network
Cross-lingual Models (1)

• **Basic Notations**
  
  – Hypernymy training sets
    
    • Source language: $D_S^{(+)} \quad |D_S^{(+)}| \gg |D_T^{(+)}|$
    
    • Target language: $D_T^{(+)}$
  
  – Non-hypernymy training sets
    
    • Source language: $D_S^{(-)} \quad |D_S^{(-)}| \gg |D_T^{(-)}|$
    
    • Target language: $D_T^{(-)}$
  
  – Unlabeled set of the target language: $U_T = \{(x_i, y_i)\}$
Cross-lingual Models (2)

- **Transfer MWP Model (TMWP)**
  - Learning hypernymy projections

\[
\min_{j} J(M^{(+)}) = \frac{\beta}{2} \sum_{j=1}^{K} \sum_{i=1}^{|D^{(+)}_S|} a_{i,j}^{(+)} y_i^{(+)} \| M_j^{(+)} S x_i - S y_i^{(+)} \|^2
\]

\[
+ \frac{1 - \beta}{2} \sum_{j=1}^{K} \sum_{i=1}^{|D^{(+)}_T|} a_{i,j}^{(+)} \| M_j^{(+)} x_i - y_i^{(+)} \|^2
\]

s. t. \( M_j^{(+)}^T M_j^{(+)} = I \), \( \sum_{i=1}^{|D^{(+)}_S|} a_{i,j}^{(+)} y_i^{(+)} = 1 \), \( \sum_{i=1}^{|D^{(+)}_T|} a_{i,j}^{(+)} = 1 \), \( j = 1, \ldots, K \)

- \( \beta \): controls the importance of training sets of source and target languages.

- \( \gamma_i^{(+)} \): controls the individual weight of each training instance of the source language

\( S \): maps the embeddings of the source language to the target language by Bilingual Lexicon Induction
Cross-lingual Models (3)

• Transfer MWP Model (TMWP)
  – Hypernymy projections in TMWP can also be converted into a high-dimensional Wahba’s problem.
  – The SVD-based closed form solution:

\[ B_j = \beta \sum_{i=1}^{D_s^{(+)}} \alpha_{i,j}^{(+)} y_i^{(+)} (S\tilde{x}_i)^T + (1-\beta) \sum_{i=1}^{D_T^{(+)}} \alpha_{i,j}^{(+)} \tilde{y}_i^{(+)} \tilde{x}_i^T; \]

\[ \text{SVD}(B_j) = U_j \Sigma_j V_j^T; \]

\[ R_j = \text{diag}(1, \ldots, 1, \det(U_j) \det(V_j)); \]

\[ M_j^{(+)} = U_j R_j V_j^T; \]
Cross-lingual Models (4)

- **Transfer MWP Model (TMWP)**
  - Learning non-hypernymy projections

\[
\min \tilde{J}(M^{(-)}) = \frac{\beta}{2} \sum_{j=1}^{K} \sum_{i=1}^{\left|D_S^{(-)}\right|} a_{i,j}^{(-)} y_i^{(-)} \|M_j^{(-)} S\tilde{x}_i - S\tilde{y}_i^{(-)}\|^2 \\
+ \frac{1 - \beta}{2} \sum_{j=1}^{K} \sum_{i=1}^{\left|D_T^{(-)}\right|} a_{i,j}^{(-)} \|M_j^{(-)} \tilde{x}_i - \tilde{y}_i^{(-)}\|^2 \\
s. t. M_j^{(-)T} M_j^{(-)} = I, \sum_{i=1}^{\left|D_S^{(-)}\right|} a_{i,j}^{(-)} y_i^{(-)} = 1, \sum_{i=1}^{\left|D_T^{(-)}\right|} a_{i,j}^{(-)} = 1, \\
j = 1, \ldots, K
\]

- Training the projection-based neural network
Cross-lingual Models (5)

- **Iterative Transfer MWP Model (ITMWP)**
  - Employ semi-supervised learning for training set augmentation

---

**Algorithm 3 Cross-lingual Hypernymy Prediction (ITMWP)**

1. Train TMWP over $D_S^{(+)}$, $D_S^{(-)}$, $D_T^{(+)}$ and $D_T^{(-)}$ by Algorithm 2;
2. **while** not converge **do**
3. **for** each pair $(x_i, y_i) \in U_T$ **do**
4. **if** $conf(x_i, y_i) > \tau$ **then**
5. **if** $f(x_i, y_i) = \text{HYPERNYMY}$ **then**
6. Update $D_T^{(+)} = D_T^{(+)} \cup \{(x_i, y_i)\}$
7. **else**
8. Update $D_T^{(-)} = D_T^{(-)} \cup \{(x_i, y_i)\}$
9. **end if**
10. Update $U_T = U_T \setminus \{(x_i, y_i)\}$
11. **end if**
12. **end for**
13. Update TMWP over $D_S^{(+)}$, $D_S^{(-)}$, $D_T^{(+)}$ and $D_T^{(-)}$ by Algorithm 2;
14. **end while**
Monolingual Experiments (1)

• **Task 1: Supervised hypernymy detection**
  – MWP outperforms state-of-the-art over two benchmark datasets (BLESS and ENTAILMENT)

<table>
<thead>
<tr>
<th>Method</th>
<th>BLESS</th>
<th>ENTAILMENT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mikolov et al. [24]</td>
<td>0.84</td>
<td>0.83</td>
</tr>
<tr>
<td>Yu et al. [54]</td>
<td>0.90</td>
<td>0.87</td>
</tr>
<tr>
<td>Luu et al. [20]</td>
<td>0.93</td>
<td>0.91</td>
</tr>
<tr>
<td>Nguyen et al. [26]</td>
<td>0.94</td>
<td>0.91</td>
</tr>
<tr>
<td><strong>MWP (Non-orthogonal)</strong></td>
<td><strong>0.95</strong></td>
<td><strong>0.90</strong></td>
</tr>
<tr>
<td><strong>MWP</strong></td>
<td><strong>0.97</strong></td>
<td><strong>0.92</strong></td>
</tr>
</tbody>
</table>


Monolingual Experiments (2)

• **Task 1: Supervised hypernymy detection**
  
  – MWP outperforms state-of-the-art over three domain-specific datasets derived from existing domain-specific taxonomies.

<table>
<thead>
<tr>
<th>Method</th>
<th>ANIMAL</th>
<th>PLANT</th>
<th>VEHICLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mikolov et al. [24]</td>
<td>0.80</td>
<td>0.81</td>
<td>0.82</td>
</tr>
<tr>
<td>Yu et al. [54]</td>
<td>0.67</td>
<td>0.65</td>
<td>0.70</td>
</tr>
<tr>
<td>Luu et al. [20]</td>
<td>0.89</td>
<td>0.92</td>
<td>0.89</td>
</tr>
<tr>
<td>Nguyen et al. [26]*</td>
<td>0.83</td>
<td>0.91</td>
<td>0.83</td>
</tr>
<tr>
<td><strong>MWP (Non-orthogonal)</strong></td>
<td><strong>0.90</strong></td>
<td><strong>0.92</strong></td>
<td><strong>0.87</strong></td>
</tr>
<tr>
<td><strong>MWP</strong></td>
<td><strong>0.92</strong></td>
<td><strong>0.94</strong></td>
<td><strong>0.90</strong></td>
</tr>
</tbody>
</table>
Monolingual Experiments (3)

- **Task 2**: Unsupervised hypernymy classification
  - Hypernymy measure: \( \tilde{s}(x_i, y_i) = \| F^-(\tilde{x}_i, \tilde{y}_i) \|_2 - \| F^+(\tilde{x}_i, \tilde{y}_i) \|_2 \)

<table>
<thead>
<tr>
<th>Measure</th>
<th>BLESS</th>
<th>WBLESS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Santus et al. [31]</td>
<td>0.87</td>
<td>-</td>
</tr>
<tr>
<td>Weeds et al. [49]</td>
<td>-</td>
<td>0.75</td>
</tr>
<tr>
<td>Kiela et al. [15]</td>
<td>0.88</td>
<td>0.75</td>
</tr>
<tr>
<td>Nguyen et al. [26]</td>
<td>0.92</td>
<td>0.87</td>
</tr>
<tr>
<td>Roller et al. [30]</td>
<td>0.96</td>
<td>0.87</td>
</tr>
<tr>
<td><strong>MWP (Non-orthogonal)</strong></td>
<td><strong>0.95</strong></td>
<td><strong>0.89</strong></td>
</tr>
<tr>
<td><strong>MWP</strong></td>
<td><strong>0.97</strong></td>
<td><strong>0.92</strong></td>
</tr>
</tbody>
</table>
Cross-lingual Experiments (1)

- **Dataset Construction**
  - English dataset: combining five human-labeled datasets (Training set)
    - 17,394 hypernymy relations
    - 67,930 non-hypernymy relations
  - Other languages: deriving from the Open Multilingual Wordnet project
    - 20% for training, 20% for development and 60% for testing

<table>
<thead>
<tr>
<th>Relation</th>
<th>Language</th>
<th>fr</th>
<th>zh</th>
<th>ja</th>
<th>it</th>
<th>th</th>
<th>fi</th>
<th>el</th>
</tr>
</thead>
<tbody>
<tr>
<td># Hypernymy relations</td>
<td></td>
<td>4,035</td>
<td>2,962</td>
<td>1,448</td>
<td>3,034</td>
<td>1,156</td>
<td>7,157</td>
<td>2,612</td>
</tr>
<tr>
<td># Non-hypernymy relations</td>
<td></td>
<td>8,947</td>
<td>6,382</td>
<td>3,203</td>
<td>6,081</td>
<td>1,977</td>
<td>9,433</td>
<td>1,454</td>
</tr>
</tbody>
</table>
Cross-lingual Experiments (2)

- **Task 1:** Cross-lingual hypernymy direction classification
  - hypernymy vs. reverse-hypernymy

<table>
<thead>
<tr>
<th>Method</th>
<th>fr</th>
<th>zh</th>
<th>ja</th>
<th>it</th>
<th>th</th>
<th>fi</th>
<th>el</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task: cross-lingual hypernymy direction classification</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Santus et al. [31]</td>
<td>0.65</td>
<td>0.65</td>
<td>0.68</td>
<td>0.61</td>
<td>0.63</td>
<td>0.70</td>
<td>0.62</td>
</tr>
<tr>
<td>Weeds et al. [49]</td>
<td>0.76</td>
<td>0.71</td>
<td>0.77</td>
<td>0.76</td>
<td>0.72</td>
<td>0.77</td>
<td>0.70</td>
</tr>
<tr>
<td>Kiela et al. [15]</td>
<td>0.67</td>
<td>0.65</td>
<td>0.71</td>
<td>0.68</td>
<td>0.65</td>
<td>0.70</td>
<td>0.62</td>
</tr>
<tr>
<td>Shwartz et al. [34]</td>
<td>0.79</td>
<td>0.67</td>
<td>0.71</td>
<td>0.72</td>
<td>0.66</td>
<td>0.75</td>
<td>0.66</td>
</tr>
<tr>
<td>TMWP (N)</td>
<td>0.78</td>
<td>0.71</td>
<td>0.75</td>
<td>0.76</td>
<td>0.73</td>
<td>0.76</td>
<td>0.71</td>
</tr>
<tr>
<td>TMWP</td>
<td>0.80</td>
<td>0.72</td>
<td>0.76</td>
<td>0.78</td>
<td>0.75</td>
<td>0.78</td>
<td>0.73</td>
</tr>
<tr>
<td>ITMWP (N)</td>
<td><strong>0.82</strong></td>
<td>0.72</td>
<td>0.76</td>
<td>0.78</td>
<td>0.75</td>
<td><strong>0.81</strong></td>
<td>0.72</td>
</tr>
<tr>
<td>ITMWP</td>
<td>0.81</td>
<td>0.74</td>
<td>0.78</td>
<td>0.81</td>
<td>0.78</td>
<td>0.81</td>
<td>0.75</td>
</tr>
</tbody>
</table>
Cross-lingual Experiments (3)

• **Task 1: Cross-lingual hypernymy detection**
  
  hypernymy vs. non-hypernymy

<table>
<thead>
<tr>
<th>Method</th>
<th>fr</th>
<th>zh</th>
<th>ja</th>
<th>it</th>
<th>th</th>
<th>fi</th>
<th>el</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task: cross-lingual hypernymy detection</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Santus et al. [31]</td>
<td>0.67</td>
<td>0.63</td>
<td>0.67</td>
<td>0.62</td>
<td>0.64</td>
<td>0.62</td>
<td>0.64</td>
</tr>
<tr>
<td>Weeds et al. [49]</td>
<td>0.74</td>
<td>0.66</td>
<td>0.68</td>
<td>0.71</td>
<td>0.62</td>
<td>0.68</td>
<td>0.69</td>
</tr>
<tr>
<td>Kiela et al. [15]</td>
<td>0.70</td>
<td>0.61</td>
<td>0.65</td>
<td>0.68</td>
<td>0.57</td>
<td>0.61</td>
<td>0.67</td>
</tr>
<tr>
<td>Shwartz et al. [34]</td>
<td>0.72</td>
<td>0.66</td>
<td>0.69</td>
<td>0.64</td>
<td>0.66</td>
<td>0.69</td>
<td>0.70</td>
</tr>
<tr>
<td>TMWP (N)</td>
<td>0.72</td>
<td>0.67</td>
<td>0.70</td>
<td>0.70</td>
<td>0.68</td>
<td>0.71</td>
<td>0.70</td>
</tr>
<tr>
<td>TMWP</td>
<td>0.75</td>
<td>0.71</td>
<td>0.76</td>
<td>0.72</td>
<td>0.69</td>
<td>0.72</td>
<td>0.71</td>
</tr>
<tr>
<td>ITMWP (N)</td>
<td>0.72</td>
<td>0.74</td>
<td>0.77</td>
<td><strong>0.74</strong></td>
<td>0.67</td>
<td>0.71</td>
<td>0.72</td>
</tr>
<tr>
<td>ITMWP</td>
<td><strong>0.76</strong></td>
<td>0.73</td>
<td><strong>0.78</strong></td>
<td><strong>0.74</strong></td>
<td><strong>0.72</strong></td>
<td><strong>0.73</strong></td>
<td><strong>0.73</strong></td>
</tr>
</tbody>
</table>
Conclusion

• **Models**
  – Monolingual hypernymy prediction: MWP
  – Cross-lingual hypernymy prediction: TMWP & ITMWP

• **Results**
  – State-of-the-art performance in monolingual experiments
  – Highly effective in cross-lingual experiments

• **Future Works**
  – Predicting multiple types of semantic relations over multiple languages
  – Improving cross-lingual hypernymy prediction via multi-lingual embeddings
Thank You!

Questions & Answers