



THE WEB
CONFERENCE

华东师范大学
EAST CHINA NORMAL UNIVERSITY

计算机科学与
软件工程学院
School of Computer Science
and Software Engineering

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

Chengyu Wang¹, Yan Fan¹, Xiaofeng He^{1*}, Aoying Zhou²

¹ School of Computer Science and Software Engineering,

² School of Data Science and Engineering,

East China Normal University

Shanghai, China



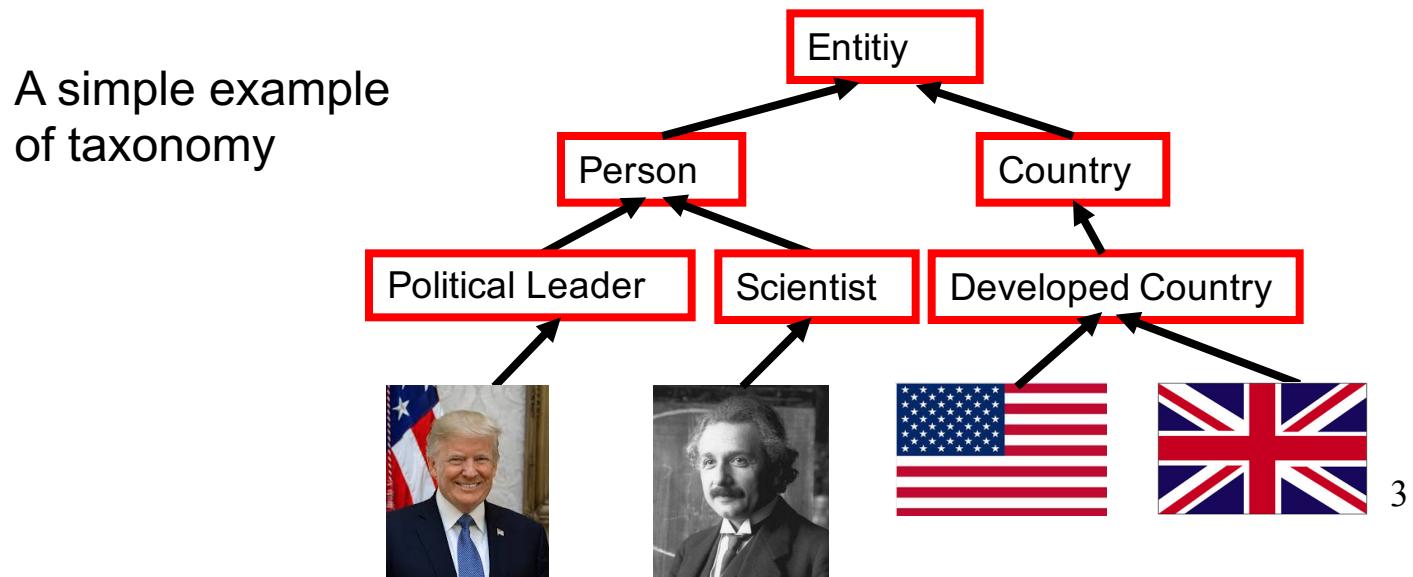
DaSE
Data Science
& Engineering

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)

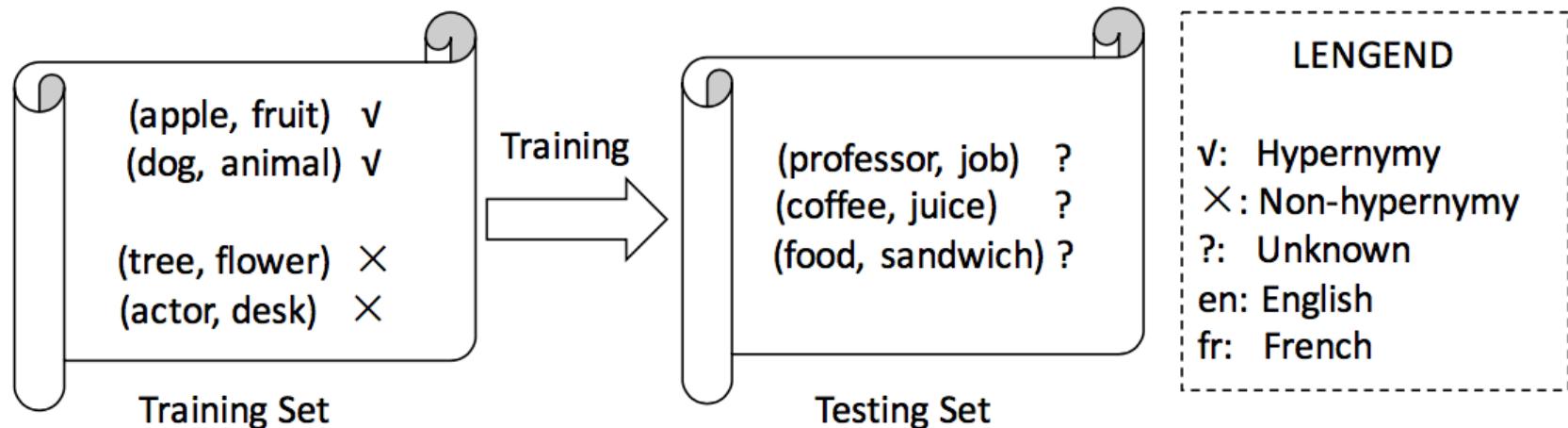
- Hypernymy (“is-a”) relations are important for NLP and Web applications.
 - Semantic resource construction: semantic hierarchies, taxonomies, knowledge graphs, etc.
 - Web-based applications: query understanding, post-search navigation, personalized recommendation, etc.



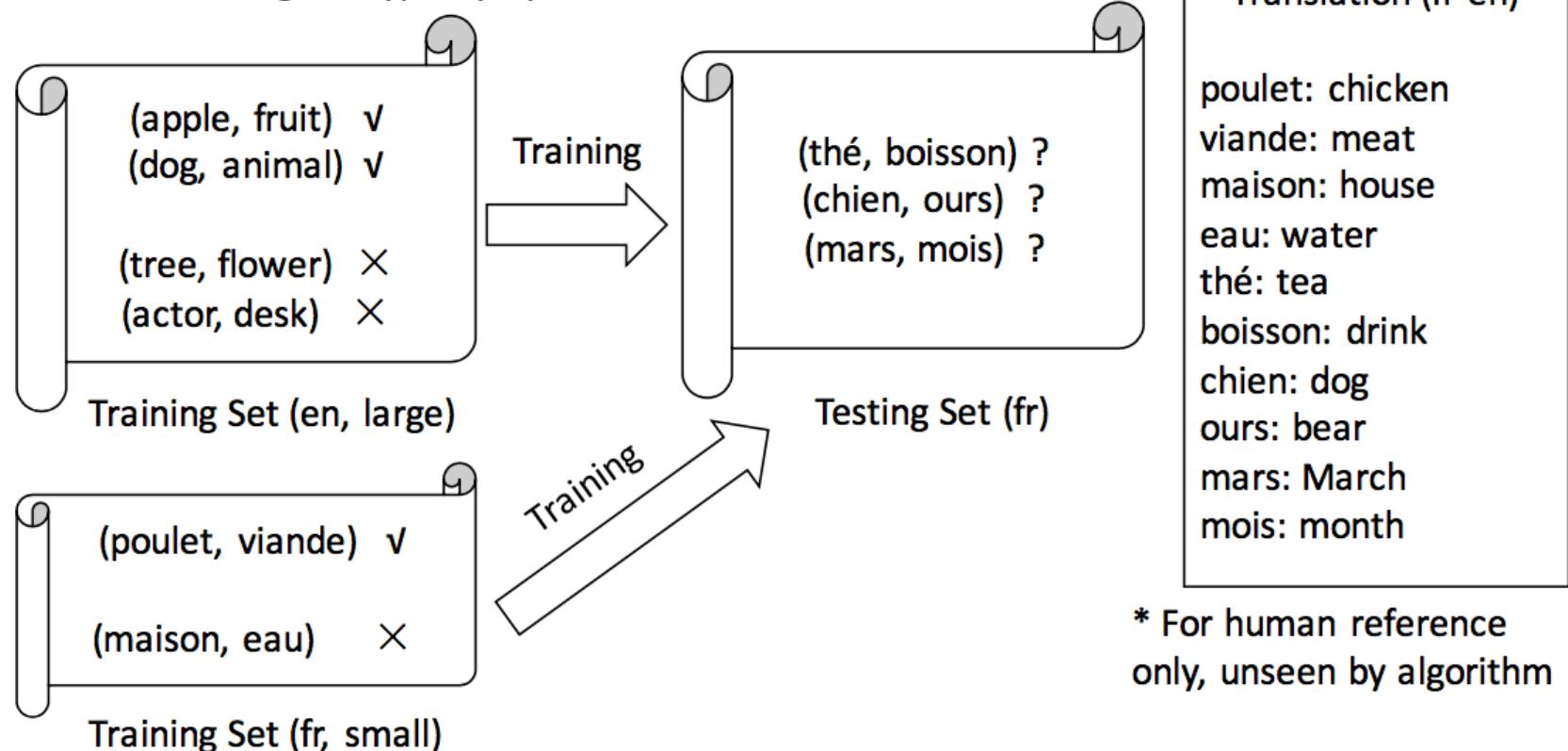
Introduction (2)

- 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



Task 2: Cross-lingual Hypernymy Prediction



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

$$\min \sum_{i=1}^{|D^{(+)}|} \|\mathbf{M}\vec{x}_i - \vec{y}_i^{(+)}\|^2 \text{ s. t. } \mathbf{M}^T \mathbf{M} = \mathbf{I}$$

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 $\vec{x}_i - \vec{y}_i^{(+)}$ with cluster centroids as $\vec{c}_1^{(+)}, \vec{c}_2^{(+)}, \dots, \vec{c}_K^{(+)}$.
 - Compute the weight of $(x_i, y_i^{(+)})$ in $D^{(+)}$ w.r.t. the j th cluster.

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

- Objective function

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

Multi-Wahba
Projection (MWP)

$$\text{s. t. } \mathbf{M}_j^{(+)^T} \mathbf{M}_j^{(+)} = \mathbf{I}, \sum_{i=1}^{|D^{(+)}|} a_{i,j}^{(+)} = 1, j = 1, \dots, K$$

Monolingual Model (3)

- Solving the MWP Problem

- Consider the j th cluster only:

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

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

A high-dimensional
Wahba's problem!

- An SVD-based closed-form solution:

$$(1) \quad \mathbf{B}_j = \sum_{i=1}^{|D^{(+)}|} a_{i,j}^{(+)} \vec{y}_i^{(+)} \vec{x}_i^T;$$

Refer to the paper for
the proof of correctness.

$$(2) \quad SVD(\mathbf{B}_j) = \mathbf{U}_j \Sigma_j \mathbf{V}_j^T;$$

$$(3) \quad \mathbf{R}_j = diag(\underbrace{1, \dots, 1}_{d-1}, \det(\mathbf{U}_j) \det(\mathbf{V}_j));$$

$$(4) \quad \mathbf{M}_j^{(+)} = \mathbf{U}_j \mathbf{R}_j \mathbf{V}_j^T;$$

Monolingual Model (4)

- Overall Procedure
 - Learning hypernymy projections

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

$$\text{s. t. } \mathbf{M}_j^{(+)^T} \mathbf{M}_j^{(+)} = \mathbf{I}, \sum_{i=1}^{|D^{(+)}|} a_{i,j}^{(+)} = 1, j = 1, \dots, K$$

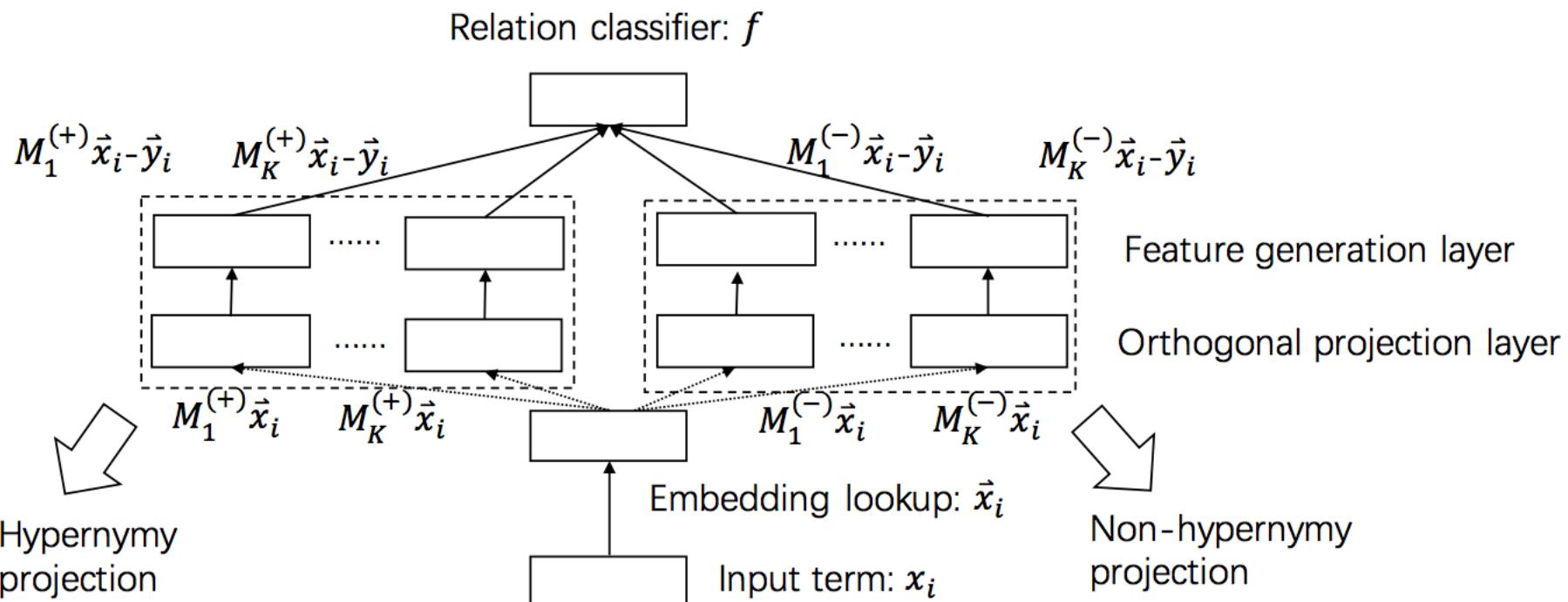
- Learning non-hypernymy projections

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

$$\text{s. t. } \mathbf{M}_j^{(-)^T} \mathbf{M}_j^{(-)} = \mathbf{I}, \sum_{i=1}^{|D^{(-)}|} a_{i,j}^{(-)} = 1, j = 1, \dots, K$$

Monolingual Model (5)

- Overall Procedure
 - Training the projection-based neural network



Cross-lingual Models (1)

- Basic Notations
 - Hypernymy training sets
 - Source language: $D_S^{(+)}$ $|D_S^{(+)}| \gg |D_T^{(+)}|$
 - Target language: $D_T^{(+)}$
 - Non-hypernymy training sets
 - Source language: $D_S^{(-)}$ $|D_S^{(-)}| \gg |D_T^{(-)}|$
 - 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

$$\min \tilde{J}(\mathcal{M}^{(+)}) = \frac{\beta}{2} \sum_{j=1}^K \sum_{i=1}^{|D_S^{(+)}|} a_{i,j}^{(+)} \gamma_i^{(+)} \|\mathbf{M}_j^{(+)} \mathbf{s}\vec{x}_i - \mathbf{s}\vec{y}_i^{(+)}\|^2 + \frac{1-\beta}{2} \sum_{j=1}^K \sum_{i=1}^{|D_T^{(+)}|} a_{i,j}^{(+)} \|\mathbf{M}_j^{(+)} \vec{x}_i - \vec{y}_i^{(+)}\|^2$$

$$\text{s. t. } \mathbf{M}_j^{(+)^T} \mathbf{M}_j^{(+)} = \mathbf{I}, \sum_{i=1}^{|D_S^{(+)}|} a_{i,j}^{(+)} \gamma_i^{(+)} = 1, \sum_{i=1}^{|D_T^{(+)}|} a_{i,j}^{(+)} = 1,$$
$$j = 1, \dots, K$$

S : maps the embeddings of the source language to the target language by Bilingual Lexicon Induction

- β : 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

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:

$$(1) \quad \mathbf{B}_j = \beta \sum_{i=1}^{|D_S^{(+)}|} a_{i,j}^{(+)} \gamma_i^{(+)} \mathbf{S} \vec{y}_i^{(+)} (\mathbf{S} \vec{x}_i)^T + (1-\beta) \sum_{i=1}^{|D_T^{(+)}|} a_{i,j}^{(+)} \vec{y}_i^{(+)} \vec{x}_i^T;$$

$$(2) \quad SVD(\mathbf{B}_j) = \mathbf{U}_j \Sigma_j \mathbf{V}_j^T;$$

$$(3) \quad \mathbf{R}_j = diag(\underbrace{1, \dots, 1}_{d-1}, \det(\mathbf{U}_j) \det(\mathbf{V}_j));$$

$$(4) \quad \mathbf{M}_j^{(+)} = \mathbf{U}_j \mathbf{R}_j \mathbf{V}_j^T;$$

Cross-lingual Models (4)

- Transfer MWP Model (TMWP)

- Learning non-hypernymy projections

$$\begin{aligned} \min \tilde{J}(\mathcal{M}^{(-)}) = & \frac{\beta}{2} \sum_{j=1}^K \sum_{i=1}^{|D_S^{(-)}|} a_{i,j}^{(-)} \gamma_i^{(-)} \|\mathbf{M}_j^{(-)} \mathbf{s}\vec{x}_i - \mathbf{s}\vec{y}_i^{(-)}\|^2 \\ & + \frac{1-\beta}{2} \sum_{j=1}^K \sum_{i=1}^{|D_T^{(-)}|} a_{i,j}^{(-)} \|\mathbf{M}_j^{(-)} \vec{x}_i - \vec{y}_i^{(-)}\|^2 \\ \text{s. t. } & \mathbf{M}_j^{(-)T} \mathbf{M}_j^{(-)} = \mathbf{I}, \quad \sum_{i=1}^{|D_S^{(-)}|} a_{i,j}^{(-)} \gamma_i^{(-)} = 1, \quad \sum_{i=1}^{|D_T^{(-)}|} a_{i,j}^{(-)} = 1, \\ & j = 1, \dots, K \end{aligned}$$

- 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)

Method	BLESS	ENTAILMENT
Mikolov et al. [24]	0.84	0.83
Yu et al. [54]	0.90	0.87
Luu et al. [20]	0.93	0.91
Nguyen et al. [26]	0.94	0.91
MWP (Non-orthogonal)	0.95	0.90
MWP	0.97	0.92

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.

Method	ANIMAL	PLANT	VEHICLE
Mikolov et al. [24]	0.80	0.81	0.82
Yu et al. [54]	0.67	0.65	0.70
Luu et al. [20]	0.89	0.92	0.89
Nguyen et al. [26]*	0.83	0.91	0.83
MWP (Non-orthogonal)	0.90	0.92	0.87
MWP	0.92	0.94	0.90

Monolingual Experiments (3)

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

Hypernymy vs. Reverse- hypernymy	Measure	BLESS	WBLESS	Hypernymy vs. Other relations
Santus et al. [31]	0.87	-		
Weeds et al. [49]	-	0.75		
Kiela et al. [15]	0.88	0.75		
Nguyen et al. [26]	0.92	0.87		
Roller et al. [30]	0.96	0.87		
MWP (Non-orthogonal)	0.95	0.89		
MWP	0.97	0.92		

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

French Chinese Japanese Italian Thai Finnish Greek

Relation ↓ Language →	fr	zh	ja	it	th	fi	el
# Hypernymy relations	4,035	2,962	1,448	3,034	1,156	7,157	2,612
# Non-hypernymy relations	8,947	6,382	3,203	6,081	1,977	9,433	1,454

Cross-lingual Experiments (2)

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

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

Cross-lingual Experiments (3)

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

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

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