Learning Sense-specific Word Embeddings and Applying for Semantic Hierarchies

Wanxiang Che

HIT-SCIR

Joint work with Jiang Guo and Ruiji Fu

2014-5-16

Word Embeddings

- Definition
  - Compact, Real-valued, Low-dimensional vector representations
  - Example
    - $Emb(\text{计算机}) = [0.12,\ 0.26,\ ...,\ 0.09]$
    - $Emb(\text{电　脑}) = [0.14,\ 0.14,\ ...,\ 0.10]$

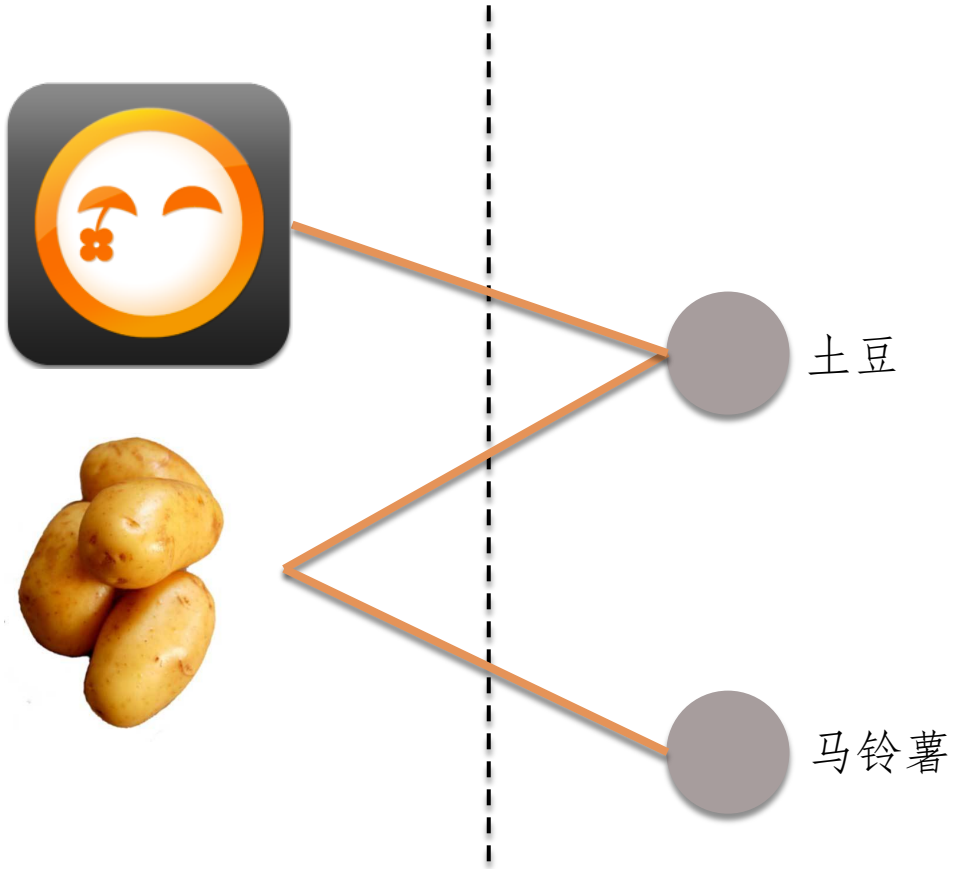
$e.g.\ 50 dim$

- Learning Architectures
  - Context-Predicting Models
    - Neural Network Language Models (NNLM)
    - Skip-Gram Model
  - Context-Counting Models
    - LSA, CCA, etc.

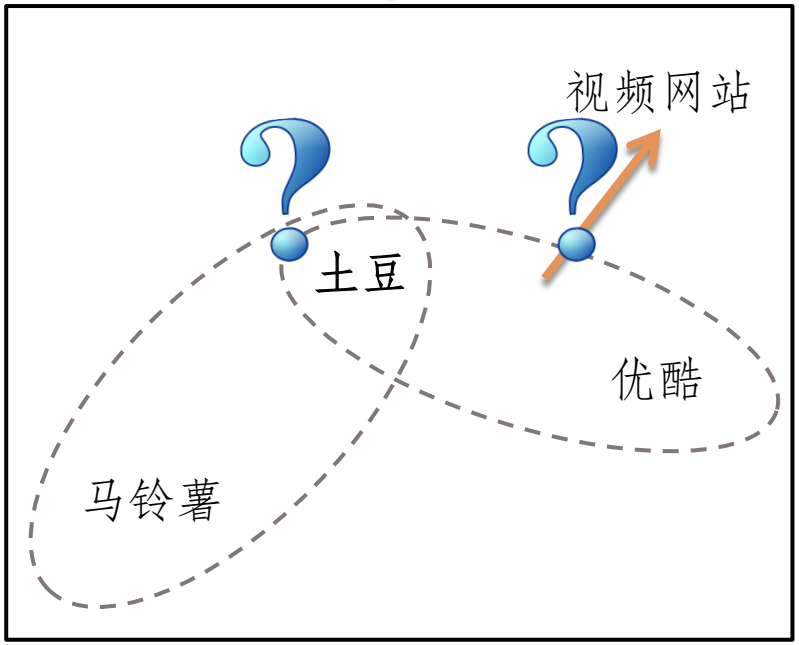
Word and Sense Mapping

Sense space

Word space



Word Embedding



Learning Sense-specific Word Embedding by Exploiting Bilingual Resources

Approach

- Represent words with sense-specific embeddings
  - Word sense induction using a bilingual approach
  - Train embeddings on the sense-tagged corpus

- Applications
  - Word similarity evaluation of polysemous words
  - Incorporating sense-specific embeddings to sequence labeling tasks (e.g. Named Entity Recognition)

Foreign Language

Sense space

Source Language

Subdue

Subjugate

Overpower

Uniform

制服

降服

工作服

工装

Foreign Language

Sense space

Source Language

Subdue

Subjugate

Overpower

Uniform

制服

降服

工作服

工装

Method --- Learning

- Word Alignment
- Extract the translation words via bidirectional translation probability
- Cluster the translation words using their embeddings
  - Affinity Propagation (AP) clustering is used, for automatically determine the cluster number
- Cross-lingual Sense Projection
  - Tag the words in source language with its sense cluster
- Train embeddings using RNNLM (recurrent NNLM)

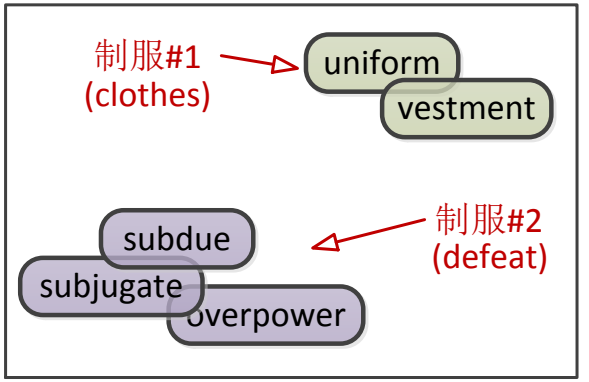
Case

Bilingual data (E: English, C: Chinese)	
1	E: The criminal is subdued at last C: 罪犯 终 被 制服
2	E: The policeman wearing uniform C: 身穿 制服 的 警察
3	E: She overpowered the burglars C: 她 制服 了 窃贼
4	E: They wore uniforms made in China C: 他们 身穿 中国 生产 的 制服
5	……

① Extract

SL word	制服
Translations	subdue uniform overpower subjugate vestment ……

② Cluster



③ Project

Monolingual sense-labeled data	
1	罪犯 终 被 制服 #2
2	身穿 制服 #1 的 警察
3	她 制服 #2 了 窃贼
4	他们 身穿 中国 生产 的 制服 #1
5	……

④ RNNLM

Sense-specific word embeddings	
制服 #1	$< v_1^{\#1}, v_2^{\#1}, ..., v_N^{\#1} >$
制服 #2	$< v_1^{\#2}, v_2^{\#2}, ..., v_N^{\#2} >$

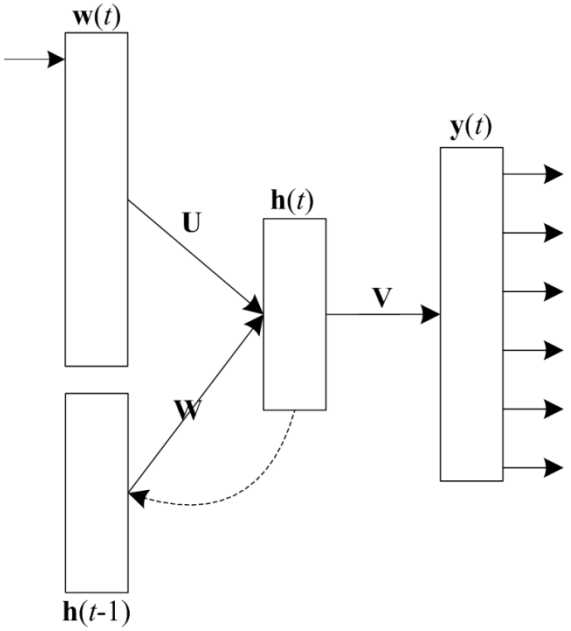
Recurrent Neural Network Language Model

- Modeling the conditional probability
  - $P(\mathrm{w(t+1)}|\mathrm{w}(t), \mathrm{h(t-1)})$
  - Compute:

$$\mathbf{h}(t) = f(\mathbf{U}\mathbf{w}(t) + \mathbf{W}\mathbf{h}(t-1))$$

$$\mathbf{y}(t) = g(\mathbf{V}\mathbf{h}(t))$$

  - U is the Embedding Matrix



Related Work

- Huang et al. (2012), Reisinger and Mooney (2010)
  - Learning Multiple-prototype Word Embeddings
    - K embeddings for each word
  - Problem
    - Ignoring the fact that the number of senses of different words are varied.

Word Similarity Evaluation

- A manually constructed Polysemous Word Similarity Dataset
- Measurement
  - Spearman Correlation
  - Kendall Correlation
- Quantitative Evaluation

System	MaxSim		AvgSim	
	$\rho \times 100$	$\tau \times 100$	$\rho \times 100$	$\tau \times 100$
Ours	**55.4**	**40.9**	49.3	35.2
SingleEmb	42.8	30.6	42.8	30.6
Multi-prototype	40.7	29.1	38.3	27.4

Word Similarity Evaluation

- Qualitative Evaluation: K-nearest neighbors

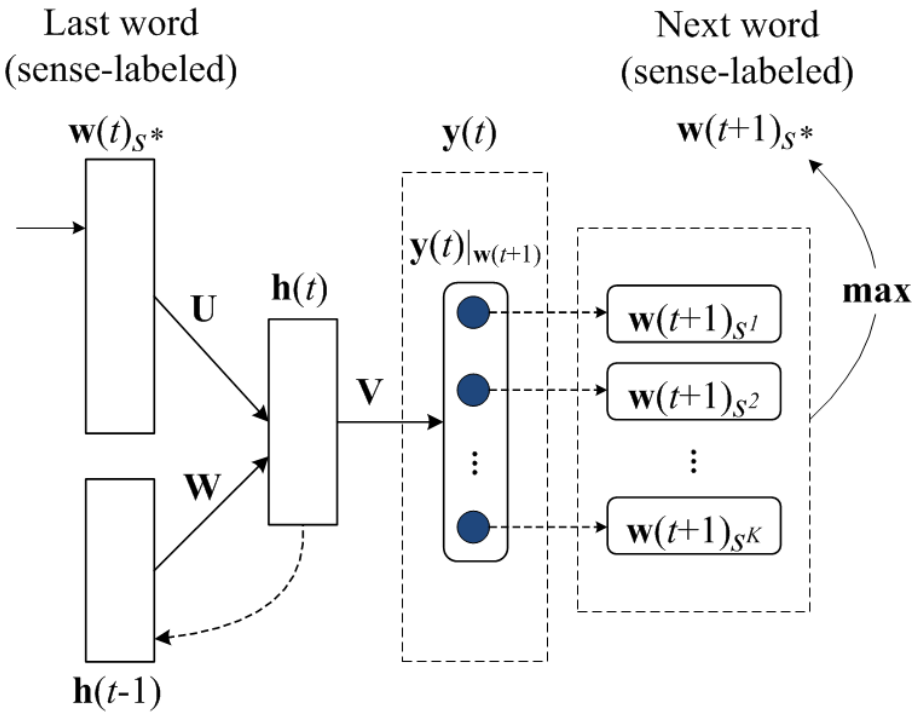
Word	Nearest Neighboors
制服$_{uniform}$	穿着$_{dress}$, 警服$_{policeman\ uniform}$
制服$_{subdue}$	打败$_{defeat}$, 击败$_{beat}$, 征服$_{conquer}$
花$_{spend}$	花费$_{cost}$, 节省$_{save}$, 剩下$_{rest}$
花$_{flower}$	菜$_{greens}$, 叶$_{leaf}$, 果实$_{fruit}$
法$_{law}$	法令$_{ordinance}$, 法案$_{bill}$, 法规$_{rule}$
法$_{method}$	概念$_{concept}$, 方案$_{scheme}$
法$_{French}$	德$_{Germany}$, 俄$_{Russia}$, 英$_{Britain}$
领导$_{lead}$	监督$_{supervise}$, 决策$_{decision}$
领导$_{leader}$	主管$_{chief}$, 上司$_{boss}$

Method --- Application

- Semi-supervised NER
  - Feed the word embeddings as additional features to a supervised learning model
  - Problem: discriminate the sense of words in NER data
  - Solution:
    - A novel Word Sense Disambiguation algorithm based on the RNNLM trained previously

Word Sense Disambiguation based on RNNLM

- Single-step decision
  - $argmax\, P(w(t+1)\downarrow s\uparrow *\ |w(t)\downarrow s\uparrow *\ , h(t-1))$



- Greedy decoding
- Beam-search decoding

NER Evaluation

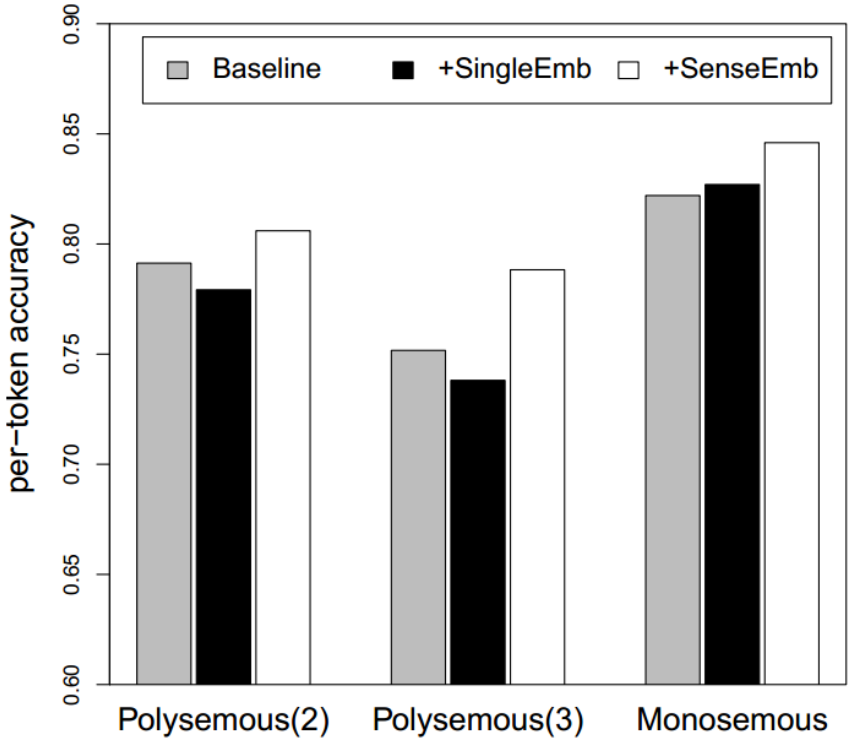
- Model: Conditional Random Fields

- Result:

System	P	R	F
Baseline	93.27	81.46	86.97
+SingleEmb	93.55	82.32	87.58
+SenseEmb (greedy)	93.38	83.56	88.20
+SenseEmb (beam search)	**93.59**	**84.05**	**88.56**

NER Evaluation

- Analysis
  - Per-token accuracy of polysemous words and monosemous words, respectively

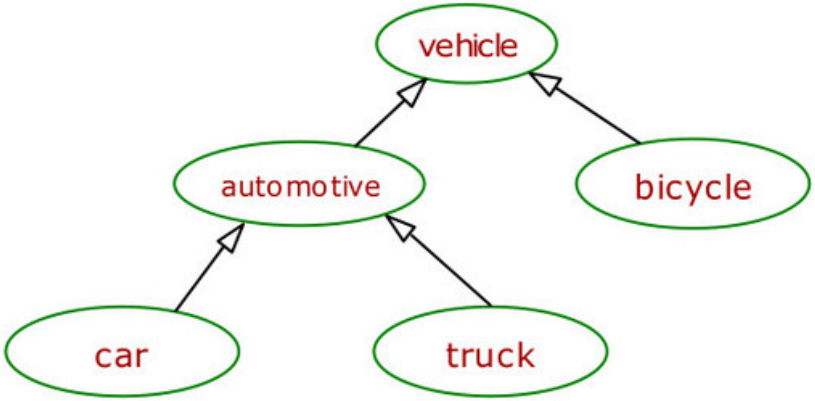


Learning Semantic Hierarchies via Word Embeddings

Semantic Hierarchies

Learning Semantic Hierarchies via Word Embeddings

- car → automotive
  - hypernym: automotive
  - hyponym: car



- manually-built semantic hierarchies
  - WordNet
  - HowNet
  - CilinE (Tongyi Cilin - Extended version)

Previous Work

- Pattern-based method
  - e.g. “such NP1 as NP2”
    - Hearst (1992)； Snow et al. (2005)

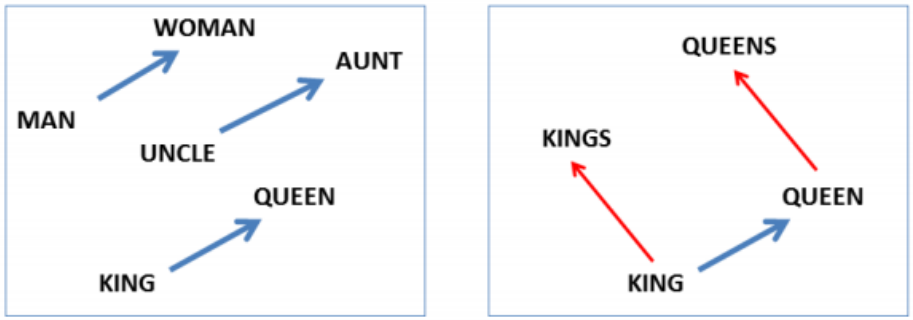
Pattern	Translation
w 是[一个|一种] h	w is a [a kind of] h
w [、] 等 h	w[,] and other h
h [，] 叫[做] w	h[,] called w
h [，] [像]如 w	h[,] such as w
h [，] 特别是 w	h[,] especially w

- Methods based on web mining
  - assuming that the hypernyms of an entity co-occur with it frequently
  - extracting hypernym candidates from multiple sources and learning to rank
    - Fu et al. (2013)

Word Embeddings

- Learning Semantic Hierarchies via Word Embeddings

$$v(\texttt{king}) - v(\texttt{queen}) \approx v(\texttt{man}) - v(\texttt{woman})$$



Mikolov et al. (2013a)

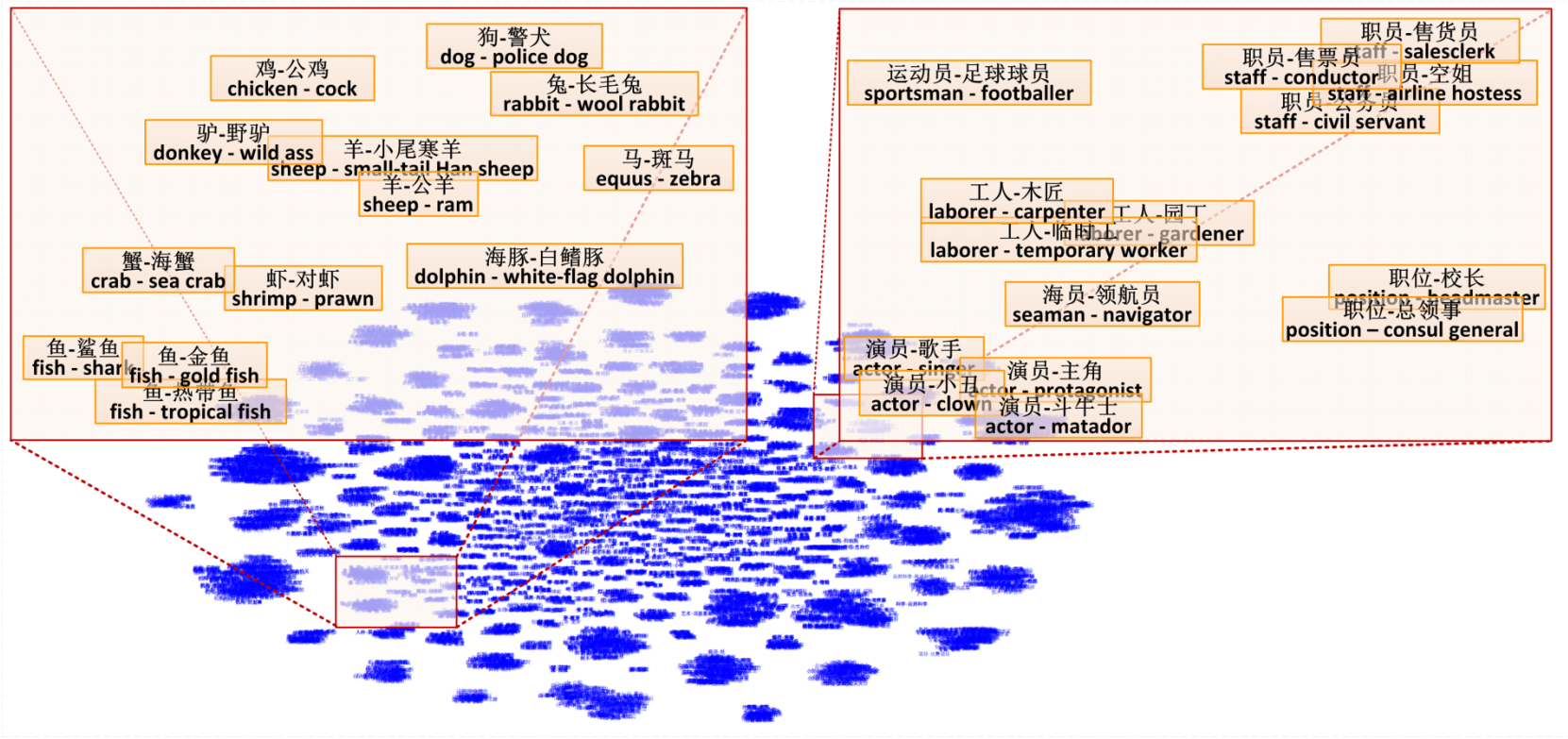
Motivation

- Does the embedding offset work well in hypernym–hyponym relations?

No.	Examples
1	$v(\text{虾}) - v(\text{对虾}) \approx v(\text{鱼}) - v(\text{金鱼})$ $v(\text{shrimp}) - v(\text{prawn}) \approx v(\text{fish}) - v(\text{gold fish})$
2	$v(\text{工人}) - v(\text{木匠}) \approx v(\text{演员}) - v(\text{小丑})$ $v(\text{laborer}) - v(\text{carpenter}) \approx v(\text{actor}) - v(\text{clown})$
3	$v(\text{工人}) - v(\text{木匠}) \not\approx v(\text{鱼}) - v(\text{金鱼})$ $v(\text{laborer}) - v(\text{carpenter}) \not\approx v(\text{fish}) - v(\text{gold fish})$

Motivation

- Clusters of the vector offsets in training data

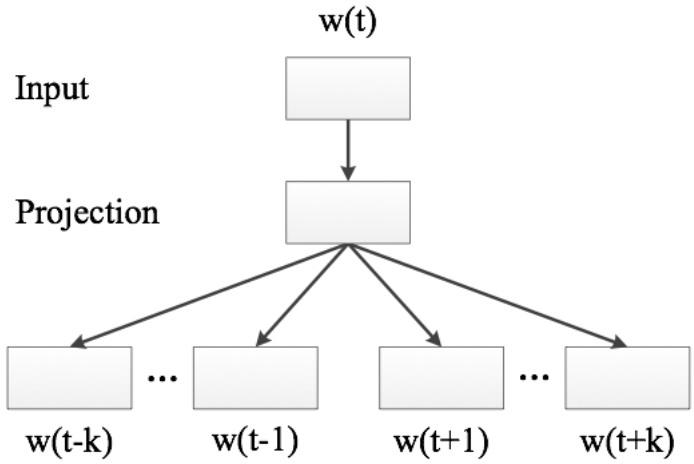


Method

- Word Embedding Training
- Projection Learning
- is-a Relation Identification

Word Embedding Training

- Skip-gram



Mikolov et al. (2013b)

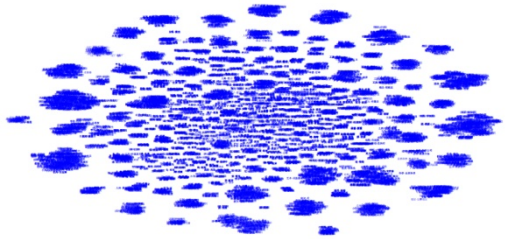
Projection Learning

- A uniform Linear Projection
  - Given a word x and its hypernym y, there exists a matrix Φ so that $y = \Phi x$.

$$\Phi^* = \arg\min_{\Phi} \frac{1}{N} \sum_{(x,y)} \| \Phi x - y \|^2$$

Projection Learning

- Piecewise Linear Projections
  - clustering y - x

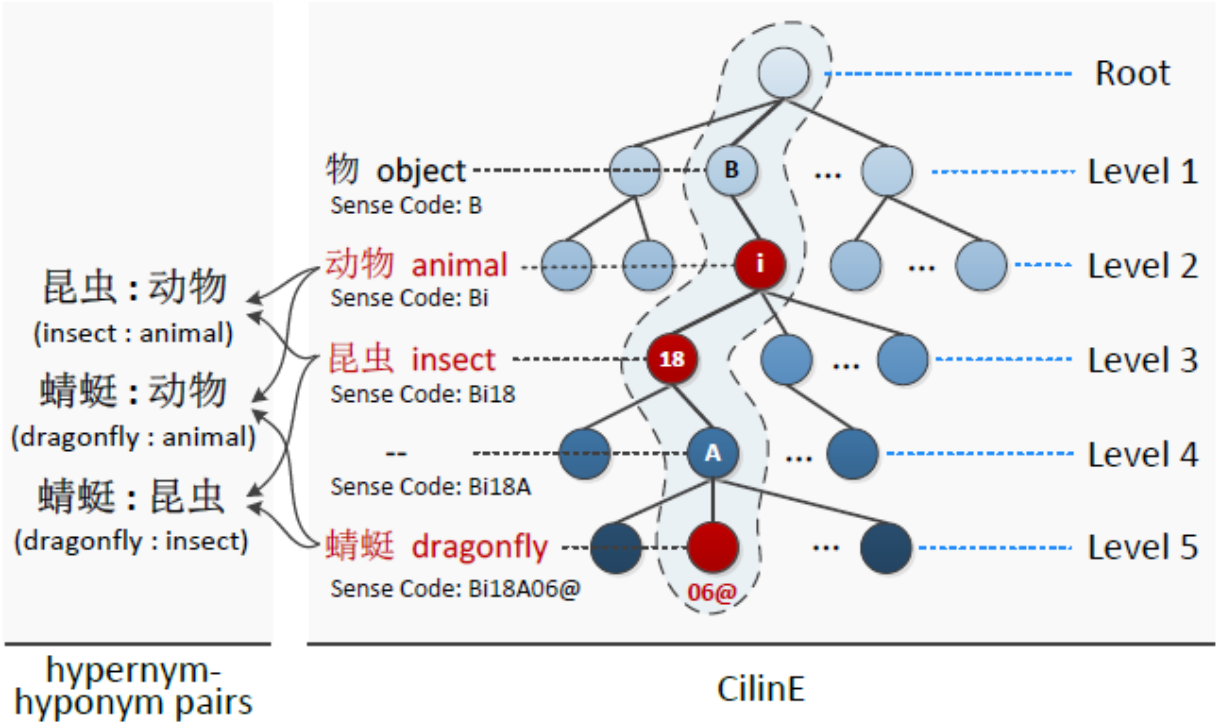


  - learning a separate projection for each cluster

$$\Phi_k^* = \arg\min_{\Phi_k} \frac{1}{N_k} \sum_{(x,y)\in C_k} \| \Phi_k x - y \|^2$$

Projection Learning

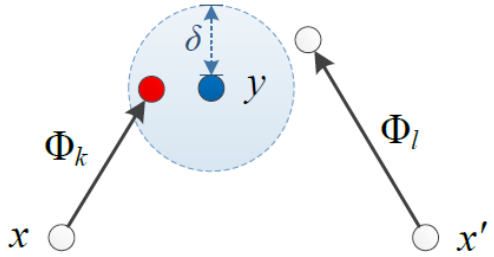
- Training data



is-a Relation Identification

- Given two words x and y
  - If y is determined as a hypernym of x, either of the two conditions must be satisfied.

Condition 1:



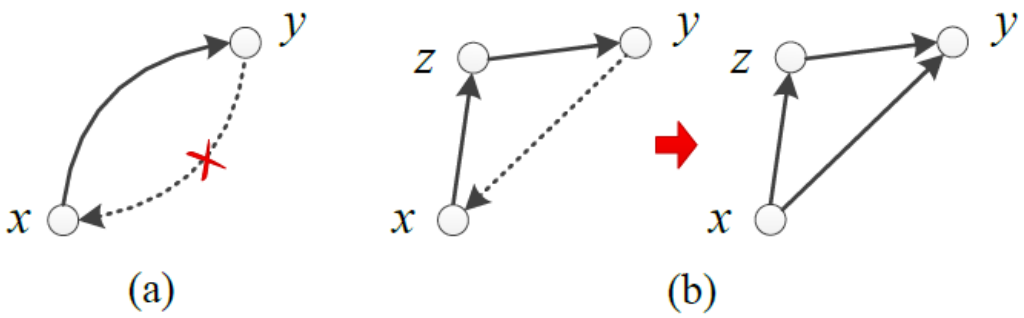
$$d(\Phi_k x, y) = \| \Phi_k x - y \|^2 < \delta$$

Condition 2:

$$x \xrightarrow{H} z \text{ and } z \xrightarrow{H} y$$

is-a Relation Identification

- Hierarchy Condition (DAG)
  - $\forall x, y \in L : x \xrightarrow{H} y \Rightarrow \neg(y \xrightarrow{H} x)$
  - $\forall x, y, z \in L : (x \xrightarrow{H} z \wedge z \xrightarrow{H} y) \Rightarrow x \xrightarrow{H} y$

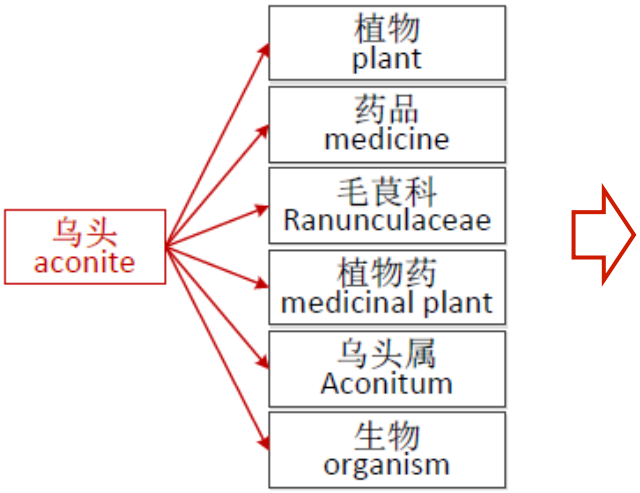


Experimental Data

- Word embedding training
  - corpus from Baidubaike
    - ~30 million sentences (~780 million words)
- Projection learning
  - CilinE
    - 15,247 is-a pairs

Experimental Data

- For evaluation



Relation	# of word pairs	
	Dev.	Test
hypernym–hyponym	312	1,079
hyponym–hypernym*	312	1,079
unrelated	1,044	3,250
Total	1,668	5,408

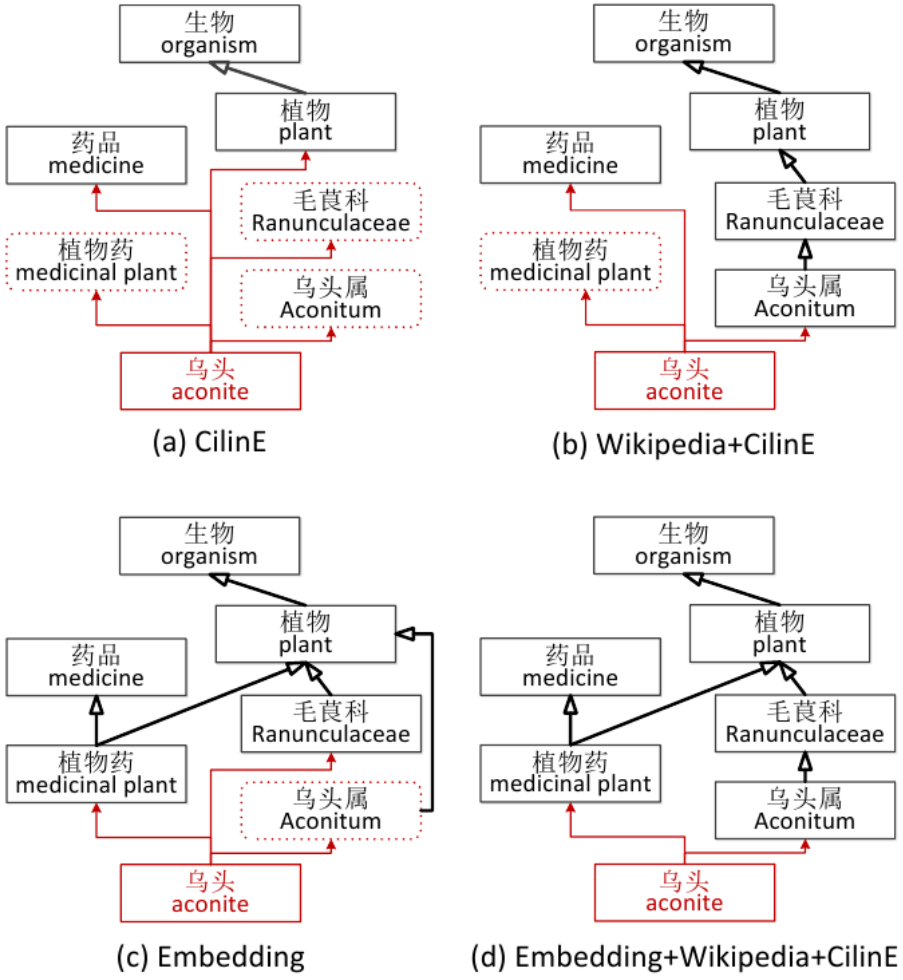
Fu et al. (2013)

Results and Analysis

- Comparison with existing methods

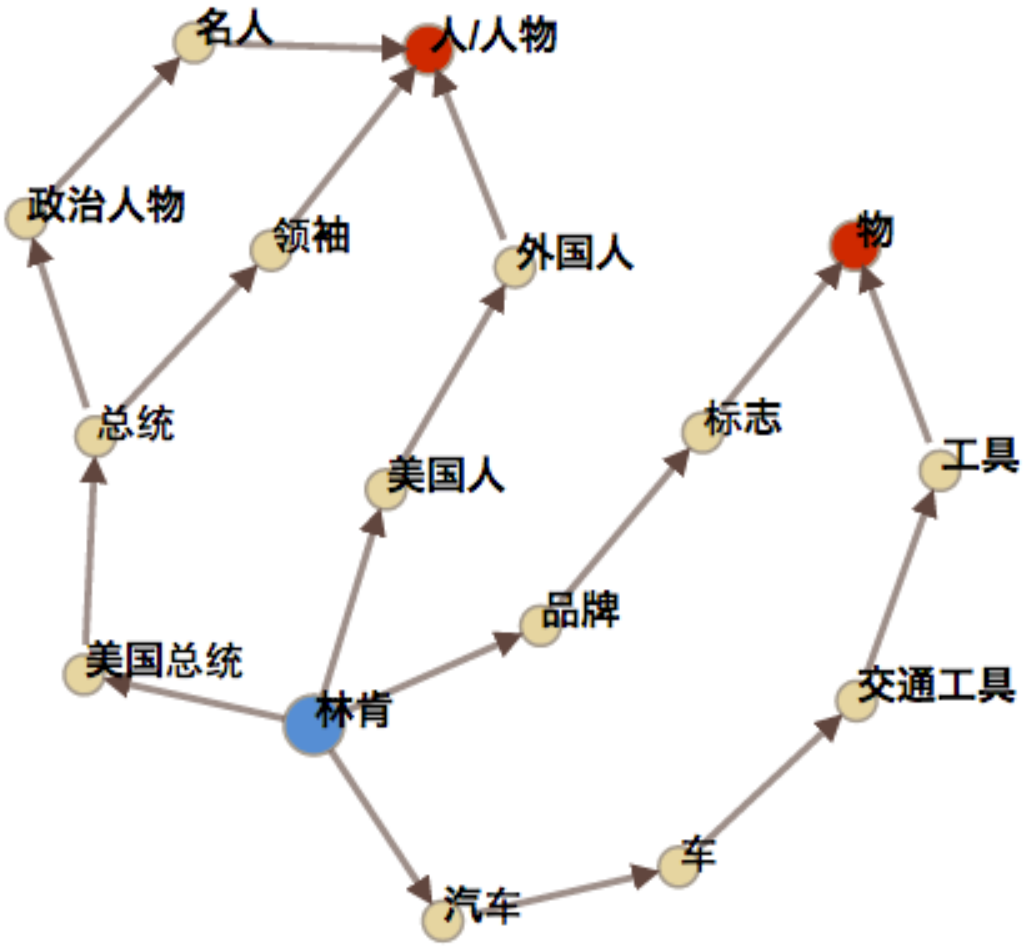
	P(%)	R(%)	F(%)	
M_{CilinE}	98.21	50.88	67.03	
$M_{Wiki+CilinE}$	92.41	60.61	73.20	Suchanek et al. (2008)
$M_{Pattern}$	97.47	21.41	35.11	Hearst (1992)
M_{Snow}	60.88	25.67	36.11	Snow et al. (2005)
$M_{balApinc}$	54.96	53.38	54.16	Kotlerman et al. (2010)
M_{invCL}	49.63	62.84	55.46	Lenci and Benotto (2012)
M_{Fu}	71.64	52.92	60.87	Fu et al. (2013)
M_{offset}	59.26	63.19	61.16	
M_{Emb}	80.54	67.99	**73.74**	
$M_{Emb+CilinE}$	80.59	72.42	76.29	
$M_{Emb+Wiki+CilinE}$	79.78	80.81	**80.29**	

Results and Analysis



Demo

- http://www.bigcilin.com



THANKS & QA