Coling 2016

Chinese Hypernym-Hyponym Extraction from User Generated Categories

Chengyu Wang, Xiaofeng He

School of Computer Science and Software Engineering,
East China Normal University
Shanghai, China

EAST CHINA NORMAL UNIVERSITY 華東師範大學

DaSE Data Science & Engineering

Outline

- **Introduction**
- Background and Related Work
- Proposed Approach
- Experiments
- Conclusion

Chinese Is-A Relation Extraction

- Chinese is-a relation extraction
  - Chinese is-a relations are essential to construct large-scale Chinese taxonomies and knowledge graphs.
  - It is difficult to extract such relations due to the flexibility of language expression.

- User generated categories
  - User generated categories are valuable knowledge sources, providing fine-grained candidate hypernyms of entities.
  - The semantic relations between an entity and its categories are not clear.

DaSE Data Science & Engineering

Baidu Baike: one of the largest online encyclopedias in China, with 13M+ entries

Barack Obama

收藏 0 261

贝拉克·奥巴马 编辑

同义词 奥巴马（美国第44任总统）一般指贝拉克·奥巴马

贝拉克·侯赛因·奥巴马（Barack Hussein Obama），1961年8月4日出生，美国民主党籍政治家，第44任美国总统，为美国历史上第一位非洲裔总统。1991年，奥巴马以优等生荣誉从哈佛法学院毕业，而后在著名的芝加哥大学法学院教授宪法长达12年（1992年-2004年）。2007年2月10日，宣布参加2008年美国总统选举。2008年11月4日正式当选为美国总统。

2009年10月9日，获得诺贝尔委员会颁发的诺贝尔和平奖 [1] 。2012年11月6日，第57届美国总统大选中，奥巴马击败共和党候选人罗姆尼，成功连任。

贝拉克·侯赛因·奥巴马于2014年11月10日至12日来华出席亚太经合组织领导人非正式会议并对中国进行国事访问。 [2] 2014年12月，奥巴马参加了由非盈利组织Code.org举办的编程大会。会上，奥巴马熟练地习得一小段JavaScript代码，并成功地画出了一个正方形。使得他成为了美国史上首位会编程的总统。

2015年3月11日，贝拉克·侯赛因·奥巴马在各国领导人工资中，排名第一位。 [3] 2015年5月，奥巴马基金会确认“奥巴马总统图书馆（Obama Presidential Library）”将落户于他曾经长期执教的芝加哥大学 [4-5] 。2015年10月，《彭博市场》公布了第五届全球金融50大最具影响力人物，美国总统奥巴马排名第六。 [6] 2015年11月4日，奥巴马名列《福布斯》全球最有权力人物排行榜第三位。 [7] 2015年12月22日，国际民调机构盖洛普调查称，奥巴马在最受欢迎的领导人排名中名列第一。 [8]

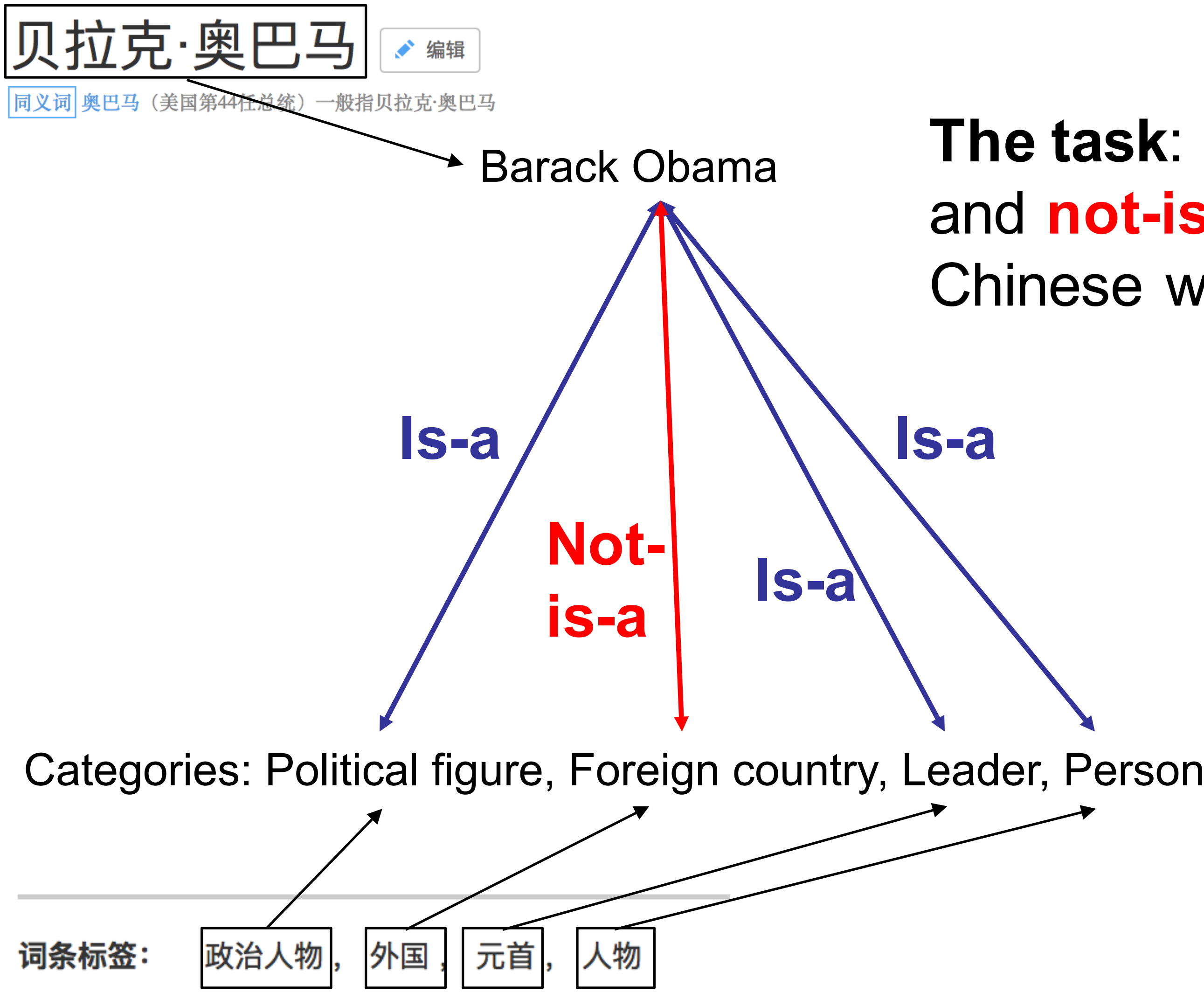
贝拉克·奥巴马图册

词条标签： 政治人物，外国，元首，人物

Categories: Political figure, Foreign country, Leader, Person

DaSE Data Science & Engineering

The task: distinguishing **is-a** and **not-is-a** relations between Chinese words/phases



Outline

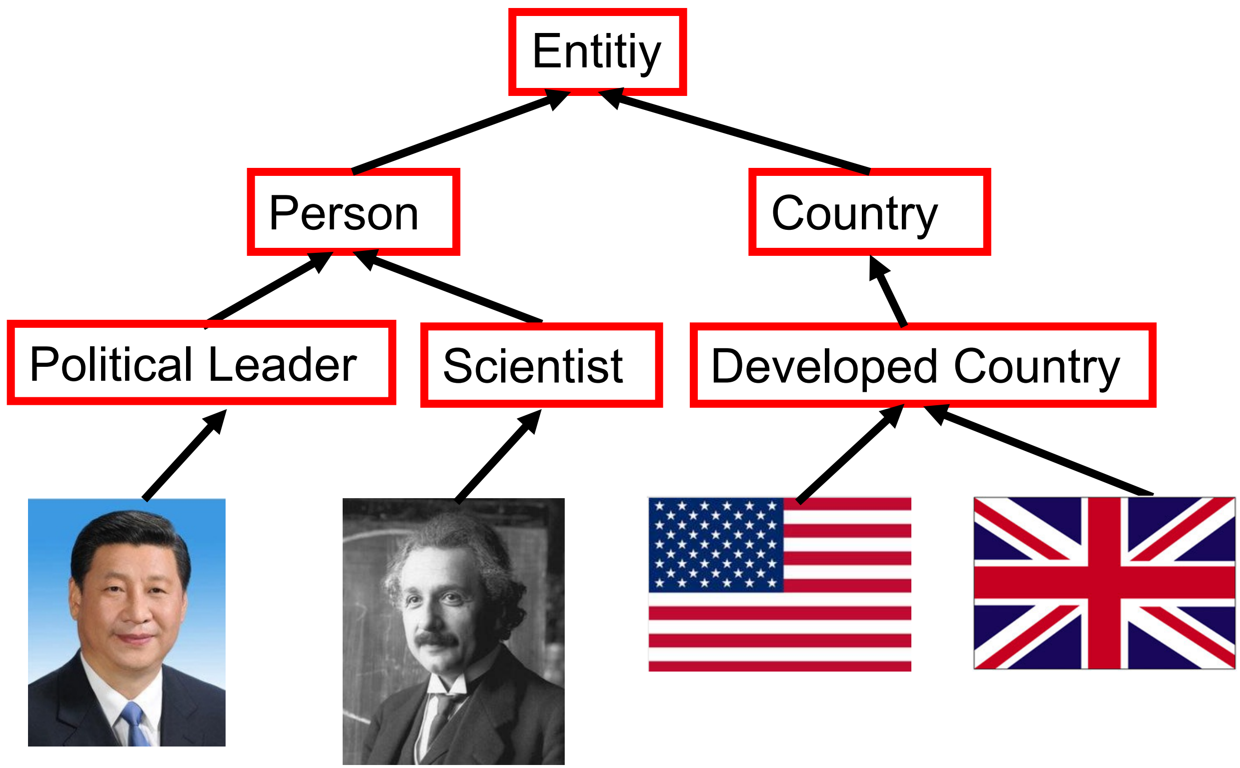
- Introduction
- **Background and Related Work**
- Proposed Approach
- Experiments
- Conclusion

Background

- Taxonomy: a hierarchical type system for knowledge graphs, consisting of is-a relations among classes and entities
  - Example

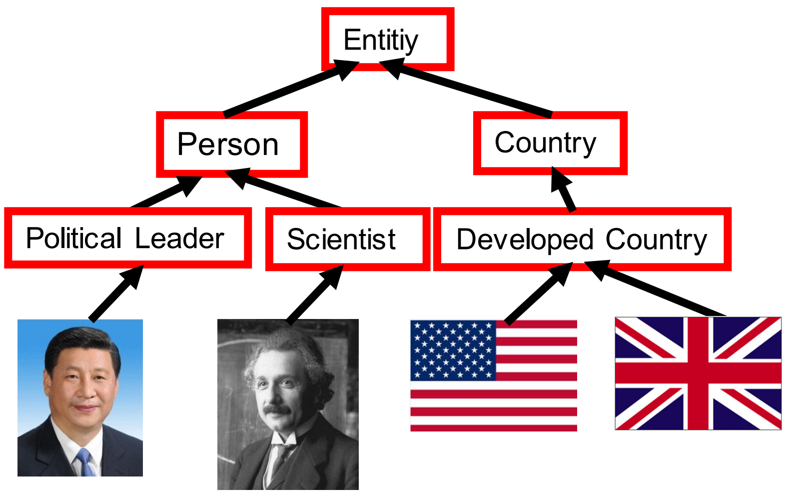
Classes

Entities



Describing the Task

- Learning *is-a* relations for taxonomy expansion



收藏 0 261

贝拉克·奥巴马 编辑

同义词 奥巴马（美国第44任总统）一般指贝拉克·奥巴马

贝拉克·侯赛因·奥巴马（Barack Hussein Obama），1961年8月4日出生，美国民主党籍政治家，第44任美国总统，为美国历史上第一位非洲裔总统。1991年，奥巴马以优等生荣誉从哈佛法学院毕业，而后在著名的芝加哥大学法学院教授宪法长达12年（1992年-2004年）。2007年2月10日，宣布参加2008年美国总统选举。2008年11月4日正式当选为美国总统。

2009年10月9日，获得诺贝尔委员会颁发的诺贝尔和平奖 [1] 。2012年11月6日，第57届美国总统大选中，奥巴马击败共和党候选人罗姆尼，成功连任。

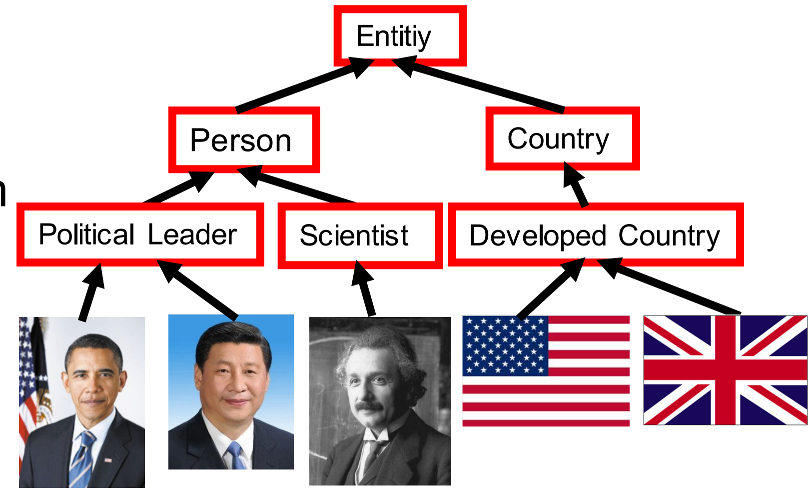
贝拉克·侯赛因·奥巴马于2014年11月10日至12日来华出席亚太经合组织领导人非正式会议并对中国进行国事访问。 [2] 2014年12月，奥巴马参加了由非盈利组织Code.org举办的编程大会。会上，奥巴马熟练地习得一小段JavaScript代码，并成功地画出了一个正方形。使得他成为了美国史上首位会编程的总统。

2015年3月11日，贝拉克·侯赛因·奥巴马在各国领导人工资中，排名第一位。 [3] 2015年5月，奥巴马基金会确认"奥巴马总统图书馆（Obama Presidential Library）"将落户于他曾经长期执教的芝加哥大学 [4-5] 。2015年10月，《彭博市场》公布了第五届全球金融50大最具影响力人物，美国总统奥巴马排名第六。 [6] 2015年11月4日，奥巴马名列《福布斯》全球最有权力人物排行榜第三位。 [7] 2015年12月22日，国际民调机构盖洛普调查称，奥巴马在最受欢迎的领导人排名中名列第一。 [8]

贝拉克·奥巴马图册

词条标签： 政治人物， 外国， 元首， 人物

Learning Algorithm



Key challenge: identify *is-a* relations from user generated categories

DaSE Data Science & Engineering

Modeling the Task

- Taxonomy
  - Direct acyclic graph $G = (E, R)$ (E: entities/classes, R: is-a relations)
- User generated categories
  - Collection of entities E^*
  - Set of user generated categories: $Cat(e)$ for $e \in E^*$
- Goal
  - Predict whether there is an is-a relation between e and c where $e \in E^*$ and $c \in Cat(e)$ based on the taxonomy G

DaSE Data Science & Engineering

Previous Approaches

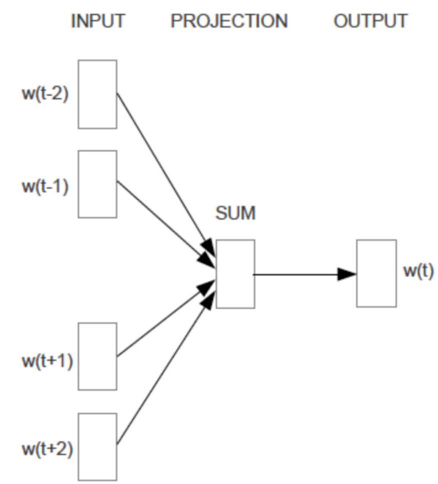
- Pattern matching-based approaches
  - Handcraft patterns: high accuracy, low coverage
    - Hearst Patterns: NP_1 *such as* NP_2
  - Automatic generated patterns: higher coverage, lower accuracy
  - Not suitable for Chinese with flexible expression
- Thesauri and encyclopedia based approaches
  - Taxonomy construction based on existing knowledge sources
    - YAGO: Wikipedia + WordNet
    - More precise but have limited scope constrained by sources
  - Chinese: relatively low-resourced
    - No Chinese version of WordNet and Freebase available

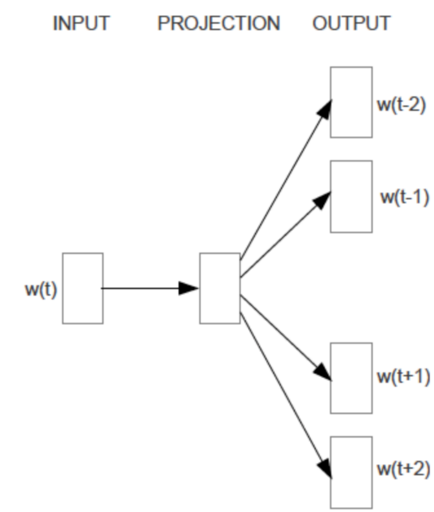
yago select knowledge

DaSE Data Science & Engineering

Previous Approaches

- Text inference based approach
  - Infer relations using distributed similarity measures
    - Assumption: a hyponym can only appear in some of the contexts of its hypernym and a hypernym can appear in all contexts of its hyponyms
  - Not suitable for Chinese with flexible and sparse contexts
- Word embedding based approach
  - Represent words as dense, low-dimensional vectors
  - Learn semantic projection models from hyponyms to hypernyms
  - State-of-the-art approach for Chinese is-a relation extraction (ACL'14)



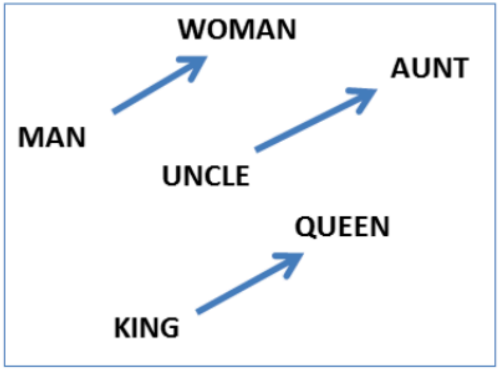


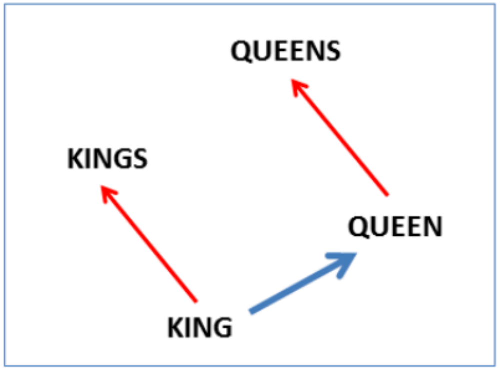
Figures taken from Mikolov et al., 2013

DaSE Data Science & Engineering

Learning from Previous Work

- Lessons learned from “state-of-the art”
  - Use word embeddings to represent words
  - Learn relations between hyponyms and hypernyms in the embedding space
- Basic approaches
  - Vector offsets
  - Linear projection





Figures taken from Mikolov et al., 2013

Observations

- Word vector offsets between Chinese is-a pairs
  - Multiple linguistic regularities may exist in is-a pairs
    - Different levels of hypernyms
    - Different types of is-a relations (instanceOf vs. subClassOf)
    - Different domains

	Example with English Translation	Vector Offsets
True Positive	$\vec{v}$(日本) $-$ $\vec{v}$(国家) $\approx$ $\vec{v}$(澳大利亚) $-$ $\vec{v}$(国家) $\vec{v}$(Japan) $-$ $\vec{v}$(Country) $\approx$ $\vec{v}$(Australia) $-$ $\vec{v}$(Country)	$1.03 \approx 0.99$
Observation 1	$\vec{v}$(日本) $-$ $\vec{v}$(国家) $\not\approx$ $\vec{v}$(日本) $-$ $\vec{v}$(亚洲国家) $\vec{v}$(Japan) $-$ $\vec{v}$(Country) $\not\approx$ $\vec{v}$(Japan) $-$ $\vec{v}$(Asian Country)	$1.03 \not\approx 0.71$
Observation 2	$\vec{v}$(日本) $-$ $\vec{v}$(国家) $\not\approx$ $\vec{v}$(主权国) $-$ $\vec{v}$(国家) $\vec{v}$(Japan) $-$ $\vec{v}$(Country) $\not\approx$ $\vec{v}$(Sovereign State) $-$ $\vec{v}$(Country)	$1.03 \not\approx 1.32$
Observation 3	$\vec{v}$(日本) $-$ $\vec{v}$(国家) $\not\approx$ $\vec{v}$(西瓜) $-$ $\vec{v}$(水果) $\vec{v}$(Japan) $-$ $\vec{v}$(Country) $\not\approx$ $\vec{v}$(Watermelon) $-$ $\vec{v}$(Fruit)	$1.03 \not\approx 0.39$

DaSE Data Science & Engineering

Outline

- Introduction
- Background and Related Work
- **Proposed Approach**
- Experiments
- Conclusion

DaSE Data Science & Engineering

General Framework

- Initial stage
  - Train piecewise linear projection models based on the Chinese taxonomy
- Iterative learning stage
  - Extract new is-a relations and adjust model parameters based on an incremental learning approach
  - Use Chinese Hypernym/Hyponym patterns to prevent "semantic drift" in each iteration

DaSE Data Science & Engineering

Initial Model Training

- Linear projection model
  - Projection model: $M\vec{v}(x_i) + \vec{b} = \vec{v}(y_i)$

    Projection matrix   Word vector   Offset vector

- Piecewise linear projection model
  - Partition a collection of is-a relations $R' \subset R^*$ into K clusters $(C_1, \cdots, C_k, \cdots, C_K)$
  - Each cluster C_k share projection matrix M_k and offset vector $\vec{b}_k$
  - Optimization function:

$$J(M_k, \vec{b}_k; C_k) = \frac{1}{|C_k|} \sum_{(x_i, y_i) \in C_k} \left\| M_k \vec{v}(x_i) + \vec{b}_k - \vec{v}(y_i) \right\|^2$$

Iterative Learning (1)

- Initialization
  - Word pairs: positive is-a set R', unlabeled set U
  - Model parameters: M_k and $\vec{b}_k$ for each cluster
- Iterative process ($t = 1, \cdots, T$)
  1. Sample $\delta|U|$ word pairs from U, denoted as $U^{(t)}$.
  2. Use the model to predict the relation between words. Denote "positive" word pairs as $U_{-}^{(t)}$.
  3. Use pattern-based relation selection method to select a subset of $U_{-}^{(t)}$ which have high confidence, denoted as $U_{+}^{(t)}$.
  4. Remove $U_{+}^{(t)}$ from U and add it to R'.

Iterative Learning (2)

- Iterative process ($t = 1, \cdots, T$)

5. Update cluster centroids incrementally based on $U_+^{(t)}$.

$$\vec{c}_k^{(t+1)} = \vec{c}_k^{(t)} + \lambda \cdot \frac{1}{\left|U_k^{(t)}\right|} \sum_{(x_i, y_i) \in U_k^{(t)}} \vec{v}(x_i) - \vec{v}(y_i) - \vec{c}_k^{(t)}$$

New centroid

Distance from centroid

Old centroid

Learning rate of centroid shift

6. Update model parameters based on new cluster assignments.

$$J\left(M_k^{(t)}, \vec{b}_k^{(t)};\ C_k^{(t)}\right) = \frac{1}{\left|C_k^{(t)}\right|} \sum_{(x_i, y_i) \in C_k^{(t)}} \left\| M_k^{(t)} \vec{v}(x_i) + \vec{b}_k^{(t)} - \vec{v}(y_i) \right\|^2$$

Iterative Learning (3)

- Model prediction
  - The prediction of the final piecewise linear projection models
  - The transitivity closure of existing is-a relations
- Discussion
  - Combination of semantic and lexical extraction of is-a relations
    - Sematic level: word embedding based projection models
    - Lexical level: pattern-based relation selection
  - Incremental learning
    - Update of cluster centroids
    - Update of model parameters

DaSE Data Science & Engineering

Pattern-based Relation Selection (1)

- Two observations
  - **Positive** evidence
    - Is-A patterns
    - Such-As patterns (between x_i/x_j and y)

    Hypothesis: x_i/x_j **is-a** y
  - **Negative** evidence
    - Such-As patterns (between x_i and x_j)
    - Co-Hyponym patterns

    Hypothesis: x_i **not-is-a** x_j x_j **not-is-a** x_i

Examples of Chinese Hypernym/Hyponym Patterns

Category	Example
Is-A	x_i是一个y x_i is a kind of y
Such-As	y，例如x_i、x_j y , such as x_i and x_j
Co-Hyponym	x_i、x_j等 x_i, x_j and others

Pattern-based Relation Selection (2)

- Positive and negative evidence scores
  - Positive score

$$PS(x_i, y_i) = \alpha\left(1 - \frac{d^{(t)}(x_i, y_i)}{\max_{(x,y)\in U_-} d^{(t)}(x,y)}\right) + (1-\alpha)\frac{n_1(x_i, y_i) + \gamma}{\max_{(x,y)\in U_-} n_1(x,y) + \gamma}$$

Confidence of model prediction   Statistics of ”positive” patterns

  - Negative score

$$NS(x_i, y_i) = \log \frac{n_2(x_i, y_i) + \gamma}{(n_2(x_i) + \gamma)\cdot(n_2(y_i) + \gamma)}$$

- Relation selection via optimization
  - Target: select m word pairs from $U_-^{(t)}$ to generate $U_+^{(t)}$

$$\max \sum_{(x_i,y_i)\in U_+^{(t)}} PS(x_i, y_i) \ \text{s.t.} \sum_{(x_i,y_i)\in U_+^{(t)}} NS(x_i, y_i) < \theta, U_+^{(t)} \subseteq U_-^{(t)}, \left|U_+^{(t)}\right| = m$$

DaSE Data Science & Engineering

Pattern-based Relation Selection (3)

- Relation selection algorithm

Algorithm 1 Greedy Relation Selection Algorithm

Initialize $U_{+}^{(t)} = \emptyset$;
while $|U_{+}^{(t)}| < m$ **do**
    Select candidate *is-a* pair with largest PS: $(x_i, y_i) = \arg\max_{(x_i, y_i) \in U_{+}^{(t)}} PS^{(t)}(x_i, y_i)$;
    Remove the pair from $U_{-}^{(t)}$: $U_{-}^{(t)} = U_{-}^{(t)} \setminus \{(x_i, y_i)\}$;
    if $NS^{(t)}(x_i, y_i) + \sum_{(x,y) \in U_{+}^{(t)}} NS^{(t)}(x, y) < \theta$ **then**
        Add the pair to $U_{+}^{(t)}$: $U_{+}^{(t)} = U_{+}^{(t)} \cup \{(x_i, y_i)\}$;
    end if
end while
return Collection of *is-a* relations $U_{+}^{(t)}$;

Outline

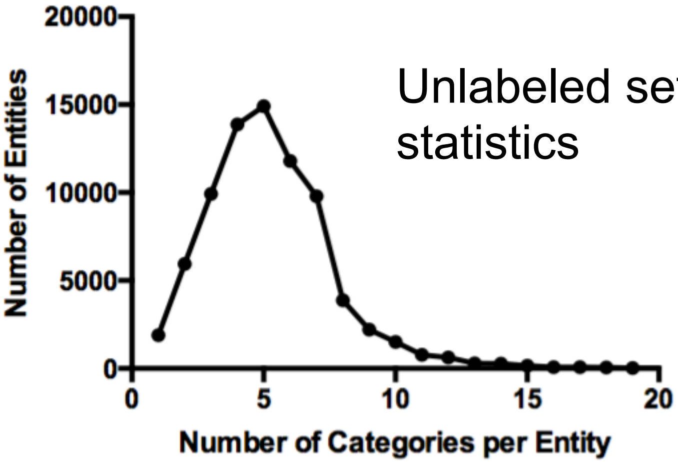
- Introduction
- Background and Related Work
- Proposed Approach
- **Experiments**
- Conclusion

DaSE Data Science & Engineering

Experimental Data

- Text corpus
  - Text contents from Baidu Baike, 1.088B words
  - Train 100-dimensional word vectors using Skip-gram model
- Is-a relation sets
  - Training: A subset of is-a relations derived from a Chinese taxonomy
  - Unlabeled: Entities and categories from Baidu Baike
  - Testing: publicly available labeled dataset (ACL'14)

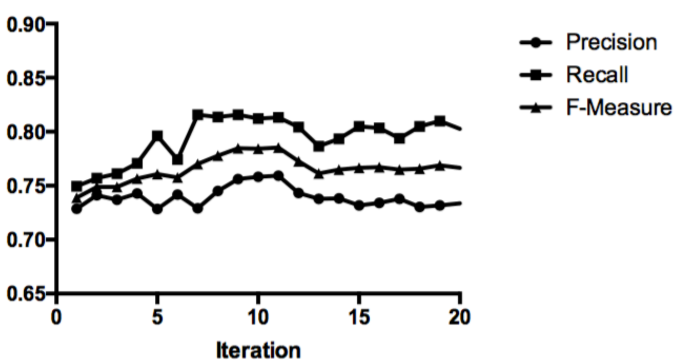
Dataset	Positive	Negative	Unknown
Wiki Taxonomy	7,312	-	-
Unlabeled Set	-	-	78,080
Validation Set	349	1,071	-
Test Set	1,042	3,223	-



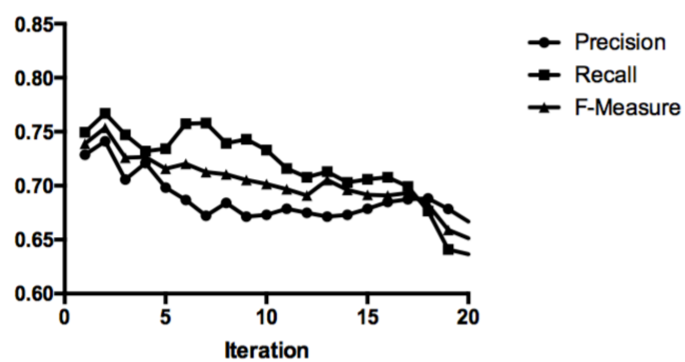
Unlabeled set statistics

Model Performance

- With pattern-based relation selection
  - The performance increases first and becomes relatively stable.
  - A few false positive pairs are still inevitably selected by our approach.
- Without pattern-based relation selection
  - The performance drops quickly despite the improvement in the first few iterations.



(a) With the pattern-based relation selection method



(b) With no relation selection method

Comparative Study

- Comparing with state-of-the-art

Pattern-based

Dictionay-based

Distributed similarity-based

Word embedding-based

Method	Precision (%)	Recall (%)	F-Measure (%)
Previous Methods			
Hearst (Hearst, 1992)	96.2	19.8	32.8
Snow (Snow et al., 2004)	67.3	28.1	39.6
Taxonomy (Li et al., 2015)	**98.5**	25.4	40.4
DSM (Lenci and Benotto, 2012)	48.5	58.1	52.9
Embedding (Fu et al., 2014)	71.7	74.9	73.3
Our Method and Its Variants			
WSRE (Initial)	74.1	76.7	75.3
WSRE (Random)	69.0	75.7	72.2
WSRE (Positive)	75.4	80.1	77.6
WSRE	75.8	81.4	78.6
WSRE+Taxonomy	78.8	**84.7**	**81.6**

Error Analysis

- Hard to distinguish *related-to* v.s. *is-a* relations (approx. 72%)
  - False positives:
    - 中药 (Traditional Chinese medicine), 药草 (Herb)
    - 元帅 (Marshal), 军事家 (Strategist)
- Inaccurate representation learning for fine-grained hypernyms (approx. 28%)
  - True positive:
    - 兰科 (Orchid) , 植物 (Plant)
  - False negative:
    - 兰科 (Orchid) , 单子叶植物纲 (Monocotyledon)

DaSE Data Science & Engineering

Outline

- Introduction
- Background and Related Work
- Proposed Approach
- Experiments
- **Conclusion**

Conclusion

- Chinese is-a relation extraction
  - Initial model training: word embedding based piecewise linear projection model
  - Iterative learning: incremental learning with pattern-based relation selection
  - Application: weakly supervised taxonomy expansion
- Future work
  - Learning generalized Chinese pattern representations for relation extraction

DaSE Data Science & Engineering

Thanks!

Questions & Answers

* The first author would like to thank COLING 2016 for the student support program.