# Detecting Synonymous Predicates from Online Encyclopedia with Rich Features

*AIRS*2016

**Zhe, Han**

Institute of Computer Science & Technology
Peking University

2016 年 11 月 30 日

# Outline

- Predicate Unification Background
- Motivation & Related Work
- Problem Definition
- Features
  - Predicate Representation
- Experiment & Analysis
- Conclusion & Future Work

# background

- Online Encyclopedia => **K**nowledge **B**ases
  - infobox/text => triple


WIKIPEDIA
The Free Encyclopedia

yago
select knowledge

DBpedia

Freebase

## problems on building structured KBs

- taxonomy construction
  - basketball player->sportsman->person
- **predicate standardize/unification**
  - birthday, birthdate; birth place, born place
- value/object standardize/purge
  - 1900-10-02, 02-10-1900
- entity linking
  - Michael Jordan -> {Michael Jordan(player), Michael Jordan(scientist)}

# background

Predicate unification is of great importance and difficulty!

- editor preference
  - too many surface forms
  - concrete vs general
- lack of Chinese KB
  - DBpedia has no linked triples / no predicate set
  - Freebase has fewer Chinese triples
- Chinese
  - lack of resources, like WordNet
  - Pronunciation/typos (坐标,座标)

# Related work

# related work

1. DBpedia
   - Handwritten rules to map wikitext to property set ($\approx 1000$)
   - each kind of infobox/template has their mapping rules
     - no mapping rules are included in Chinese DBpedia
     - birthdate would be written many times if exists in different templates
2. YAGO
   - Lmited predicate set ($\approx 120$) to avoid inside predicate unification
3. Freebase
   - (Tan 2014) detect synonyms based on user domain expertise and co-occurrence of objects and subjects
   - object type info is needed

# related work

4. Abedjan treats syn-pred detection as a association rule mining problem
5. Baroni and Wei find co-occurrence of synonym candidates in web documents
6. Naumann proves effectiveness of aggregate features
7. Li's experiment shows **weak** performance using dictionaries only

# Problem Definition

# Wikipedia Resources


section

predicate

infobox name

wikitext predicate

```
{{expand|time=2015-02-17T12:45:53+00:00}}
{{noteTA|G1=NBA
|1=zh:科里; zh-hans:库里;zh-hk:居里;zh-tw:柯瑞}}
{{Infobox NBA Player
| image = Stephen Curry 2.jpg
| name = 斯蒂芬·科里<br>Stephen Curry
| nickname = 咖哩王子<br>萌神
| position = [[控球後衛]]
| height_ft = 6
| height_in = 3
| weight_lbs = 190
| team = 金州勇士
| number = 30
| nationality = {{USA}}
| birth_date = {{birth date and age|1988|3|14|}}
| birth_place = [[俄亥俄州]][[阿克倫 (俄亥俄州)|阿克倫城]]
| college = {{link-en|大卫森学院|Davidson Wildcats men's basketba
| draft_round  = 1
| draft_pick   = 7
| draft_year   = 2009
| draft_team = [[金州勇士]]
| career_start = 2009年
| former_teams = [[金州勇士]]（2009 - 至今）
| awards =
* [[NBA最有價值球員]]（{{nbay|2014|end}}）
* 2次NBA全明星賽（{{nasg|2014}}–{{nasg|2015}}）
* [[NBA最佳陣容]]第一隊（{{nbay|2014|end}}）
```

图: Wikipedia web info

图: wikitext info

## Problem Definition

- binary classification problem
- given a pair of predicates $pred1$ , $pred2$ from Wikipedia web infoboxes, predicting whether these two are synonyms

- process
  1. give the **representation vector** of each predicate
  2. calculate the **feature vector** from the vector pair
  3. give the association score of this pair from pre-trained classifier
- different from other's work.
  - no structured Chinese KB based on Wikipedia(non-structured objects)
  - other works are on DBpedia/Freebase with **structured objects (type info)**
  - directly on web predicate

# Features

— Predicate Representation

- 7 kinds of features
  - surface form features
  - pinyin features
  - bilingual dictionary features
  - wikitext features
  - wikiSection featureses
  - wikiInfobox features
  - Freebase category features
- combine to a large feature vector

# Features

- **surface form** features & **pinyin** features

| | | | |
|---|---|---|---|
| surfaceForm | 1.$unigram_{(0,1)}$ | 3.$edit\_distance_{(0,1)}$ | 5.$length\_ratio$ |
| | 2.$unigram_{(1,0)}$ | 4.$edit\ distance_{(1,0)}$ | |
| Pinyin | 6.$pinyin\_unigram_{(0,1)}$ | 8.$pinyin\_edit\_distance_{(0,1)}$ | 10.$pinyin\_length\_ratio$ |
| | 7.$pinyin\_unigram_{(1,0)}$ | 9.$pinyin\_edit\_distance_{(1,0)}$ | |

$$unigram_{(1,0)}(pred_1, pred_2) = \frac{character\_overlap(pred_1, pred_2)}{character\_count(pred_1)} \quad (1)$$

$$edit\_distance_{(0,1)}(pred_1, pred_2) = \frac{edit\_distance(pred_1, pred_2)}{character\_count(pred_2)} \quad (2)$$

- **bilingual dictionary** features
  - translate the Chinese predicates to English words
  - same as surface form features

# Features

- **wikitext** features
  - we mapped wikitext-predicates to corresponding web predicates based on objec/value similarity.
  - the **wikitext-predicates-distribution** of predicate (normalized to a unit vector).
  - The **wikitext-predicates-distribution** of predicate 面积*(area)* :

| **wikitext** | 面积*(area)* | area | areatotal | arearank | population total | tarea | 面积排名 | area imperial | ... |
|---|---|---|---|---|---|---|---|---|---|
| **aligned frequency** | 2860 | 1251 | 272 | 163 | 124 | 93 | 72 | 24 | ... |

- **wikiSection** & **wikiInfobox** features
  - similar to wikitext features
  - first section/infobox name ditribution of each web predicates
  - normalized to unit vectors

# Features

- **Freebase category** features
  - Wikipedia orignial category hierarchy is **rejected**
    - circles exists: 冰島 *(Iceland)->* 冰岛地理 *(Iceland geography)->* 冰島島嶼 *(Iceland islands)->* 冰島 *(Iceland)*
    - confusion categories: 含有希伯来语的条目 (*articles containing Hebrew*)
  - collect all the subjects' Freebase types of web predicates, normalized to a unit vector
  - compress to 200-dimensions vector using SVG

# Experiment

# Experiment

## semi-structured KB

- extracted from zh.Wikipedia
- 3.5m s-p-o from 33.8k infoboxes
- subject is entity while object is not
- 11k web predicates

## dataset

- 1500 web predicates pairs
- positive:negtive = 2:1
- selected on the whole predi-set
- 1000 pairs for trainning

- 3 experiments are conducted
  1. Single kind feature experiment
  2. Minus one kind feature experiment
  3. Best feature combination experiment

# Experiment

1. Single kind feature experiment

| **feature** | Accuracy | | | |
|---|---|---|---|---|
| | AdaBoost | SVMR | SVML | VP |
| pinyin | **0.662** | **0.664** | **0.610** | 0.618 |
| surfaceForm | 0.634 | 0.584 | 0.586 | **0.626** |
| Bi-Dictionary | 0.594 | 0.598 | 0.598 | 0.586 |
| FB-Category | 0.568 | 0.580 | 0.562 | 0.582 |
| wikiText | 0.562 | 0.572 | 0.586 | 0.562 |
| wikiSection | 0.518 | 0.526 | 0.522 | 0.532 |
| wikiInfobox | 0.518 | 0.526 | 0.522 | 0.532 |

- SVMR: svm rbf
- SVML: svm linear
- VP: Voted Perceptron

- **pinyin** takes spell mistakes and differenct expressions into account
- **surface form** and **Bi-Dictionary** are good single features

# Experiment

2. Minus one kind feature experiment

| **reduced feature** | Accuracy | | | |
|---|---|---|---|---|
| | SVMR | AdaBoost | SVML | VP |
| -surfaceForm | **0.634** | **0.642** | **0.634** | **0.624** |
| -wikiText | 0.656 | 0.666 | 0.648 | 0.666 |
| -wikiInfobox | 0.680 | 0.670 | 0.676 | 0.670 |
| -wikiSection | 0.680 | 0.670 | 0.676 | 0.670 |
| -Pinyin | 0.688 | 0.666 | 0.688 | 0.676 |
| -Freebase Category | 0.684 | 0.686 | 0.696 | 0.668 |
| -Bilingual Dictionary | 0.698 | 0.666 | 0.678 | 0.692 |

- **surface form** and **wikitext** features are irreplaceable
- **Bi-Dictionary** $\subset$ **wikitext**

# Experiment

3. Best feature combination experiment

| features | accuracy |
|---|---|
| pinyin, surfaceForm, wikiText, wikiSection, wikiInfobox, FB-category | 0.698 |
| pinyin, surfaceForm, wikiText, wikiInfobox, FB-category | 0.694 |
| pinyin, surfaceForm, wikiText, wikiSection, FB-category | 0.694 |
| surfaceForm, wikiText, wikiInfobox, FB-category | 0.688 |
| surfaceForm, wikiText, wikiSection, FB-category | 0.688 |
| surfaceForm, wikiText, wikiSection, wikiInfobox, Bi-Dictionary, FB-category | 0.688 |

- surfaceForm and wikiText are fundamentally useful
- wikiInfobox and wikiSection show efficacy in complex feature combinations

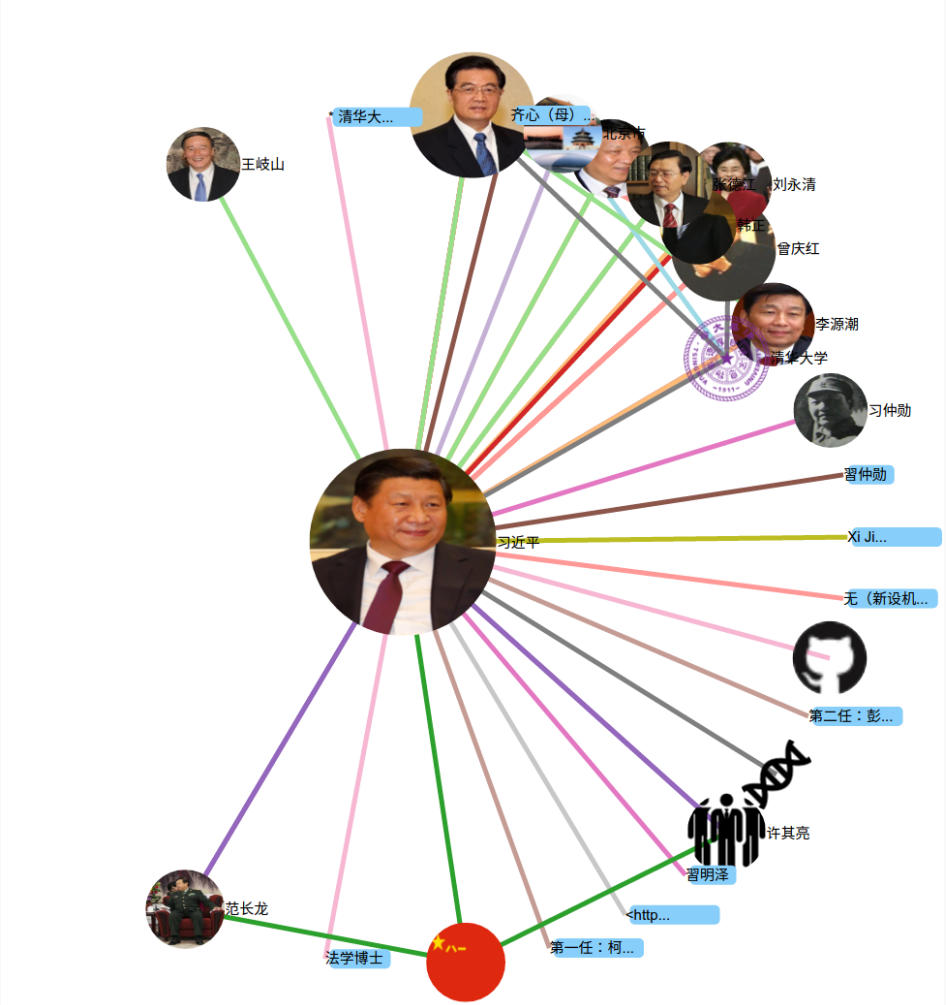
# Conclusion & Future Work

# Conclusion & Future Work

- Full-fledged method on detecting predicate synonyms
  - Thorough study has been done on wikitext
  - wikitext with **frequency independent** features are good combination
  - surface form, category and section information can be used by other encyclopedias
  - groundwork for building Chinese structured KB
- Improvement
  - real-time predicate suggestion when add new triples
  - top 3 relevant wikitext/section name/infobox name in distribution
  - leverage object information
    - basic type: date, candidate entity types, string, number, ...

# Conclusion & Future Work

- Constructing an open-domain Chinese KB
  - Taxonomy, **predicate set**, linked to DBpedia, ...



Thanks

Q & A

# Append