Challenges in Chinese Knowledge Graph Construction

Chengyu Wang, Ming Gao, Xiaofeng He, Rong Zhang

Institute for Data Science and Engineering
East China Normal University
Shanghai, China

Knowledge Graph -

Modeling Knowledge as a Graph

Entities
- Concepts
- Instances
- Values

Relations
- IsA
- Co-occurrence
- Others

Knowledge Graph

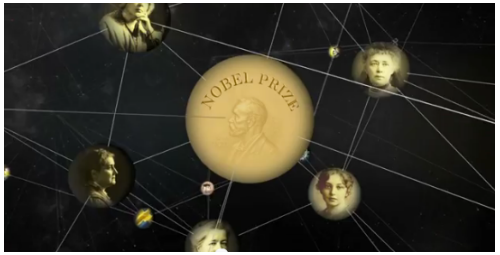
Nodes: entities (concept, named entity, …)
Edges: semantic relationships

DBpedia

yago
select knowledge

ProBase

Freebase



Google Knowledge Graph

Satori (Bing Search)

Chinese Knowledge Graph
Data Sources & Challenges

- Sources
  - Heterogeneous data sources
  - No public knowledge repositories or semantic networks
- Methods
  - Machine translation: low quality
  - Information extraction: difficult

Chinese Wikis:

Chinese Wikipedia (0.8M+ articles)

Baidu Baike (10M+ articles)

Hudong Baike (11M+ articles)

Data Sparsity

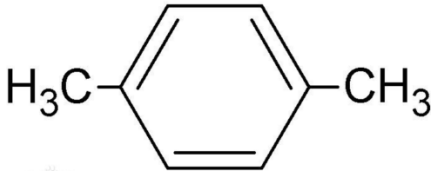
- Comparison between Chinese & English Wikipedias

	Chinese Wikipedia	English Wikipedia	
#Articles	~0.8M	~4M	**5 times!**
#Infoboxes	~0.1M	~1.6M	**13 times!**

- Challenges
  - Entities: extracting long-tailed entities
  - Relations: construction of a "dense" KG
- Solution
  - Data fusion from different sources

Information Accuracy

- "Editing war" on PX (P-Xylene)
  - Polarized attitudes towards plan of a PX factory in city of Xiamen, China
  - Edited 76 times in total
  - Supporters: PX is slightly toxic.
  - Protesters: PX is extremely toxic!
- Challenges
  - Mining editing logs
  - Detecting inaccurate attributes



Editing log on PX

2014-04-02 16:37	查看	溺水三千s
2014-04-02 13:53	查看	1162007677
2014-04-02 12:27	查看	楓之群動
2014-04-01 11:20	查看	cycrc7
2014-03-31 21:56	查看	邓靖轩
2014-03-31 20:41	查看	zhiyuanep1
2014-03-31 18:49	查看	道牙子没事
2014-03-31 17:05	查看	sunbingame
2014-03-31 12:05	查看	道牙子没事
2014-03-31 10:54	查看	道牙子没事
2014-03-30 20:33	查看	847872000
2014-03-30 18:41	查看	亚豆亚豆
2014-03-30 00:09		幻想书生wjc

Link Quality

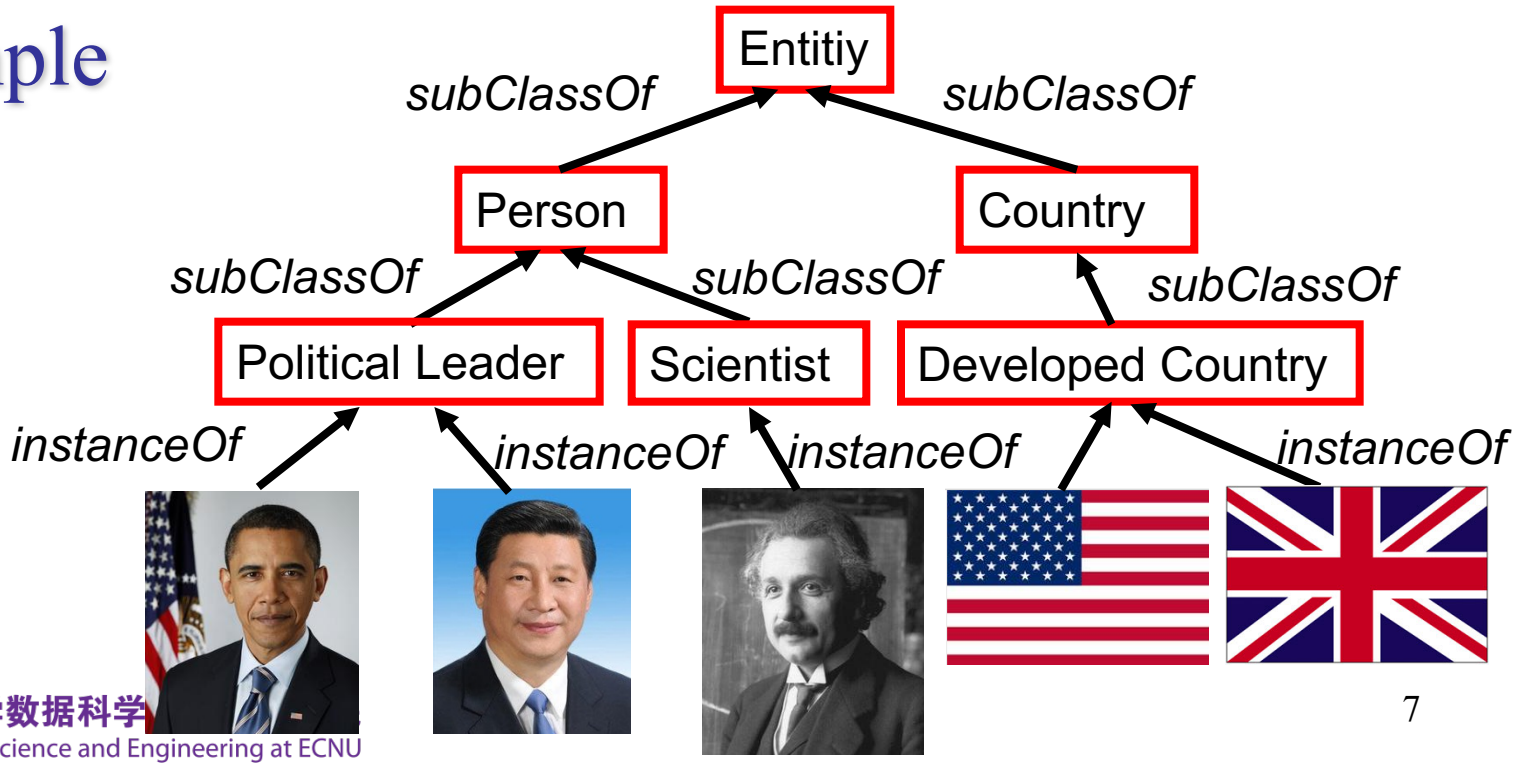
- Hyperlinks in Wikipedia
  - Link entity mentions in texts with corresponding Wikipedia pages
  - Serve as evidence to perform entity linking

    Barack Hussein Obama II is the 44th and current President of the United States, and the first African American to hold the office.

- Wrongly annotated links in Chinese Wikipedia
  - **Wu Mei** (Prof of Peking Univ.) in page **May Fourth Movement** linked to **Wu Mei** (dubbing actress in Hong Kong)
  - Automatic detection of error links in Wikipedia

Taxonomy Derivation

- Taxonomy: a hierarchical type system for KGs
  - **subClassOf** relations (subject: class, object: class)
  - **instanceOf** relations (subject: entity, object: class)
- Example



Taxonomy Derivation

- Challenges in Chinese taxonomy derivation
  - Lack of resources (No Chinese equivalent of WordNet)
  - Hard to map entities to their categories

Research directions

- Language patterns
- Classification
- Machine translation
- Complete taxonomy construction

Xi Jinping (Chinese President)

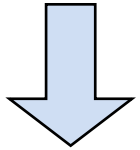
Labels: Person, Politician, Politics, Official

relatedTo? topicOf?
subClassOf? instanceOf?

IsA Extraction

- Hearst patterns (Hearst. COLING'92)
  - such NP as NP,* or|and NP
  - NP such as NP, NP, ..., and|or NP
  - NP, including NP,* or | and NP
  - …

Countries such as **China**, **France** and **Germany**



China isA **Country**
France isA **Country**
Germany isA **Country**

- Chinese IsA patterns
  - Poor NLP analysis in Chinese Web text
  - Lack of explicit high-quality isA patterns
  - Implicit expressions of isA relations

ProBase

Largest taxonomy in English
- 2.6M+ concepts
- 20M+ isA pairs

General Relation Extraction

- Relation extraction systems

  Snowball (SIGMOD'01) KnowItAll (WWW'04)

  LELIA (KDD'06) TextRunner (IJCAI'07)

  StatSnowball (WWW'09) Many others…

  – Focus on English language
- Chinese relation extraction
  – Extract knowledge from semi-structured and structured data
  – Design statistical and NLP-based features for Chinese text
  – Use facts of high precision to supervise RE process (distant supervision)

Conclusion

- Web-scale Chinese KG construction
  - Quality of data sources: data fusion and cleaning
  - Taxonomy derivation: study on taxonomic relations in Chinese
  - Knowledge harvesting: isA patterns, Chinese RE systems

Lots of challenges

Lots to do!

Thanks!

Questions & Answers