Cross-lingual Knowledge Validation Based Taxonomy Derivation from Heterogeneous OnlineWikis

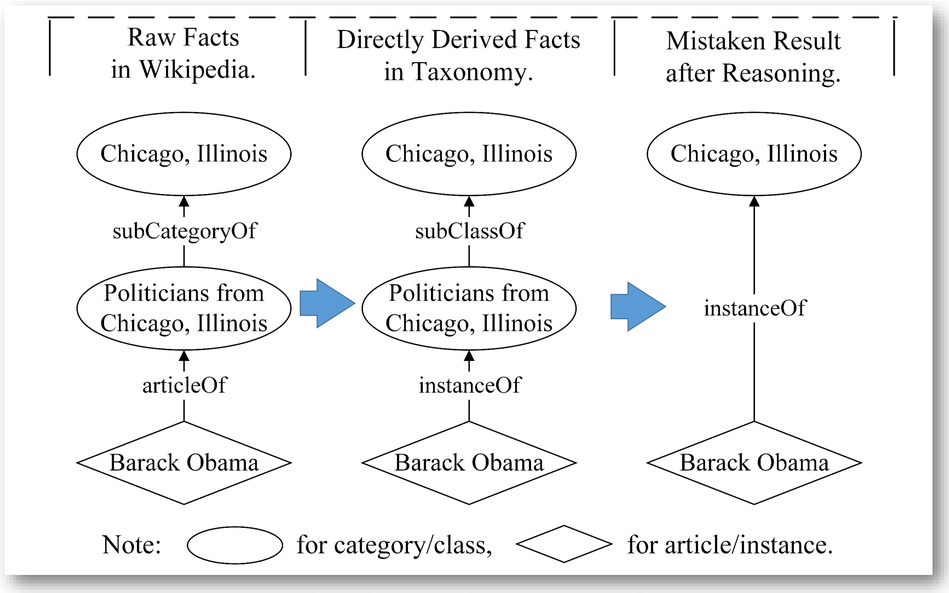
Zhigang Wang[1], Juanzi Li[1], Shuangjie Li[1], Mingyang Li[1], Jie Tang[1], Kuo Zhang[2], Kun Zhang[2]

1 TSINGHUA UNIVERSITY, BEIJING, CHINA **2** SOGOU INCORPORATION, BEIJING, CHINA

Introduction

- Creating KBs based on the crowd-sourced wikis has attracted significant research interest in the field of intelligent Web.
- However, the *user-generated subsumption relations* in the wikis and the *semantic taxonomic relations* in the KBs are not exactly the same.
- Current taxonomy derivation approaches include:
  - The heuristic-based methods
  - The corpus-based methods

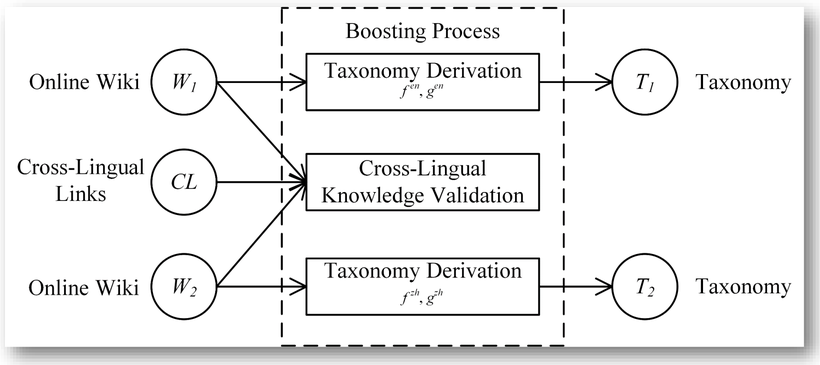
- Here, we systematically study the problem of **cross-lingual knowledge validation based taxonomy derivation** from heterogeneous online wikis.
- The problem of **cross-lingual taxonomic relation prediction** is at the heart of our work.



Example of Mistaken Derived Facts

Approach

Given two wikis W_1, W_2 in different languages (English and Chinese here) and the set of cross-lingual links CL, **Cross-lingual Taxonomy Derivation** is a cross-lingual knowledge validation based boosting process, by simultaneously learning four taxonomic prediction functions f^{en}, f^{zh}, g^{en} and g^{zh} in T iterations.



Framework

where f^{en}, f^{zh}, g^{en} and g^{zh} denote the English ***subClassOf***, the Chinese ***subClassOf***, the English ***instanceOf***, and the Chinese ***instanceOf*** prediction functions respectively.

Dynamic Adaptive Boosting (**DAB**) model is to maintain a dynamic changed training set to achieve a better generalization ability via knowledge validation with *cross-lingual links*.

1. Weak Classifier

We utilize the binary classifier for the basic learner and use the Decision Tree as our implementation.

Linguistic Heuristic Features

Feature 1: English Features.
Whether the head words of *label* are plural or singular.

Feature 2: Chinese Features.
Whether the super-category's *label* is the prefix/suffix of the sub-category's *label*. Or, whether the category's *label* is the prefix/suffix of the article's *label*.

Feature 3: Common Features for *instanceOf*.
Whether the *comment* contains the *label* or not.

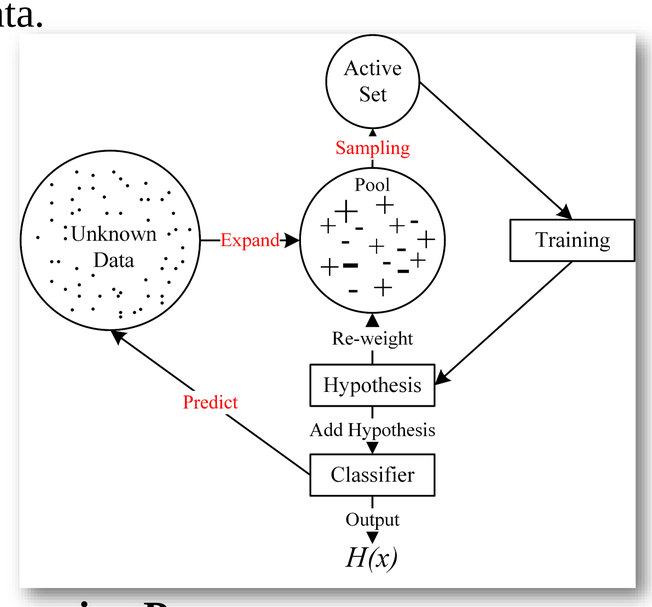
Structural Features

Six *Normalized Google Distance* based structural features are defined on *articles*, *properties* and *categories*.

$$d_a(c, c') = \frac{\log(\max(|A(c)|, |A(c')|)) - \log(|A(c) \cap A(c')|)}{\log(|A|) - \log(\min(|A(c)|, |A(c')|))}$$

2. Boosting Model

Active Set A: the set of training data.
Pool P: the set of all labeled data.
Unknown Data Set U: the set of unlabeled data.



Learning Process.

- Train a hypothesis on current active set.
- Re-weight the weight vector.
- **Predict** U using current classifier and validate the results using CL.
- **Expand** P and **update** U.
- **Resample** A with the constant size.

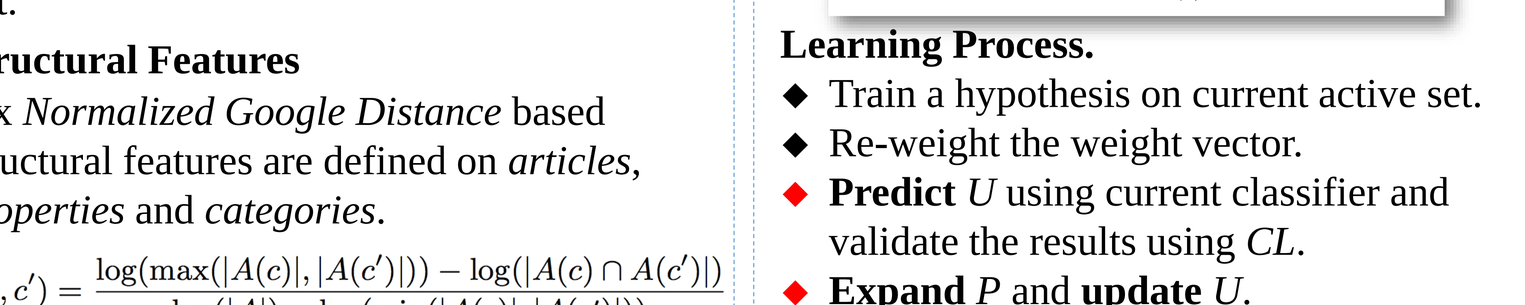
Experiments

Comparison Methods

- **Heuristic Linking (HL):** only uses the linguistic heuristic features, and trains the taxonomic relation prediction functions separately using the decision tree model.
- **Decision Tree (DT):** uses both the linguistic heuristic features and the structural features, and trains the taxonomic relation prediction functions separately using the decision tree model.
- **Adaptive Boosting (AdaBoost):** uses the same basic learner, and iteratively trains the taxonomic relation prediction functions using the real AdaBoost model.

Performance of Cross-lingual Taxonomy Derivation with Different Methods (%)

Methods	English SubClassOf			Chinese SubClassOf			English InstanceOf			Chinese InstanceOf		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
HL	87.1	81.3	84.1	91.4	91.4	91.4	94.3	89.4	91.8	42.4	51.9	46.7
DT	88.7	86.9	87.8	90.9	92.0	91.4	91.9	95.6	93.7	46.8	58.1	51.8
AdaBoost	90.8	90.9	90.9	91.4	92.3	91.8	94.3	94.1	94.2	51.4	63.9	57.0
DAB	90.7	91.8	**91.2**	91.1	95.2	**93.1**	94.1	97.7	**95.9**	77.8	75.0	**76.4**



(a) English **subClassOf** (b) Chinese **subClassOf** (c) English **instanceOf** (d) Chinese **instanceOf**

Boosting Contribution Comparison

Conclusion and Future Work

- DAB gives a new way for language processing tasks using cross-language resources.
- The future work contains automatically learning more cross-lingual validation rules and conducting more experiments in other languages.

References

- de Melo, G., and Weikum, G. 2010. Menta: Inducing multilingual taxonomies from Wikipedia. In CIKM'10.
- Potthast, M., Stein, B., and Anderka, M. 2008. A Wikipedia-based multilingual retrieval model. In ECIR'08.
- Wang, Z.; Li, J.; Wang, Z.; and Tang, J. 2012. Cross-lingual knowledge linking across wiki knowledge bases. In WWW'12.