Boosting to Bulid a Large-scale Cross-lingual Ontology

Zhigang Wang, Liangming Pan, Juanzi Li
Knowledge Engineering Group
Department of Computer Science & Technology
Tsinghua University

Tsinghua University

Outline

Motivation

Problem Definition

Methods

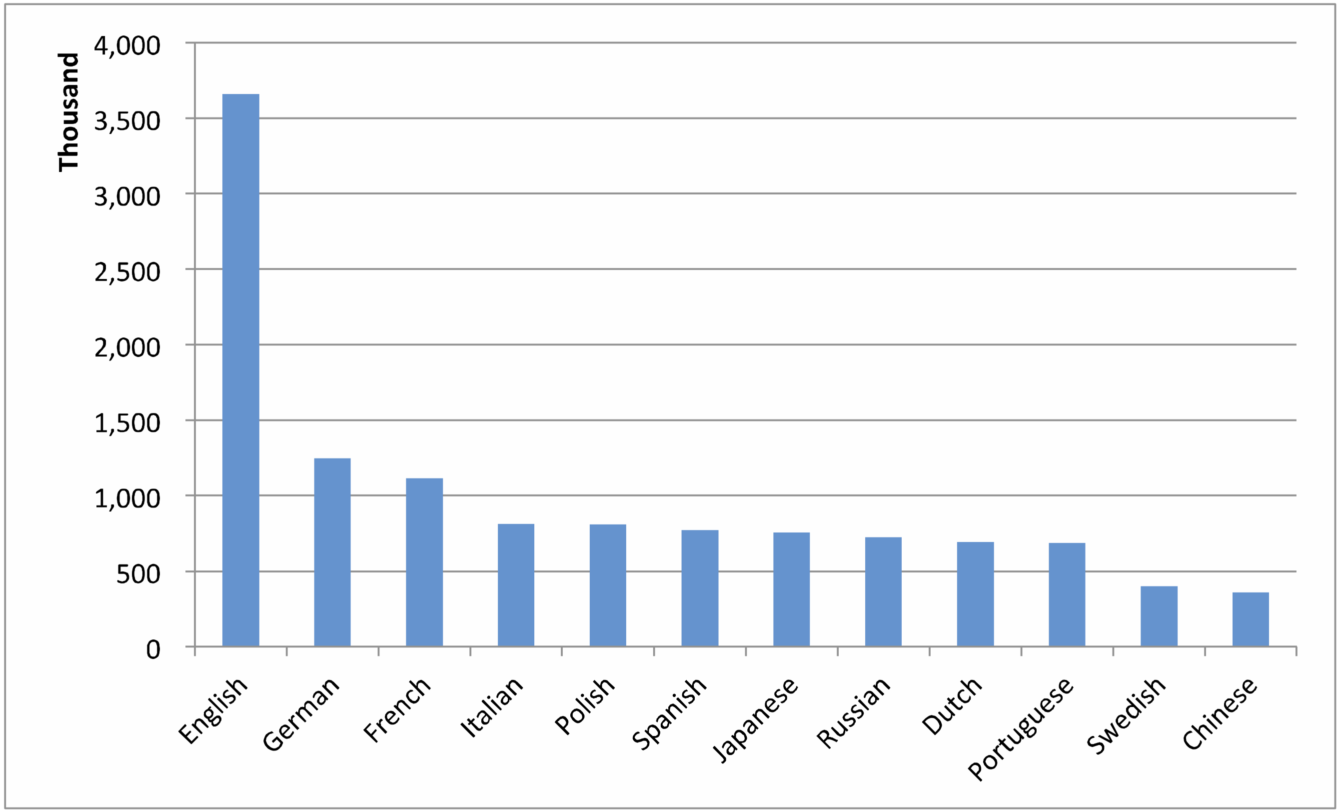
Experiments and Analysis

Conclusion

Motivation

Motivation 1. The scarcity of non-English knowledge

- The imbalanced sizes of different Wikipedia language versions lead to the highly imbalanced knowledge distribution in different languages.



Motivation

Motivation 2. The noise in the multi-lingual ontology schema relations

Table 1. Examples of Semantic Relations

Entity 1	Relation	Entity 2	Right or Wrong
European Microstates	`instanceOf`	Microstates	Right
European Microstates	`instanceOf`	Europe	Wrong
教育人物(Educational Person)	`subClassOf`	人物(Person)	Right
教育人物(Educational Person)	`subClassOf`	教育(Education)	Wrong

Motivation 3. The limited coverage of cross-lingual links

- The amount of integrated multilingual knowledge totally depends on these existing cross-lingual links
- There are less existing CLs or none at all between different wikis.

Motivation

- **Summary**
  - Current status of famous multi-lingual ontologies
    - the scarcity of non-English knowledge
    - the noise in the multi-lingual ontology schema relations
    - the limited coverage of cross-lingual equivalent relations
  - Possible solutions
    - Build an ontology using the large-scale non-English resources
    - Predict the correct semantic relations between two entities
    - Iteratively mine more cross-lingual links

Outline

Motivation

Problem Definition

Methods

Experiments and Analysis

Conclusion

Problem Formalization

Input

- Given two cross-lingual online encyclopedias,
  - $G_1 = (V, E)$ (In English), $G_2 = (V', E')$ (In Chinese)
  - $v \in V$
    - v denotes an *entity*
    - Each *entity* has an corresponding *document*
  - $E = V \times V$
    - $e_{ij} = 1$ represents that v_i $\boldsymbol{subCategoryOf}$ *or* $\boldsymbol{articleOf}$ v_j
- and an input alignment
  - $A = \{a_i\}_{i=1}^{m}$
  - $a_i = (v, v')$

Problem Formalization

Output

- Our target is to build a cross-lingual ontology, which contains
  - $O_1 = (X, Y), O_2 = (X', Y'),\ \ A' = \{a_i'\}_{i=1}^{n}$
- where,
  - $X \subseteq V,\ X' \subseteq V',\ Y = Y' = \{0,1\}$
    - $x \in X \Leftrightarrow x \leftrightarrow concept$ or $x \leftrightarrow instance$
    - $y_{ij} = 1 \Leftrightarrow x_i\ subClassOf\ x_j, x_i x_j \leftrightarrow concept$

      Or $x_i\ instanceOf\ x_j, x_i \leftrightarrow instance, x_j \leftrightarrow concept$
  - $a_i' = (x, x')$
  - $\mathrm{n} \gg m$

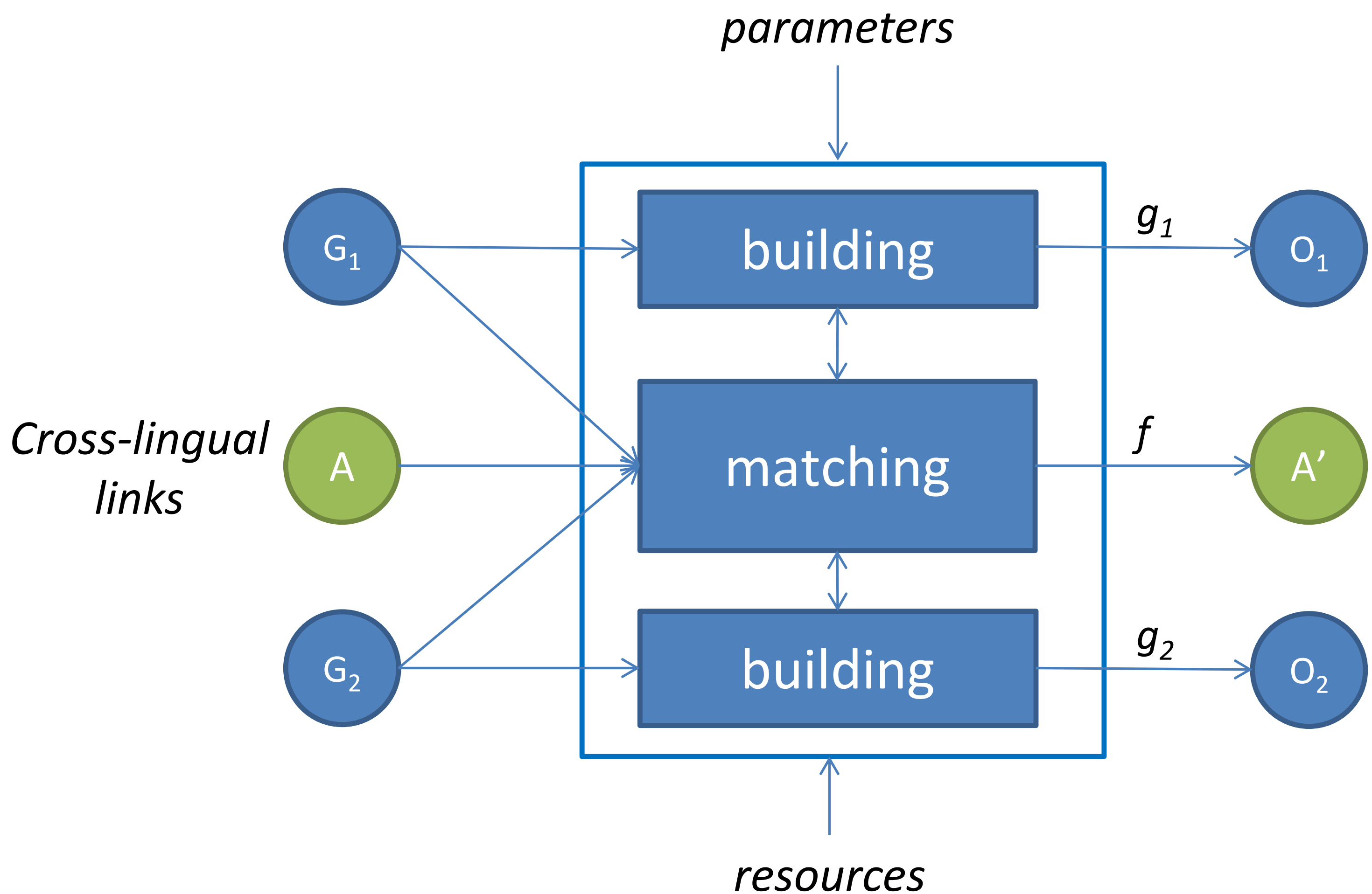
Outline

- Motivation
- Problem Definition
- **Methods**
- Experiments and Analysis
- Conclusion

Method: Framework

- We attempt to learn two kinds of functions,
  - $f: V \times V' \rightarrow [0,1]$ (***ontology matching***)
    - $w \in [0,1]$ represents the confidence to be equivalent between *v* and *v'*
  - $g: V \times V \rightarrow [0,1]$ (***ontology building***)
    - $w \in [0,1]$ represents the confidence to be *subClassOf or InstanceOf* between *v* and *v'*

Method: Framework



Learning *f*

Boosting of *f*, *g*

Learning *g*

Method: Mono-lingual Ontology Building

- We define several useful features and apply the Logistic Regression model to learn the classifier.

Table 2. Feature Definition for g_1

ID	Feature	Range
1	Is the head word of super-category plural?	$\{0,1\}$
2	Is the head word of sub-category plural?	$\{0,1\}$
3	Word length of super-category	Integer
4	Word length of sub-category	Integer
5	Word length of head words of super-category	Integer
6	Word length of head words of sub-category	Integer
7	Relation between the head words of super-category and sub-category	$\{\equiv, \subseteq, \supseteq, \perp, \triangle\}$
8	Does the non-head words of sub-category contain the head words of super-category?	$\{0,1\}$
9	Does the non-head words of super-category contain the head words of sub-category?	$\{0,1\}$
10	Score of sub-category	Numeric
11	Score of super-category	Numeric

$\equiv$ equivalent, $\subseteq$ smaller, $\supseteq$ larger, $\perp$ disjoint, $\triangle$ otherwise.

Method: Cross-lingual Instance Matching

- We first define the Set Similarity between entity a and b as follows.

$$s(a,b) = \frac{2 \cdot |\phi_{1\to 2}(S_a \cap S_b)|}{|\phi_{1\to 2}(S(a))| + |S(b)|}$$

  - S_a and S_b are the related sets of entities a and b
  - where $\phi_{1\to 2}(\cdot)$ maps the set of entities in G_1 (or O_1) to their equivalent entities in G_2 (or O_2) if the alignment exists.

Method: Cross-lingual Instance Matching

- Similar as Ontology Building, we also define some features and apply the Logistic Regression model to learn the classifier.
- Both the ***lexical similarities*** and ***link-based structural similarities*** are defined.

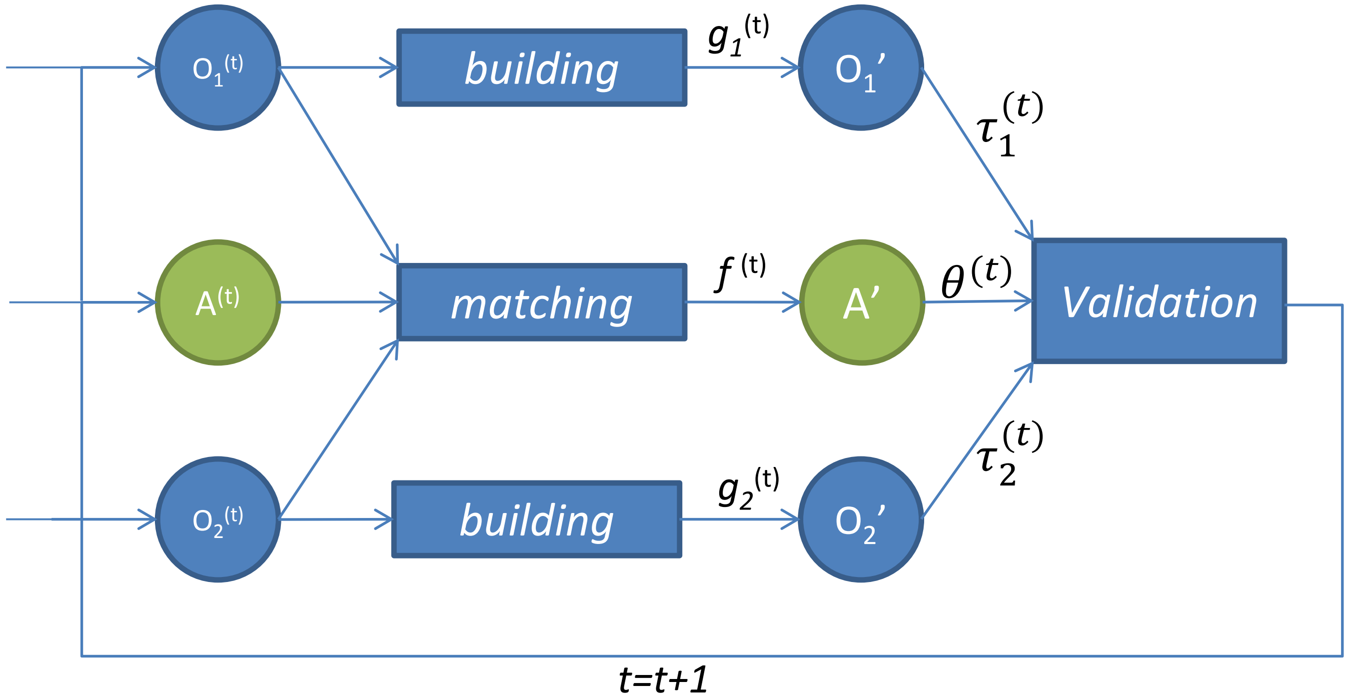
Table 3. Feature Definition for f

Type	ID	Feature	Description
Lexical	1	Edit-distance of titles without translation	Return 0 if there are no common characters.
	2	Difference in word length	$\lvert English\_Word\_Length - Chinese\_Character\_Length \rvert$.
Structural	3	*Set Similarity* of categories	Calculated between $\mathcal{G}_1$ and $\mathcal{G}_2$
	4	*Set Similarity* of outlinks	Calculated between $\mathcal{G}_1$ and $\mathcal{G}_2$
	5	*Set Similarity* of inlinks	Calculated between $\mathcal{G}_1$ and $\mathcal{G}_2$
	6	*Set Similarity* of concepts	Calculated between $\mathcal{O}_1$ and $\mathcal{O}_2$

Method: Boosting to Build a Large-scale Ontology

- To boost a large-scale cross-lingual ontology, we iteratively learn the ***ontology building functions*** and the ***instance matching function***.

For each iteration $t = 1 \sim N$,



Method: Boosting to Build a Large-scale Ontology

- Update O_1' and O_2'
  - Train $g_1^{(t)}$ and $g_2^{(t)}$ based on current training set
  - If $f^{(t)}(x_1, x_1') > \theta^{(t)}$ and $f^{(t)}(x_2, x_2') > \theta^{(t)}$ then

$$\begin{cases} g_1^{(t)}(x_1, x_2) = g_2^{(t)}(x_1', x_2') = 1 & \text{If } g_1^{(t)}(x_1, x_2) + g_2^{(t)}(x_1', x_2') > (\tau_1^{(t)} + \tau_2^{(t)}) \\ g_1^{(t)}(x_1, x_2) = g_2^{(t)}(x_1', x_2') = 0 & \text{OtherWise} \end{cases}$$

  - Expand the training set

- Update A'
  - Train $f^{(t)}$ using current alignments
  - If $f^{(t)}(x_1, x_1') > \theta^{(t)}$ then $f^{(t)}(x_1, x_1') = 1$
  - Expand the alignment set

Outline

- Motivation
- Problem Definition
- Methods
- **Experiments and Analysis**
- Conclusion

Experiments

- DataSets
  - English Wikipedia dump
    - Archived in August 2012
  - Hudong Baike dump
    - Crawled from Huong Baike's website in May 2012

Table 4. Statistics of Cleaned Data Sets

Online Wiki	#Categories	#Articles	#Links	#Links/#Articles
English Wikipedia	561,819	3,711,928	63,504,926	17.1
Hudong Baike	28,933	980,411	23,294,390	23.8

  - 126,221 alignments between English Wikipedia and Hudong Baike
  - Labeled data for Ontology Building

Table 5. Labeled Data for Mono-lingual Ontology Building.

Examples	subClassOf en	subClassOf zh	instanceOf en	instanceOf zh
Positive	2,123	780	2,097	638
Negative	787	263	381	518

en: English, **zh**: Chinese.

Experiments

Experimental Results

Results of Mono-lingual Ontology Building

Table 6. Results of Mono-lingual Ontology Building. (%)

Methods	subClassOf en			subClassOf zh			instanceOf en			instanceOf zh		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
NB	87.1	62.5	72.8	87.1	85.9	86.5	95.8	42.7	59.1	60.2	55.7	57.9
SVM	80.8	86.7	83.6	83.8	98.6	**90.6**	84.5	100	91.6	53.1	82.1	64.5
LR	80.6	87.1	**83.7**	84.0	97.7	90.3	87.4	98.4	**92.6**	56.5	80.1	**66.3**

P: precision, R: recall, F1: F1-measure, en: English, zh: Chinese.

Results of Cross-lingual Instance Matching

Table 7. Results of Cross-lingual Instance Matching. (%)

#Alignments	Before HP			After HP		
	Precision	Recall	F1-measure	Precision	Recall	F1-measure
0.03 Mil.	81.5	5.6	10.5	91.4	5.6	10.6
0.06 Mil.	86.4	6.0	11.3	91.9	6.0	11.3
0.09 Mil.	**89.7**	6.5	12.0	**93.9**	6.5	12.2
0.12 Mil.	86.5	6.8	**12.6**	88.9	6.8	**12.6**

Experiments

Experimental Results

- Results of building subClassOf relation

No. of Iteration	English *subClassOf*			Chinese *subClassOf*		
	Precision	Recall	F1-measure	Precision	Recall	F1-measure
Iteration 1.	80.82%	88.24%	84.36%	81.99%	100.00%	90.10%
Iteration 2.	87.28%	91.77%	89.46%	91.77%	98.64%	95.08%
Iteration 3.	87.70%	93.38%	90.45%	94.81%	99.26%	96.98%

- Results of building instanceOf relation

No. of Iteration	English *instanceOf*			Chinese *instanceOf*		
	Precision	Recall	F1-measure	Precision	Recall	F1-measure
Iteration 1.	87.37%	97.14%	92.00%	65.04%	63.00%	64.00%
Iteration 2.	93.33%	98.44%	95.81%	91.38%	89.08%	90.218%
Iteration 3.	97.25%	99.61%	98.42%	96.72%	97.04%	96.88%

- Results of built ontology

	Number of Concepts	Number of Instances	Number of *subClassOf*	Number of *instanceOf*
English	479,040	3,520,765	751,154	11,339,698
Chinese	24,243	803,278	29,655	2,144,000

Outline

Motivation

Problem Definition

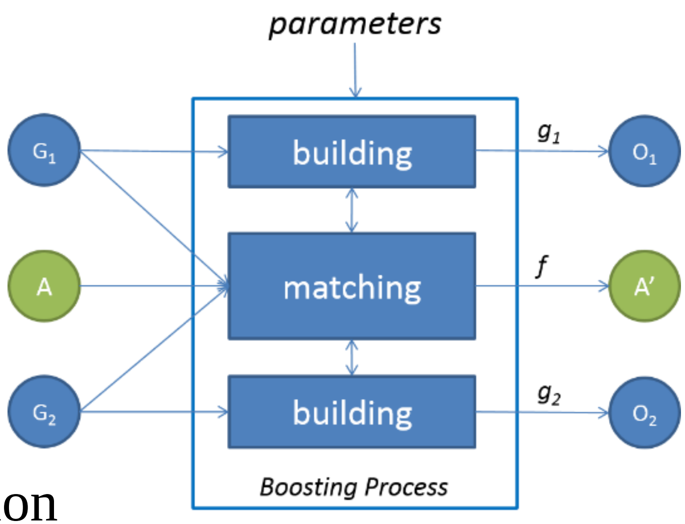
Methods

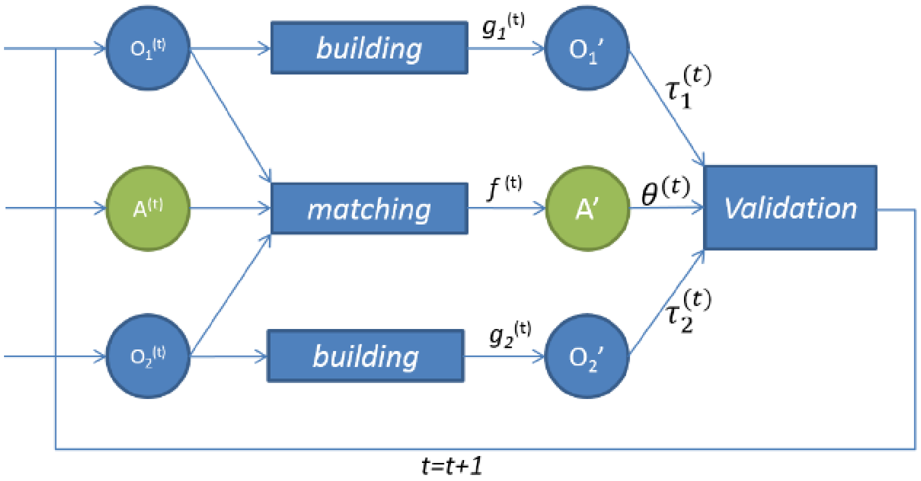
Experiments and Analysis

Conclusion

Conclusion

- **Approach**
  - Learn two kinds of functions
    - Ontology building function
      - $g_1: V \times V \rightarrow [0,1]$ and $g_2: V' \times V' \rightarrow [0,1]$
    - Instance matching function
      - $f: X \times X' \rightarrow [0,1]$
  - Boosting method with cross-lingual validation
    - Heuristics-based cross-lingual validation





Thanks!

Liangming Pan
KEG, THU
peterpan10211020@163.com