Boosting to Build a Large-Scale Cross-Lingual Ontology

Zhigang Wang(✉), Liangming Pan, Juanzi Li, Shuangjie Li, Mingyang Li,
and Jie Tang

Department of Computer Science and Technology,
Tsinghua University, Beijing 100084, People's Republic of China
{wzhigang,plm,ljz,lsj,lmy,tangjie}@keg.cs.tsinghua.edu.cn

Abstract. The global knowledge sharing makes large-scale multi-lingual knowledge bases an extremely valuable resource in the Big Data era. However, current mainstream Wikipedia-based multi-lingual ontologies still face the following problems: the scarcity of non-English knowledge, the noise in the multi-lingual ontology schema relations and the limited coverage of cross-lingual `owl:sameAs` relations. Building a cross-lingual ontology based on other large-scale heterogenous online wikis is a promising solution for those problems. In this paper, we propose a cross-lingually boosting approach to iteratively reinforce the performance of ontology building and instance matching. Experiments output an ontology containing over 3,520,000 English instances, 800,000 Chinese instances, and over 150,000 cross-lingual instance alignments. The F1-measure improvement of Chinese `instanceOf` prediction achieve the highest 32%.

Keywords: Ontology building · Instance matching · Cross-lingual

1 Introduction

As the Web is evolving to a highly globalized information space, sharing knowledge across different languages is attracting increasing attentions. Multilingual ontologies, in which the cross-lingual equivalent concepts or relationships are linked together using `owl:sameAs`, are important sources for harvesting cross-lingual knowledge from the Web and have significant applications such as multilingual information retrieval, machine translation and deep question answering. DBpedia [1], by extracting structured information from Wikipedia in 111 different languages, is a multi-lingual multi-domain knowledge base and becomes the nucleus of LOD. Obtained from WordNet and Wikipedia, YAGO, MENTA, and BabelNet are other famous large multi-lingual ontologies [6,7,12].

Though lots of researches have been done, there are still some problems to be solved. Firstly, the imbalance of different Wikipedia language versions leads to the highly unbalanced knowledge distribution in different languages. Figure 1 shows a simplified long tail distribution of the number of articles on

H. Chen et al. (Eds.): CCKS 2016, CCIS 650, pp. 41–53, 2016.
DOI: 10.1007/978-981-10-3168-7_5

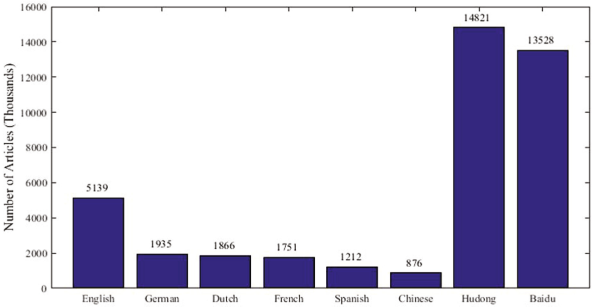

Fig. 1. Number of articles on major wikipedias, Hudong Baike and Baidu Baike

six major Wikipedia language versions. Most non-English knowledge in these ontologies is pretty scarce. Secondly, the noise of the large category system in Wikipedia leads to the incorrect semantic relations in these ontologies. For example, "Wikipedia-books-on-people is the `subCategoryOf` People" will lead to the wrong "Wikipedia-books-on-people is `subClassOf` People" in DBpedia's SKOS schema. And the relatively precise WordNet only cover some aspects of domains in English. Finally, because those ontologies are integrated directly by Wikipedia's cross-lingual links, the coverage of cross-lingual `owl:sameAs` relations in those ontologies is limited by the number of existing cross-lingual links.

On the other hand, there are more and more similar large-scale non-English online wikis in big data era. For example, the Chinese Hudong Baike and Baidu Baike, both containing more than 6 million articles, are even larger than the English Wikipedia (the largest Wikipedia language version). If multi-lingual ontology could be established between two large online wikis, such as English Wikipedia and Chinese Hudong Baike, multi-lingual ontologies with much higher coverage can be constructed.

In this paper, we try to build a large-scale cross-lingual ontology based on two heterogeneous online wikis in different languages. To our best of knowledge, we are the first to combine the processes of mono-lingual ontology building and cross-lingual instance matching together to build a cross-lingual ontology. Our work is motivated by two observations on the multilingual knowledge distributions. ***Cross-lingual Knowledge Consistency***. A lot of facts are considered as correct all over the world, e.g. the facts about Science. Mining consistency across different languages not only helps to match equivalent cross-lingual knowledge, but also assists to improve the performance of mono-lingual ontology building each other. ***Cross-lingual Knowledge Discordance***. The facts people concern or believe are quite different. E.g. the Chinese instance "China" is more linked to the Chinese locations but the English instance "China" is more linked to the counties in the world. Consideration of this problem in depth can help avoid incorrect matching.

This non-trivial task poses the challenges as follows, how to build two large-scale mono-lingual ontologies with correct semantic relations? How to construct an effective and efficient language-independent instance matching model? And how to boost the building of the cross-lingual ontology iteratively? Driven by these challenges, we propose a unified boosting framework to iteratively build a cross-lingual ontology. Our contributions are as follows.

1. We propose a binary classification-based method for large-scale mono-lingual ontology building, and a language-independent instance matching method. The ontology building method is able to eliminate the noise inside the wikis by predicting the correct `subClassOf` and `instanceOf` relations. The ontology matching method works for two highly heterogenous cross-lingual ontologies effectively and efficiently.
2. We propose a cross-lingually boosting method to reinforce the processes of ontology building and instance matching. The cross-lingual knowledge consistency and discordance are analyzed in depth. We iteratively expand the volume of labeled data for ontology building and expand the cross-lingual alignments for instance matching to improve the quality of built ontology simultaneously.
3. We conduct an experiment using the English Wikipedia and Hudong Baike data sets. Experimental results show that our boosting method outperforms the non-iterative method. The F1-measure of ontology building functions has an improvement of above 6%. In particular, the performance of Chinese `instanceOf` function get a high 32% improvement for F1-measure. A large ontology containing 3,520,000 English instances and 800,000 Chinese instances is built. Over 150,000 cross-lingual instance alignments are constructed.

2 Preliminaries

Basic Concepts. Given two online wikis in different languages and an initial aligment set, our target is to build two mono-lingual ontologies and find the equivalent alignments between them.

Definition 1. An ***online wiki*** is a graph containing a set of entities and a set of links between two entities. It can be formally represented as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $v \in \mathcal{V}$ denotes an *entity* and has an related *document*. We have $\mathcal{E} = \mathcal{V} \times \mathcal{V}$, and $e_{ij} \in \mathcal{E}$ indicate whether there exists a `subCategoryOf` or `articleOf`[1] relation from v_i to v_j (1 for yes, 0 for no).

Definition 2. An ***ontology*** is defined as a 2-tuple of the set of entities and the set of semantic relations. It can be formally represented as $\mathcal{O} = (\mathcal{X}, \mathcal{Y})$, where $x \in \mathcal{X}$ denotes a *concept* in the schema-level or an *instance* in the instance-level. We have $\mathcal{Y} = \mathcal{X} \times \mathcal{X}$ and $y_{ij} \in \mathcal{Y}$ indicate whether there exists a legal semantic relation from y_i to y_j (1 for yes, 0 for no). We only consider two kinds of semantic relations, which are `subClassOf` between two concepts and `instanceOf` from one instance to one concept.

[1] We use category and article to denote the concept and instance in the online wiki respectively.

Definition 3. The ***alignment set*** is the set of equivalent instances between two ontologies. It can be formally represented as $\mathcal{A} = \{a_i\}$, where $a_i = (x, x')$ denotes the equivalent instances between two ontologies respectively.

Problem Formulation. Given two online wikis $\mathcal{G}_1 = (\mathcal{V}, \mathcal{E})$, $\mathcal{G}_2 = (\mathcal{V}', \mathcal{E}')$ and an initial alignment set $\mathcal{A} = \{a_i\}_{i=1}^m$, we aim at constructing two monolingual ontologies $\mathcal{O}_1 = (\mathcal{X}, \mathcal{Y})$, $\mathcal{O}_2 = (\mathcal{X}', \mathcal{Y}')$ and a cross-lingual alignment set $\mathcal{A}' = \{a_i\}_{i=1}^n$. We have $n > m$, and $\mathcal{G}_1$, $\mathcal{G}_2$ are in two different languages[2]. The entities of the constructed ontologies are from the entities of online wikis, where $\mathcal{X} \subseteq \mathcal{V}$ and $\mathcal{X}' \subseteq \mathcal{V}'$. Thus, our major issue is to predict three kinds of relations, which are `subClassOf` between two concepts in each ontology, `instanceOf` from one instance to one concept in each ontology, and `equalTo` between two instances from two ontologies.

We formalize this problem as multiple binary classification problems. More formally, we are to learn two kinds of classification functions with a confidence output as follows.

- ***Instance Matching Function*** $f : \mathcal{X} \times \mathcal{X}' \mapsto [0, 1]$ to predict the probability to be `equalTo` relation between two instances x and x' from $\mathcal{O}_1$ and $\mathcal{O}_2$ respectively.
- ***Ontology Building Function*** $g_1 : \mathcal{V} \times \mathcal{V} \mapsto [0, 1]$ to predict the probability to be `subClassOf` or `instanceOf` relation between two entities v_i and v_j in $\mathcal{G}_1$, or $g_2 : \mathcal{V}' \times \mathcal{V}' \mapsto [0, 1]$ in $\mathcal{G}_2$.

To improve the performance of the isolated functions, we boost to mutually reinforce the learning of the building and matching functions.

3 Approach

As shown in Fig. 2, our approach is a boosting method. In each iteration we use the results of ontology building g_1, g_2 and instance matching f to reinforce the learning performance in the next iteration.

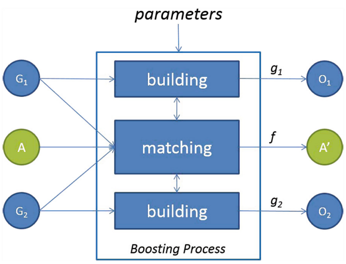

Fig. 2. Overview of the proposed approach

[2] We use $\mathcal{G}_1$ to represent the English online wiki, and use $\mathcal{G}_2$ to represent the Chinese online wiki.

3.1 Mono-lingual Ontology Building

We take the entities of $\mathcal{V}$, $\mathcal{V}'$ in the online wikis $\mathcal{G}_1$ and $\mathcal{G}_2$ as the entities of $\mathcal{X}$, $\mathcal{X}'$ in the ontoloties $\mathcal{O}_1$ and $\mathcal{O}_2$. Concretely, we take the categories in wikis as the concepts, and take the articles as the instances. Hence, our task is to learn the ontology building functions g_1 and g_2 to predict the correct `subClassOf` or `instanceOf` relations between two entities. We view both the correct `subClassOf` relation between two concepts and the correct `instanceOf` relation from an instance to a concept as an `is-a` relation. Table 1 shows some examples about the semantic relations generated from the online wikis.

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

In this paper, we are to learn two series of functions $g_1 : \mathcal{V} \times \mathcal{V} \mapsto [0, 1]$ and $g_2 : \mathcal{V}' \times \mathcal{V}' \mapsto [0, 1]$ to predict the probabilities to be an `is-a`relation between two entities (1 for completely positive, 0 for completely negative). Notice that, we actually train four functions which are English `subClassOf`, English `instanceOf`, Chinese `subClassOf` an Chinese `instanceOf`, but we uniformly represent the ontology building functions of `subClassOf` and `instanceOf` in one language the same. The unique difference between them is that the input entities of `subClassOf` are two concepts but the input entities of `instanceOf` are one instance and one concept.

By manually labeling some training examples, we can learn the Logistic Regression models to get the ontology building functions g_1 and g_2. Table 2 shows the feature definition of g_1 function. The 10th feature is calculated as follows. We firstly list all of the sub-categories of current super-category. Then we calculate the frequency of each word in all of the sub-categories. The score of current sub-category is the sum of the frequency of each word in current sub-category. This feature is similar to a voting process, in which the more frequent words denote a higher probability. Similar as the 11th feature is.

The features in Table 2 are for learning the `subClassOf` predictor of g_1. The `instanceOf` features are similar, in which we replace the super-category into category and replace the sub-category into article. The head words can be extracted using a NLP parser. Note that, for features of g_2, we revise the 1st and 2nd features into "Is the sub-category starting with super-category" and "Is the sub-category ending with super-category" respectively. Besides, the basic unit for g_2 is one Chinese character but not a word. E.g. the 3rd feature is "the length of super-category characters".

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.

3.2 Cross-Lingual Instance Matching

Given the initial alignment set $\mathcal{A} = \{a_i\}_{i=1}^{m}$, cross-lingual instance matching is to generate a much larger alignment set $\mathcal{A}' = \{a_i\}_{i=1}^{n}$ $(n >> m)$ between $\mathcal{O}_1$ and $\mathcal{O}_2$. We are to learn the function $f : \mathcal{X} \times \mathcal{X}' \mapsto [0, 1]$ to predict the probability to be `equalTo` relation between two instances x and x'.

By automatically sampling a part of alignments from $\mathcal{A}$ as the training examples, we can learn the Logistic Regression model to get the function f. We firstly present the features for instance matching, and then introduce two preprocessing methods, namely maximum clique pruning and link annotation. Finally, we present the post-processing method.

Feature Definition. The features used in f are designed by the observation of cross-lingual knowledge consistency. Both the lexical similarities and link-based structural similarities are defined. We use the following *Set Similarity* as the basic metric for structural similarities, which has been proven to be quite effective in [15]. Given two instances a and b, let S_a and S_b be their related sets of entities, the ***Set Similarity*** between a and b is calculated as

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

where $\phi_{1\to 2}(\cdot)$ maps the set of entities in $\mathcal{G}_1$ (or $\mathcal{O}_1$) to their equivalent entities in $\mathcal{G}_2$ (or $\mathcal{O}_2$) if the alignment exists.

Table 3 shows the feature definition of f. As we can see, both the structural similarities in the online wikis and in the ontologies are used.

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 $\mathcal{O}_2$

To overcome the link sparseness, we use a smoothing method in our experiments when computing those structural features.

Maximum Clique Pruning. Due to the cross-lingual knowledge discordance, the knowledge distributions across different languages differs a lot. Our feature definition is apt to choose the correspondences sharing more common related entities. However, we observe that a lot of neighbor entities are not very related in online wikis. E.g. in Hudong Baike, the article “ “1月1日” ” (1st, Jan.) is linked to many dates without much relatedness. This will lead to some erroneous correspondences such as “ “1月1日” ”(1st, Jan.) `equalTo` “3rd, May”. We propose a maximum clique pruning to remove those structurally high linked but semantically low related structures. For each article in $\mathcal{G}_1$ or $\mathcal{G}_2$, we construct a local graph using this article and its linked articles. Then we calculate the maximum clique of this local graph. If the size of the maximum clique is larger 5, we prune the links between any two articles in the clique. In this way, lots of noise can be pruned from the online wikis. We add the similarities on the pruned network as new features for instance matching.

Link Annotation. Due to the link sparseness, the structural similarities across two heterogenous online wikis are quite sparse. To overcome this problem, we conduct a n-gram link annotation process to mine more links. The precision of link annotation is not sensitive, because we use the annotated links as new features for instance matching.

Heuristic Post-processing. Based on our observations, we propose the following rules to filter out some unreliable matching results: (1) *Multiple Correspondence.* If one English instance has been aligned to more than one Chinese instance, we remove all of those correspondences. (2) *Digits or Letters Co-occurrence.* If the Chinese instance's title contains a substring of more than two continuous digits or upper-case letters, we remove the correspondence if the English instance's title doesn't contain the same substring.

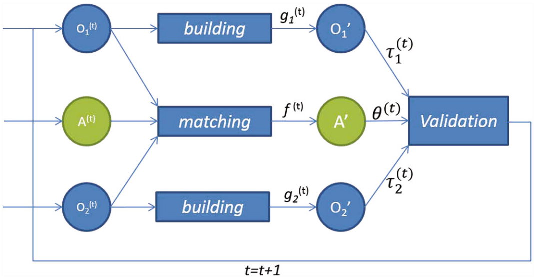

Fig. 3. Overview of boosting process in the iteration of t

3.3 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. Figure 3 shows the overview of our boosting method in the iteration of t. Our boosting strategies are different for the building and matching functions.

Boosting the Ontology Building Process. The performance of ontology building functions is related to the volume of manually labeled data sets. Our idea is to expand the training data sets automatically after each iteration by using a cross-lingual semantic validation method. The detailed strategies are as follows.

- Train the ontology building functions $g_1^{(t)}$, $g_2^{(t)}$ using current training data sets.
- Predict the unlabeled data sets using the learned $g_1^{(t)}$, $g_2^{(t)}$.
- Validate the predicted data using current cross-lingual alignments as follows: if $f^{(t)}(x_1, x_1^{'}) > \theta^{(t)}$ and $f^{(t)}(x_2, x_2') > \theta^{(t)}$, then we have $g_1^{(t)}(x_1, x_2) = g_2^{(t)}(x_1', x_2') = 1$ if $g_1^{(t)}(x_1, x_2) + g_2^{(t)}(x_1', x_2') > (\tau_1^{(t)} + \tau_2^{(t)})$, and $g_1^{(t)}(x_1, x_2) = g_2^{(t)}(x_1', x_2') = 0$ if $g_1^{(t)}(x_1, x_2) + g_2^{(t)}(x_1', x_2') < (\tau_1^{(t)} + \tau_2^{(t)})$ (we experimentally set $\theta^{(t)}$, $\tau_1^{(t)}$ and $\tau_2^{(t)}$ to be 0.9, 0.5 and 0.5 respectively. A higher parameter value generates a stricter validation result).
- Expand the training data sets using the cross-lingually validated data.
- Iteratively repeat this process for the next iteration.

Boosting the Instance Matching Process. The structural features of instance matching process are calculated based on the initial alignment set. More alignments help to harvest more precise features. Thus, our idea is to expand the alignment set automatically after each iteration. The detailed strategies are as follows.

- Train the instance matching function $f^{(t)}$ using current alignments.
- Predict the unlabeled data sets using $f^{(t)}$.
- Validate the predicted data sets as follows: if $f^{(t)}(x, x') > \theta^{(t)}$, then we have $f^{(t)}(x, x') = 1$ (we experimentally set $\theta^{(t)}$ to be 0.9).
- Expand the alignment set using the validated alignments.
- Iteratively repeat this process for the next iteration.

4 Experiments

We conduct the experiments using English Wikipedia and Hudong Baike. The English Wikipedia dump is archived in August 2012, and the Hudong Baike dump is crawled from Huong Baike's website in May 2012. We remove all those entities in English Wikipedia, whose titles contain the following strings: *wikipedia*, *wikiprojects*, *lists*, *mediawiki*, *template*, *user*, *portal*, *categories*, *articles*, *pages*, *by*. We also remove the articles in Hudong Baike, which do not belong to any categories of Hudong. Table 4 shows the statistics of the cleaned online wikis.

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

Using the cross-lingual links between English and Chinese Wikipedias, we get an initial alignment set containing 126,221 alignments between English Wikipedia and Hudong Baike. We use Stanford Parser [2] for extracting the head words and use the Weka [3] toolkit for implementing the learning algorithms. We first evaluate the effectiveness of proposed mono-lingual ontology building and cross-lingual instance matching methods respectively, and then evaluate the proposed boosting approach as a whole.

4.1 Mono-Lingual Ontology Building

For the evaluation of mono-lingual ontology building, we randomly selected 3,000 English `subClassOf`, 1,500 Chinese `subClassOf`, 3,000 English `instanceOf`, and 1,500 Chinese `instanceOf` examples. We ask 5 graduate students of Tsinghua University to help us manually label those examples. The examples consented by more than 3 students are kept. Table 5 shows the detail of our labeled examples.

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.

We conduct our experiments with a 5-fold cross-validation, and compare our Logistic Regression (LR) model with two baselines, namely Naïve Bayes (NB) and Support Vector Machines (SVM), using the same features defined in Sect. 3.1. As shown in Table 6, LR outperforms NB a lot and achieves comparative performance as the SVM (in most cases also outperforms SVM on F1-measure). In consideration of computation cost of the boosting process, our LR method is a good choice owing to its excellent learning efficiency.

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.

Table 6 also shows the cross-lingual performance comparison of `subClassOf` and `instanceOf` respectively. We find that English `instanceOf` performs better than Chinese `instanceOf`, but Chinese `subClassOf` is better than English `subClassOf`. This is because the 2nd and 3rd features in learning the building functions are linguistic related. The features are quite effective in learning English `instanceOf` and Chinese `subClassOf` respectively. That indicates the possibility to mutually improve the performance by the boosting process.

4.2 Cross-Lingual Instance Matching

In order to evaluate the cross-lingual instance matching method, we randomly select 3,000 initial alignments as the ground truth. We also automatically sample 10,000 random positive and 25,000 random negative alignments as the training data sets. In the experiments, we aim to investigate how the instance matching method performs before and after the heuristic post-processing (**HP**), and how the instance matching performs with different numbers of alignments. Therefore, we conduct four groups of experiments, each of which uses different number of alignments. In each group, we also compare the performance of our method before and after the heuristic post-processing. Table 7 shows the detailed results. The precision of our method is relatively high but the recall is rather low. We think this still works for our boosting method because the recalled alignments can be enriched iteratively even the recall is relatively low. However, a low precise alignment results will deteriorate the boosting process rapidly.

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**

As we can see from Table 7, in each group of the experiments, our method always performs better after the heuristic post-processing (especially for the precision). It shows the heuristic post-processing method can effectively filter out the unreliable matching results. On the other side, the F1-measure of our approach always increases when more alignments are used. Therefore, expanding the initial alignment set iteratively is important for improving the instance matching performance.

4.3 Boosting to Building a Large-Scale Ontology

At last, we evaluate our approach as a whole. For ontology building, we use the same labeled data sets and iteratively boost our approach. Table 8 shows that the performance of the four ontology building functions increases in each iteration. In particular, the precision and recall of Chinese `instanceOf` function goes from 65.0% and 63.0% to 96.7% and 96.9% respectively. As we can see, the performance after three iterations is excellent.

Table 8. Results of boosting to build the ontology. (%)

Iteration	`subClassOf` **en**			`subClassOf` **zh**			`instanceOf` **en**			`instanceOf` **zh**		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Iteration 1	80.8	88.2	84.4	82.0	100	90.1	87.4	97.1	92.0	65.0	63.0	64.0
Iteration 2	87.3	91.8	89.5	91.8	98.6	95.1	93.3	98.4	95.8	91.4	89.1	90.2
Iteration 3	87.7	93.4	**90.5**	94.8	99.3	**97.0**	97.3	99.6	**98.4**	96.7	97.0	**96.9**

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

In our experiments, we stop after the third iteration and successfully get two ontologies as shown in Table 9. For ontology matching, we use the same training data sets and all of the 126,221 alignments as the initial alignment set. We iteratively repeat the boosting process and 31,108 new alignments are found after 100 iterations. Due to the high computation cost, more iterations are still ongoing to find more alignments.

Table 9. Results of built ontology

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

5 Related Work

Multi-lingual Ontology Building. Ontology building is to generate an ontology concerning some specific domains in the form of Resource Description Framework. Current ontology building strategies can be grouped into three categories, namely manual construction, crowdsourcing based approach [13] and open Web extraction approach. The costly manual constructed ontologies, such as WordNet, HowNet and Cyc, are relatively high-quality but usually only cover parts of facts and are costly to maintain. Crowdsourcing based approach is becoming a prevalent method for building a large-scale and regularly updated ontology. DBpedia, by making the Wikipedia machine-readable, is a representative of this approach [1]. YAGO [12], MENTA [6] and BabelNet [7] are other multi-lingual ontologies based on WordNet and Wikipedia. Zhishi.me [8] is a Chinese knowledge base by integrating Hudong Baike, Baidu Baike and Chinese Wikipedia. XLORE [16] is a multilingual ontology generated from Hudong Baike, Baidu Baike, Chinese Wikipedia and English Wikipedia. Ponzetto and Strube have proposed some methods based on connectivity in the network and lexico-syntactic matching to derive a taxonomy from Wikipedia [9]. The open Web extraction approach aims to find a wider range of knowledge in the Web. This method gives us more opportunities to harvest more knowledge, but involves more noise and need to build an ontology from scratch. Probase [17] and TextRunner [18] are representatives of open Web extraction approach. Our proposed approach is a crowdsourcing based cross-lingual ontology building method.

Cross-lingual Ontology Matching. Ontology matching is to find equivalent correspondences between semantically related entities of ontologies [4,11]. Current ontology matching strategies can be grouped into two categoies, namely heuristic-based approach and machine learning-based approach. By manually defining some weights or threshold values, such heuristic-based approaches as similarity flooding and similarity aggregation can resolve the ontology matching problem quite efficiently and effectively. RiMOM [5] is a multi-strategy ontology alignment framework. The machine learning-based approach is to learn the weights and threshold values automatically. Rong et al. have proposed a transfer learning-based binary classification approach for instance matching [10]. Wang et al. have proposed a linkage factor graph model to match the instances across heterogenous wiki knowledge bases [15]. Current cross-lingual ontology matching approaches usually employ a generic two-step method, where ontology labels are translated into the target natural language first and monolingual matching techniques are applied next [5,14]. Wang et al. proposed a language-independent linkage factor graph model for instance matching [15]. Our proposed approach is a classification-based language-independent boosting method.

6 Conclusion and Future Work

In this paper, we propose a boosting method to build a large-scale cross-lingual ontology. The performance of ontology building and instance matching is reinforced iteratively. In particular, the performance of Chinese `instanceOf` function get a high 32% improvement for F1-measure. In our future work, we will iteratively find more cross-lingual instance alignments and crawl more Hudong Baike articles to enrich the Chinese instances. We will also improve our cross-lingual instance matching model to improve the recall, which is relatively low currently.

References

1. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: DBpedia: a nucleus for a web of open data. In: Aberer, K., Choi, K.-S., Noy, N., Allemang, D., Lee, K.-I., Nixon, L., Golbeck, J., Mika, P., Maynard, D., Mizoguchi, R., Schreiber, G., Cudré-Mauroux, P. (eds.) ASWC/ISWC -2007. LNCS, vol. 4825, pp. 722–735. Springer, Heidelberg (2007). doi:10.1007/978-3-540-76298-0_52
2. Green, S., de Marneffe, M.C., Bauer, J., Manning, C.D.: Multiword expression identification with tree substitution grammars: a parsing tour de force with french. In: EMNLP (2011)
3. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. SIGKDD **11**, 10–18 (2009)
4. Jean-Mary, Y.R., Shironoshita, E.P., Kabuka, M.R.: Ontology matching with semantic verification. Web Semant. **7**, 235–251 (2009)
5. Li, J., Tang, J., Li, Y., Luo, Q.: RiMOM: a dynamic multistrategy ontology alignment framework. TKDE **21**, 1218–1232 (2009)
6. de Melo, G., Weikum, G.: MENTA: inducing multilingual taxonomies from wikipedia. In: CIKM (2010)
7. Navigli, R., Ponzetto, S.P.: BabelNet: the automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artif. Intell. **193**, 217–250 (2012)
8. Niu, X., Sun, X., Wang, H., Rong, S., Qi, G., Yu, Y.: Zhishi.me - weaving Chinese linking open data. In: Aroyo, L., Welty, C., Alani, H., Taylor, J., Bernstein, A., Kagal, L., Noy, N., Blomqvist, E. (eds.) ISWC 2011. LNCS, vol. 7032, pp. 205–220. Springer, Heidelberg (2011). doi:10.1007/978-3-642-25093-4_14
9. Ponzetto, S.P., Strube, M.: Deriving a large scale taxonomy from wikipedia. In: AAAI (2007)
10. Rong, S., Niu, X., Xiang, E.W., Wang, H., Yang, Q., Yu, Y.: A machine learning approach for instance matching based on similarity metrics. In: Cudré-Mauroux, P., Heflin, J., Sirin, E., Tudorache, T., Euzenat, J., Hauswirth, M., Parreira, J.X., Hendler, J., Schreiber, G., Bernstein, A., Blomqvist, E. (eds.) ISWC 2012. LNCS, vol. 7649, pp. 460–475. Springer, Heidelberg (2012). doi:10.1007/978-3-642-35176-1_29
11. Shvaiko, P., Euzenat, J.: Ontology matching: state of the art and future challenges. TKDE **25**, 158–176 (2013)
12. Suchanek, F.M., Kasneci, G., Weikum, G.: Yago: a core of semantic knowledge. In: WWW (2007)
13. Tang, J., Leung, H.f., Luo, Q., Chen, D., Gong, J.: Towards ontology learning from folksonomies. In: IJCAI (2009)
14. Trojahn, C., Quaresma, P., Vieira, R.: A framework for multilingual ontology mapping. In: LREC (2008)
15. Wang, Z., Li, J., Wang, Z., Tang, J.: Cross-lingual knowledge linking across wiki knowledge bases. In: WWW (2012)
16. Wang, Z., Li, J., Wang, Z., Li, S., Li, M., Zhang, D., Shi, Y., Liu, Y., Zhang, P., Tang, J.: XLore: A large-scale English-Chinese bilingual knowledge graph. In: ISWC (2013)
17. Wu, W., Li, H., Wang, H., Zhu, K.Q.: Probase: a probabilistic taxonomy for text understanding. In: SIGMOD (2012)
18. Yates, A., Cafarella, M., Banko, M., Etzioni, O., Broadhead, M., Soderland, S.: Textrunner: open information extraction on the web. In: NAACL-Demonstrations (2007)