Cross-Lingual Knowledge Validation Based Taxonomy Derivation from Heterogeneous Online Wikis

Zhigang Wang[†] Juanzi Li[†] Shuangjie Li[†] Mingyang Li[†] Jie Tang[†] Kuo Zhang[‡] Kun Zhang[‡]
[†] Department of Computer Science and Technology, Tsinghua University, Beijing, China
{wzhigang, ljz, lsj, lmy, tangjie}@keg.cs.tsinghua.edu.cn
[‡] Sogou Incorporation, Beijing, China
{zhangkuo, zhangkun}@sogou-inc.com

Abstract

Creating knowledge bases based on the crowd-sourced wikis, like Wikipedia, has attracted significant research interest in the field of intelligent Web. However, the derived taxonomies usually contain many mistakenly imported taxonomic relations due to the difference between the user-generated subsumption relations and the semantic taxonomic relations. Current approaches to solving the problem still suffer the following issues: (i) the heuristic-based methods strongly rely on specific language dependent rules. (ii) the corpus-based methods depend on a large-scale high-quality corpus, which is often unavailable. In this paper, we formulate the cross-lingual taxonomy derivation problem as the problem of cross-lingual taxonomic relation prediction. We investigate different linguistic heuristics and language independent features, and propose a cross-lingual knowledge validation based dynamic adaptive boosting model to iteratively reinforce the performance of taxonomic relation prediction. The proposed approach successfully overcome the above issues, and experiments show that our approach significantly outperforms the designed state-of-the-art comparison methods.

Introduction

Global multi-lingual knowledge bases, which semantically represent the world's truth in the form of machine-readable graphs composed of classes, instances and relations, are at the heart of the intelligent Web, and significantly improve many applications such as cross-lingual information retrieval (Potthast, Stein, and Anderka 2008), machine translation (Wentland et al. 2008), and deep question answering (Yahya et al. 2012; Agirre, de Lacalle, and Soroa 2013). Projects like DBpedia (Auer et al. 2007), YAGO (Hoffart et al. 2013), MENTA (de Melo and Weikum 2010), XLORE (Wang et al. 2013) and BabelNet (Navigli and Ponzetto 2012) are constructing such knowledge bases by extracting structured information from Wikipedia, which is one of the most popular crowd-soured online wikis on the Web. Compared with the manually created knowledge bases such as Cyc (Sharma and Forbus 2013) and WordNet (Miller 1995), those knowledge bases constructed based on the online wikis become the more and more popular, and own the following advantages: automatically or semi-automatically generated, domain independent, easy to be maintained, user interest oriented, and usually with high accuracy and high coverage.

By treating each category and disambiguated article as one candidate class and instance respectively, the taxonomies are directly derived from the online wikis by transforming the user-generated subsumption relations, namely `subCategoryOf` between two categories and `articleOf` from one article to one category, into the semantic taxonomic relations, which are `subClassOf` between two classes and `instanceOf` from one instance to one class. However, the user-generated subsumption relations in the wikis and the semantic taxonomic relations in the knowledge bases are not exactly the same. The well-defined `subClassOf` and `instanceOf` essentially represent the `isA` relation, while freely edited `subCategory` and `articleOf` cover another `topicOf` relation, which denotes the topic related relation and generates the noise in the derived taxonomy. As Figure 1 shows, reasoning based on the directly derived taxonomy, we mistakenly conclude the fact that '`Barack Obama`' (person) `isA` '`Chicago, Illinois`' (location), which apparently should be the topic related relation.

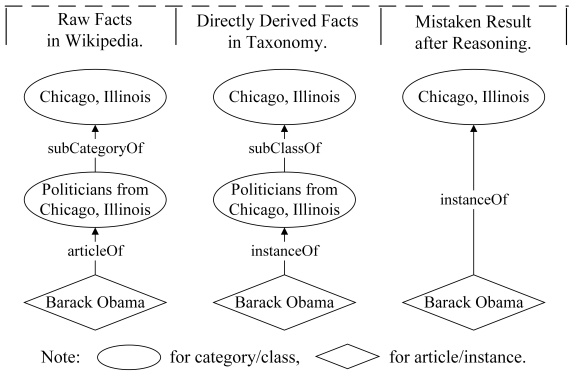


Figure 1: Example of Mistaken Derived Facts.

Recognizing this problem, several ontology learning approaches have been proposed to identify the correct taxonomic or `isA` relations. De Melo and Weikum (2010) pro-

posed a heuristic-based linking method to discover potential taxonomic relations in Wikipedia. The central idea is that "*the subsumption relation in Wikipedia can be found with high accuracy by determining whether the head words of the categories or articles are plural or singular, and countable or uncountable*" (Hoffart et al. 2013). Traditional corpus-based methods deal with this problem by taking texual corpus as inputs and inducing the corresponding taxonomic relations from the texts. Ponzetto and Strube (2007) utilized both the heuristic-based method and the corpus-based method to derive a taxonomy by distinguishing the `isA` relations from the `topicOf` relations in the English Wikipedia.

However, the previous approaches still suffer the following problems: (i) the heuristic-based methods (de Melo and Weikum 2010; Ponzetto and Strube 2007) strongly rely on the accuracy of head word recognization algorithm, and the language dependent rules could not handle some languages with no explicit plural/singual forms, such as Chinese and Japanese; (ii) the corpus-based methods (Ponzetto and Strube 2007) depend on a large-scale corpus with high quality, which in fact is often unavailable. Thus, the generated taxonomies are often small, mostly domain dependent, and with a rather poor performance (Ponzetto and Strube 2007; Buitelaar, Cimiano, and Magnini 2005).

In this paper, we systematically study the problem of cross-lingual knowledge validation based taxonomy derivation from heterogeneous online wikis. Based on the found cross-lingual links by Wang et al. (2012; 2013), the problem of cross-lingual taxonomic relation prediction is at the heart of our work. The problem is non-trivial and poses the following challenges.

Linguistics. Existing heurisitc-based methods heavily depend on the linguistic-specific rules. Are there some heuristic rules that work for other languages? Can we find some language-independent features for taxonomic relation prediction?

Model. Current corpus-based methods strongly rely on an input corpus with high quality, which is often unavailable. If we have a small number of labeled relations, how can we define a uniform model to predict the taxonomic relations based on kinds of features?

Performance. Is it possible to learn a robust model on the limited number of labeled relations? In consideration that the model may have different advantages on different languages, can we iteratively utilize the cross-lingual information to mutually reinforce the learning performance across different languages?

Driven by these challenges, we empirically investigate several important features and propose a unified boosting model to solve the problem of cross-lingual taxonomic relation prediction. Our contributions include:

- We formally formulate the problem of cross-lingual taxonomy derivation from heterogeneous online wikis, and analyze language dependent heuristics and language independent features for taxonomic relation prediction.
- We propose the **D**ynamic **A**daptive **B**oosting (**DAB**) model for cross-lingual taxonomy derivation. To improve the learning performance of taxonomic relation prediction, our model is trained iteratively on a dynamic active training set, where the training examples are weighted sampled from the pre-labeled data and the cross-lingual validated predicted data. We utilize a cross-lingual validation method to avoid potential performance deterioration.
- We evaluated our DAB model on an elaborately labeled dataset from English Wikipedia and Chinese Hudong Baike. Our model outperforms the heuristic linking method, non-boosting method and the AdaBoost method in both precision and recall.

The rest of this paper is organized as follows. Section 2 formally defines the problem of cross-lingual knowledge validation based taxonomy derivation and some related concepts. Section 3 reveals our proposed Dynamic Adaptive Boosting model in detail. Section 4 presents the evaluation results and Section 5 outlines some related work. Finally we conclude our work in Section 6.

Problem Formulation

In this section, we formally define the cross-lingual taxonomy derivation problem. Here, we first define the input wikis according to the mechanism of crowd-sourced wikis.

A ***wiki*** is a directed graph containing the collaboratively edited categories, articles and user-generated subsumption relations. It can be formally represented as $W = \{C, A\}$, where $c \in C$ denotes a *category* and $a \in A$ represents an *article*.

Each disambiguated ***article*** in the wiki describes a specific thing, and is one candidate instance in the taxonomy to be derived. Each article is linked to some categories by the `articleOf` relations. We formally represent each article a as a 5-tuple $a = \{label(a), comment, C(a), A(a), P(a)\}$, where $label(a)$ denotes the title of the article; $comment$ is the summarized texual description of a; $C(a)$ represents categories of a; $A(a)$ is the set of linked articles of a, and $P(a)$ is the set of attributes (or properties) used in the infobox of a.

Each ***category*** in the wiki is to group the articles on similar topics together, and is one candidate class in the taxonomy to be derived. Categories on similar topics are organized as a tree by the `subCategoryOf` relations. Each category can be represented as a 4-tuple $c = \{label(c), A(c), C(c), P(c)\}$, where $label(c)$ denotes the title of the category; $A(c)$ represents articles connected to c; $C(c)$ is the set of categories of $A(c)$, and $P(c)$ is the set of attributes (or properties) used in the infoboxes of $A(c)$.

We take each category c as one class and each article a as one instance. As shown in Figure 1, the freely user-generated `subCategory` and `articleOf` relations cover another `articleOf` relation other than the semantic `isA` relation. Thus, we define the problem of deriving a semantically organized taxonomy from the online wikis as follows.

Taxonomy Derivation. Given the input wiki W, taxonomy derivation is the process of recognizing whether there is a correct `subClassOf` relation for each `subCategoryOf` from $c_i \in C$ to $c_j \in C$, and recognizing whether there is a correct `instanceOf` relation for each `articleOf` from $a \in A$ to $c \in C$.

The problem of taxonomic relation prediction is at the heart of taxonomy derivation. In this paper, we tackle this problem as a binary classification problem by learning the following two functions.

- **subClassOf Prediction Function**. $f : C \times C \mapsto \{+1, -1\}$ to predict whether the `subCategoryOf` relation from the category $c_i \in C$ to the category $c_j \in C$ is a correct `subClassOf` or not (+1 for positive and -1 for negative).
- **instanceOf Prediction Function**. $g : A \times C \mapsto \{+1, -1\}$ to predict whether the `articleOf` relation from the aritcle $a \in A$ to the category $c \in C$ is a correct `instanceOf` or not (+1 for positive and -1 for negative).

The online wikis contain lots of different subsumption relations and it is challenging to build an enough robust model for one particular taxonomic relation in one language based on limited number of labeled data. On the other hand, as we introduced and will present later, different heuristic rules have different advantages in taxonomic relation derivation across different languages. Thus, it is promising to iteratively improve the learning performance of taxonomic relation prediction across different languages via knowledge validation using the cross-lingual information.

The ***cross-lingual links*** is the set of equivalent categories/articles between two wikis in different languages. It can be formally represented as $CL = \{(x, x^{'})\}$, where x and $x^{'}$ are two equivalent categories/articles from two wikis in different languages. Besides, we have that $CL = cCL \cup aCL$, where cCL denotes the set of cross-lingual categories and aCL represents the set of cross-lingual articles.

Cross-lingual Taxonomy Derivation Given two input wikis W_1, W_2 in different languages (English and Chinese in this paper) and the set of cross-lingual links CL, cross-lingual taxonomy derivation is a cross-lingual knowledge validation based boosting process of inducing two taxonomies simultaneously. Figure 2 shows the framework of cross-lingual taxonomy derivation.

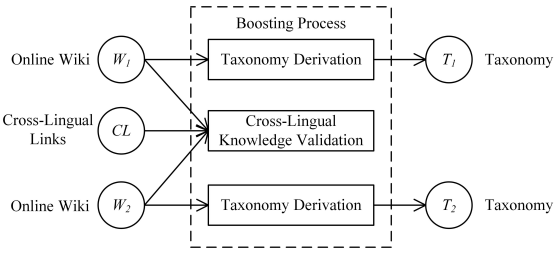


Figure 2: Framework.

The Proposed Approach

Dynamic Adaptive Boosting (DAB) model is to simultaneously learn four taxonomic prediction functions f^{en}, f^{zh}, g^{en} and g^{zh} in T iterations, 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. For each taxonomic relation prediction function, DAB tries to learn a linear combination of a series of weak classifiers. In this section, we will present the detailed approach for weak classifiers, and then introduce our boosting model as a whole.

Weak Classifier

We utilize the binary classifier for the basic learner and use the Decision Tree (Yuan and Shaw 1995) as our implementation. The defined features include the linguistic heuristic features and the language independent structural features.

Linguistic Heuristic Features We define three kinds of linguistic features as: English features, Chinese features and common features for `instanceOf`.

Feature 1: English Features.

Whether the head words of $label(a)$/$label(c)$ are plural or singular. The occurrence of singular form usually implies a negative taxonomic relation. E.g. '`Educational People`' `subClassOf` '`Education`' (singular) is a negative example.

Feature 2: Chinese Features.

For `subClassOf`, whether the super-category's label is the prefix/suffix of the sub-category's label. For `instanceOf`, whether the category's label is the prefix/suffix of the article's label. The prefix usually implies the negative relation, while the suffix implies the positive one. E.g. In Chinese '`Educational People`' `subClassOf` '`Education`' (prefix) is a negative example.

Feature 3: Common Features for `instanceOf`.

Given $a \in A$ and $c \in C$, whether the *comment* of a contains the $label(c)$ or not. The containing relation usually implies the positive example. E.g. '`Barack Obama`' `instanceOf` '`President`' is a positive example because the comment of '`Barack Obama`' contains "*Barack Obama is the 44th and current President ...*".

Structural Features We define six language independent structural features based on the Normalized Google Distance (Cilibrasi and Vitanyi 2007): three for `subClassOf` and the other three for `instanceOf`.

Feature 4: Article Divergence for `subClassOf`.

Articles $A(c)$ are the set of articles which connect to the category c. The article divergence computes the semantic relatedness between the articles of two categories. Given two categories $c \in C$ and $c' \in C$, the article divergence is computed as

$$d_a(c, c') = \frac{\log(\max(|A(c)|, |A(c')|)) - \log(|A(c) \cap A(c')|)}{\log(|A|) - \log(\min(|A(c)|, |A(c')|))} \quad (1)$$

Feature 5: Property Divergence for `subClassOf`.

Properties $P(c)$ are the attributes defined in the infoboxes of $A(c)$. The property divergence computes the semantic relatedness between the properties of two categories.

$$d_p(c, c') = \frac{\log(\max(|P(c)|, |P(c')|)) - \log(|P(c) \cap P(c')|)}{\log(|P|) - \log(\min(|P(c)|, |P(c')|))} \quad (2)$$

Feature 6: Category Divergence for `subClassOf`.

Categories $C(c)$ are the categories of $A(c)$. The category divergence is computed as

$$d_c(c, c') = \frac{\log(\max(|C(c)|, |C(c')|)) - \log(|C(c) \cap C(c')|)}{\log(|C|) - \log(\min(|C(c)|, |C(c')|))} \tag{3}$$

Feature 7: Article Divergence for `instanceOf`.

Given the article $a \in A$ and the category $c \in C$, the article divergence is calculated as

$$d_a(a, c) = \frac{\log(\max(|A(a)|, |A(c)|)) - \log(|A(a) \cap A(c)|)}{\log(|A|) - \log(\min(|A(a)|, |A(c)|))} \tag{4}$$

Feature 8: Property Divergence for `instanceOf`.

Properties $P(a)$ are the attributes in the infobox of a. Given the article $a \in A$ and the category $c \in C$, the property divergence is

$$d_p(a, c) = \frac{\log(\max(|P(a)|, |P(c)|)) - \log(|P(a) \cap P(c)|)}{\log(|P|) - \log(\min(|P(a)|, |P(c)|))} \tag{5}$$

Feature 9: Category Divergence for `instanceOf`.

Given the article a, the expanded categories $CC(a)$ are the categories to which the articles $A(a)$ connect. The category divergence is computed as

$$d_c(a, c) = \frac{\log(\max(|CC(a)|, |C(c)|)) - \log(|CC(a) \cap C(c)|)}{\log(|C|) - \log(\min(|CC(a)|, |C(c)|))} \tag{6}$$

Boosting Model

As Figure 3 shows, for each taxonomic relation prediction function, DAB repeatedly calls a weak learner h_t trained on the dynamic changed training set A_t in $t = 1, ..., T$ rounds. The main idea of the algorithm is to maintain a set of weights D_t over A_t to iteratively improve the learning performance, and to maintain a dynamic changed training set A_t, namely the active set, to achieve a better generalization ability. In detail, DAB iteratively learns each taxonomic relation prediction function as follows.

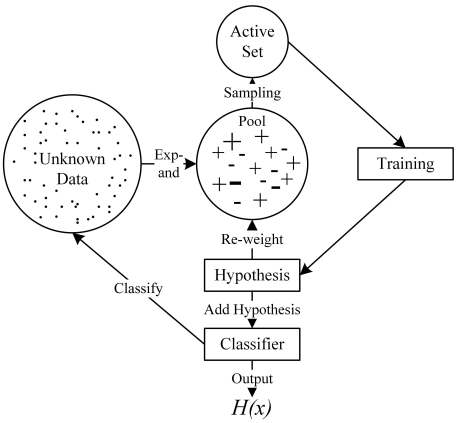


Figure 3: Dynamic Adaptive Boosting Model.

Definition. **Active Set** A represents the set of labeled data for the learning of the relation prediction function. **Pool** P denotes the set of all labeled data and we have $A \subset P$. **Unknown Data Set** U represents the set of all unlabeled data and we have $|U| \gg |P|$. CL is the set of cross-lingual links between categories/articles.

Initialization. $P_1 = \{(x_1, y_1), (x_2, y_2), ..., (x_m, y_m)\}$ is the set of all labeled data, where x denotes the labeled example and $m = |P_1|$ is the initial size of the pool. We set $A_1 = P_1$ and set the weights D_1 over A_1 as $D_1(i) = \frac{1}{m}$ for $i = 1, 2, ..., m$.

Learning Process. Each function $f^{(k)}$ is a linear combination of a series of weak classifiers $\{h_t^{(k)}\}_{t=1}^T$ trained on the dynamic changed training set $A_t^{(k)}$, namely the active set. For $t = 1, 2, ..., T$,

- Train a basic classifier with the minimal weighted error rate on current active set A_t,

$$h_t(x) = \arg\min_{h_j \in H} \epsilon_t \tag{7}$$

  where the error rate

$$\epsilon_t = \sum_{x_i \in A_t} D_t(i) \cdot I\{h_t(x_i) \neq y_i\} \tag{8}$$

- Check whether $\epsilon_t < \frac{1}{2}$. If not, stop the learning process.
- Choose $\alpha_t = \frac{1}{2} \cdot \ln(\frac{1-\epsilon_t}{\epsilon_t})$ and re-weight the weight vector as

$$\begin{aligned} D_{t+1}(i) &= \frac{D_t(i)}{Z_t} \cdot e^{-\alpha_t y_i h_t(x_i)} \\ &= \frac{D_t(i)}{Z_t} \cdot \begin{cases} e^{-\alpha_t}, & h_t(x_i) = y_i \\ e^{\alpha_t}, & h_t(x_i) \neq y_i \end{cases} \end{aligned} \tag{9}$$

  where Z_t is the normalization factor and

$$Z_t = \sum_i D_t(i) \cdot e^{-\alpha_t y_i h_t(x_i)} \tag{10}$$

- Predict U_t using $l_t(x) = \sum_{i=1}^t \alpha_i \cdot h_i(x)$ and validate the predicted results using CL to get the validated results V_t. In detail, we use the English and Chinese predicted results together to validate the results.

  If the elements (categories or articles) of x (English) and $x^{'}$ (Chinese) are linked by the cross-lingual links, we get the validated results as: $y^{en}(x) = y^{zh}(x^{'}) = 1$ if both $l_t^{en}(x)$ and $l_t^{zh}(x^{'})$ are greater than the threshold θ, and $y^{en}(x) = y^{zh}(x^{'}) = -1$ if both $l_t^{en}(x)$ and $l_t^{zh}(x^{'})$ are less than the threshold $-\theta$, where the threshold θ is experimentally set as 0.93.
- Expand the pool as $P_{t+1} = P_t + V_t$. And update the unknown dataset as $U_{t+1} = U_t - V_t$.
- Resample the active set as $A_{t+1} = sample(A_t, \delta, V_t)$. The size of the active set is constant, and we randomly sample some examples from the former active set, and replace them with the validated examples.

  In detail, we divide the validated results V_t as the correctly classified examples by $h_t(x)$ and the wrongly classified examples by $h_t(x)$. For each example in these two

parts, we randomly replace an correctly or wrongly classified example by $h_t(x)$ from A_t with it. The parameter δ is used to limit the update speed, where in each iteration no more than $\delta \cdot m$ examples are replaced from A_t. δ is experimentally set as 0.2.

Model Analysis. Compared to the real AdaBoost model (See Equation 7-10), DAB model utilized a similar weight vector updating strategy and thus owns the similar training/generalization bounds (Schapire 1999). On the other hand, owing to the dynamic changed active set, the DAB model has a better generalization ability than the real AdaBoost, which will be presented in the next section.

Experiments

In this paper, our proposed approach for cross-lingual taxonomy derivation is a general model, and can be used for any two wikis in different languages. To evaluate our approach, we conduct our experiments using English Wikipedia and Chinese Hudong Baike. The English Wikipedia dump is archived in August 2012, and the Hudong Baike dump is crawled from Huong's website in May 2012. We remove the Wikipedia entities whose titles contain one of the following strings: *wikipedia*, *wikiprojects*, *lists*, *mediawiki*, *template*, *user*, *portal*, *categories*, *articles*, *pages*, and *by*. We also remove the Hudong articles which don't belong to any categories. Finally, we get the English Wikipedia dataset containing 561,819 categories and 3,711,928 articles. The Chinese Hudong Baike dataset contains 28,933 categories and 980,411 articles.

Experiment Settings

Data Sets We randomly selected 3,000 English `subCategoryOf` examples, 1,500 Chinese `subCategoryOf` examples, 3,000 English `articleOf` examples and 1,500 Chinese `articleOf` examples. We ask a professional team, composed of 5 graduate students in Tsinghua University, to help us manually label those relations. The examples which are consented by more than 4 students are kept. Table 1 shows the detail of the labeled dataset. We can see that the initial user-generated relations in the online wikis contain plenty of wrong semantic taxonomic relations.

Table 1: Labeled Data

Taxonomic Relation	#Positive	#Negative
English subClassOf	2,123	787
Chinese subClassOf	780	263
English instanceOf	2,097	381
Chinese instanceOf	638	518

Using the cross-lingual links between English and Chinese Wikipedias, we get 126,221 cross-lingual links between English Wikipedia and Hudong Baike. The data sets are available at *http://xlore.org/publications.action*.

Comparison Methods We define three state-of-the-art taxonomy derivation methods, which are Heuristic Linking (HL), Decision Tree (DT) and Adaptive Boosting with no cross-lingual knowledge validation (AdaBoost).

- **Heuristic Linking (HL)**. This method only uses the linguistic heuristic features as defined in Section 3, and trains the taxonomic relation prediction functions separately using the decision tree model.
- **Decision Tree (DT)**. This method uses both the linguistic heuristic features and the structural features as defined in Section 3, and trains the taxonomic relation prediction functions separately using the decision tree model.
- **Adaptive Boosting (AdaBoost)**. This method uses the same basic learner as defined in Section 3, and iteratively trains the taxonomic relation prediction functions using the real AdaBoost model (Schapire 1999).

Evaluation Metrics Two series of evaluation metrics are used to evaluate our approach.

- We use precision (P), recall (R) and F1-score (F1) to evaluate different taxonomy derivation methods.
- We use the error rate ϵ (see Equation 8) to evaluate the detailed boosting performance in each iteration of AdaBoost and DAB.

Result Analysis

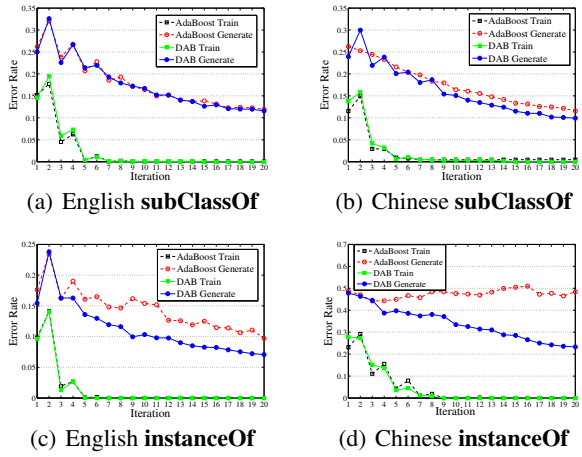
Performance Comparison To demonstrate the better generalization ability of DAB model, we conduct 2-fold cross-validation on the labeled dataset. Besides, in each iteration, we separate the cross-lingual validated results V_t into 2 fold and add one of them into the testing dataset, and use the other part to expand the pool P_{t+1}. We first run the DAB method and use the final testing dataset of DAB to evaluate the comparison methods. Both the comparison methods and DAB model use the default settings of Decision Tree in Weka (Hall et al. 2009). We use the Stanford Parser (Green et al. 2011) for head word extraction. The AdaBoost and DAB methods run 20 iterations. Table 2 gives the detailed results of four methods. It can be clearly observed that:

- Decision Tree method outperforms Heuristic Linking method in precision and recall, which demonstrate the effects of our defined structural features.
- Both AdaBoost method and DAB method are better than the Decision Tree method. It is clear that the boosting process strongly improves the learning performance.
- DAB method significantly outperforms AdaBoost method especially in the prediction of Chinese `instanceOf` relation, which proves the remarkable effects of the dynamic boosting process and the cross-lingual knowledge validation.
- The `subClassOf` features perform better in English, while the `instanceOf` features perform better in Chinese. That is because that the English heuristic features cover more `subClassOf` relations than the `instanceOf` relations, while the Chinese heuristic features cover more `instanceOf` relations. This also reveals the potential possibility for cross-lingual learning performance improvement.

Table 2: 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**

Boosting Contribution Figure 4 shows the error rates of four prediction functions in each iteration. We can see that the training performance of the DAB method and the AdaBoost method are comparable, while the generalization performance of DAB method is much better, which is because the dynamic changed active set A_t greatly improves the model's generalization ability. What's more, we also use the Naive Bayes learner as the weak classifier but generates much worse results, which corresponds to the fact that the boosting process prefer the learner with high variance and low bias like decision tree (Ting and Zheng 2003).



(a) English **subClassOf** (b) Chinese **subClassOf**

(c) English **instanceOf** (d) Chinese **instanceOf**

Figure 4: Boosting Contribution Comparison.

Related Work

In this section, we review some related work including taxonomy derivation and cross-lingual knowledge linking.

Taxonomy Derivation

Taxonomy derivation is necessary to build ontological knowledge (Tang et al. 2009). Many large-scale knowledge bases are built based on taxonomy derivation, such as DBpedia (Auer et al. 2007), YAGO (Hoffart et al. 2013), MENTA (de Melo and Weikum 2010), XLORE (Wang et al. 2013) and BabelNet (Navigli and Ponzetto 2012). BabelNet uses disambiguation context of Wikipages to find their equivalent WordNet (Miller 1995) senses. YAGO combines concept system in WordNet and instance system in Wikipedia by mapping leaf categories in Wikipedia to WordNet synsets. MENTA uses a heuristic based linking method and Markov chain-based ranking approach to integrate entities from Wikipedia and WordNet into a single coherent taxonomic class hierarchy. These methods are mainly designed for the English wikis. Another kind of approach dealing with this problem is taking advantage of information from rich textual corpus. Simone et al. (2012) utilize methods based on the connectivity of the wiki network and on applying lexicon-syntactic patterns to very large corpora to distinguish between `isA` and `notIsA` relations. Our dynamic adaptive boosting model doesn't rely on the background corpus and uses the language independent structural features to support wikis in non-English languages.

Cross-lingual Knowledge Linking

Current approaches for inducing cross-lingual knowledge links usually employ a generic two-step method, selecting missing link candidates using link structure of articles first and classifying those links with graph-based and linguistic-based information next. By defining proper features, Sorg et al. (2008) and Oh et al. (2008) employ such approaches and resolve the problem of discovering missing cross-lingual links quite effciently and effectively without information from other lexical resources. Wang et al. (Wang et al. 2012) employ a factor graph model which only used link-based features to find cross-lingual links between a Chinese wiki and English Wikipedia. They also take advantage of concept annotation which reveals relations between instances and concepts besides of inner links within instances, and a regression-based learning model to learn weights to aggregate similarities (Wang, Li, and Tang 2013). Traditional ontology matching also works for the task of knowledge linking (Trojahn, Quaresma, and Vieira 2008; Grace 2002; Jean-Mary, Shironoshita, and Kabuka 2009; Shvaiko and Euzenat 2013). Current knowledge linking methods, which focus on mining more links by utilizing these cross-lingual links, are a promising direction. Our method utilizes the found cross-lingual links to boost the learning performance of taxonomy derivation.

Conclusion and Future Work

In this paper, we propose a cross-lingual knowledge validation based dynamic adaptive boosting model to iteratively reinforce the performance of taxonomy derivation. The proposed approach significantly outperforms the designed state-of-the-art comparison methods. In our future work, we will automatically learn more cross-lingual validation rules and other reasoning strategies to improve the boosting process. We will also conduct more experiments on wikis in other languages.

Acknowledgements

The work is supported by 973 Program(No. 2014CB340504), NSFC (No. 61035004, No. 61222212), NSFC-ANR(No. 61261130588), National High-tech R&D Program (No. 2014AA015103), National Basic Research Program of China (No. 2012CB316006), Tsinghua University Initiative Scientific Research Program (20131089256, 20121088096), FP7-288342, MCM20130321, a research fund supported by Huawei Inc., Beijing Key Lab of Networked Multimedia, THU-NUS NExT Co-Lab and THU-Sogou Co-Lab.

References

Agirre, E.; de Lacalle, O. L.; and Soroa, A. 2013. Random walks for knowledge-based word sense disambiguation. In *Computational Linguistics '13*.

Auer, S.; Bizer, C.; Kobilarov, G.; Lehmann, J.; Cyganiak, R.; and Ives, Z. 2007. Dbpedia: A nucleus for a web of open data. In *ISWC'07*, 11–15. Springer.

Buitelaar, P.; Cimiano, P.; and Magnini, B. 2005. *Ontology Learning from Text: Methods, Evaluation and Applications*, volume 123 of *Frontiers in Artificial Intelligence and Applications Series*. Amsterdam: IOS Press.

Cilibrasi, R. L., and Vitanyi, P. M. 2007. The google similarity distance. *IEEE Transactions on Knowledge and Data Engineering* 19:370–383.

de Melo, G., and Weikum, G. 2010. Menta: Inducing multilingual taxonomies from wikipedia. In *CIKM'10*, 1099–1108. ACM.

Grace, M. C. 2002. Creating a bilingual ontology: A corpus-based approach for aligning wordnet and hownet. In *Proceedings of the 1st Global WordNet Conference*, 284–292.

Green, S.; de Marneffe, M.-C.; Bauer, J.; and Manning, C. D. 2011. Multiword expression identification with tree substitution grammars: A parsing tour de force with french. In *EMNLP'11*, 725–735. ACL.

Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; and Witten, I. H. 2009. The weka data mining software: an update. *SIGKDD Explor. Newsl.* 11(1):10–18.

Hoffart, J.; Suchanek, F. M.; Berberich, K.; and Weikum, G. 2013. Yago2: A spatially and temporally enhanced knowledge base from wikipedia. *Artif. Intell.*

Jean-Mary, Y. R.; Shironoshita, E. P.; and Kabuka, M. R. 2009. Ontology matching with semantic verification. *Web Semant.* 7(3):235–251.

Miller, G. A. 1995. Wordnet: A lexical database for english. *Commun. ACM*.

Navigli, R., and Ponzetto, S. P. 2012. Babelnet: the automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artif. Intell.*

Oh, J.-H.; Kawahara, D.; Uchimoto, K.; Kazama, J.; and Torisawa, K. 2008. Enriching multilingual language resources by discovering missing cross-language links in wikipedia. In *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*, 322–328.

Ponzetto, S. P., and Strube, M. 2007. Deriving a large scale taxonomy from wikipedia. In *AAAI'07*, 1440–1445. AAAI Press.

Potthast, M.; Stein, B.; and Anderka, M. 2008. A wikipedia-based multilingual retrieval model. In *Proceedings of the IR Research, 30th European Conference on Advances in Information Retrieval*, ECIR'08, 522–530. Springer-Verlag.

Schapire, R. E. 1999. A brief introduction to boosting. In *IJCAI'99*, 1401–1406.

Sharma, A. B., and Forbus, K. D. 2013. Automatic extraction of efficient axiom sets from large knowledge bases. In *AAAI'13*. AAAI Press.

Shvaiko, P., and Euzenat, J. 2013. Ontology matching: State of the art and future challenges. *IEEE Transactions on Knowledge and Data Engineering* 25(1):158–176.

Sorg, P., and Cimiano, P. 2008. Enriching the crosslingual link structure of wikipedia - a classification-based approach. In *Proceedings of the AAAI 2008 Workshop on Wikipedia and Artificial Intelligence*.

Tang, J.; Leung, H.-f.; Luo, Q.; Chen, D.; and Gong, J. 2009. Towards ontology learning from folksonomies. In *IJCAI'09*, 2089–2094. Morgan Kaufmann Publishers Inc.

Ting, K. M., and Zheng, Z. 2003. A study of adaboost with naive bayesian classifiers: Weakness and improvement. *Computational Intelligence* 19(2):186–200.

Trojahn, C.; Quaresma, P.; and Vieira, R. 2008. A framework for multilingual ontology mapping. In *LREC'08*.

Wang, Z.; Li, J.; Wang, Z.; and Tang, J. 2012. Cross-lingual knowledge linking across wiki knowledge bases. In *WWW'12*, 459–468. ACM.

Wang, Z.; Li, J.; Wang, Z.; Li, S.; Li, M.; Zhang, D.; Shi, Y.; Liu, Y.; Zhang, P.; and Tang, J. 2013. Xlore: A large-scale english-chinese bilingual knowledge graph. In *Posters and Demos of ISWC'13*.

Wang, Z.; Li, J.; and Tang, J. 2013. Boosting cross-lingual knowledge linking via concept annotation. In *IJCAI'13*.

Wentland, W.; Knopp, J.; Silberer, C.; and Hartung, M. 2008. Building a multilingual lexical resource for named entity disambiguation, translation and transliteration. In *LREC'08*. European Language Resources Association.

Yahya, M.; Berberich, K.; Elbassuoni, S.; Ramanath, M.; Tresp, V.; and Weikum, G. 2012. Deep answers for naturally asked questions on the web of data. In *Proceedings of the 21st International Conference Companion on World Wide Web*, WWW'12 Companion, 445–449. ACM.

Yuan, Y., and Shaw, M. J. 1995. Induction of fuzzy decision trees. *Fuzzy Sets Syst.* 69:125–139.